Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages

Rashmi Gangadharaiah
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, USA
rgangadh@cs.cmu.edu

N. Balakrishnan
SERC,
Indian Institute of Science
Bangalore, India
balki@serc.iisc.ernet.in

Abstract

This paper shows how certain perceptible linguistic rules help in the process of making Machine Translation simpler and computationally inexpensive when translating from English to a class of Indian Languages. The paper explains the method in detail for Kannada (Language spoken in Karnataka) and can be extended to other Indian Languages. A combination of a simple Generalized Example Based Machine Translation (G-EBMT) system and these linguistic rules gave a score of 0.7164 using the BLEU evaluation metric on a set of 100 sentences while translating sentences from English to Kannada.

1 Introduction

The demand for translators has been growing and the interest in translators has grown markedly with computer programs translating documents from one language to another. The commercial interests are growing at a very high rate. Google has started offering machine translation of their search results.

There are no perfect machine translators and all of the existing ones require human intervention for correcting the output. This is because of the restrictions inherent in each method. In Rule Based Machine Translation, adding all possible rules is time consuming and while adding a new rule, it has to be checked against all old rules for consistency. Transfer-oriented Machine Translation systems rely on similarity in syntactic structure between the source and the target sentence, which is rare. In Example Based Machine translation, it is hard to have a broad coverage of sentences and may be limited due to the size of the translation memory. Statistical Based Machine Translation cannot take into account long distance dependencies.

In (Robert Frederking and Sergei Nirenburg, 1994), it has been shown that the quality of MT systems is improved by using the best results obtained from a variety of systems working on the same text simultaneously. With all these factors in mind, now the problem is to know how to get “a good enough” translation in limited resources domain. This paper solves the problem by applying a minimum set of rules for Indian Languages. It obtains its results from (a) a simple G-EBMT (Ralf D. Brown, 1999) system and (b) a system built with small set of linguistic rules obtained from perceptible common features in linguistic constructs for Indian Languages.

Today, India has fifteen official languages. These languages originated from the Indo-Iranian branch of the Indo-European language family, the non-Indo-European Dravidian-family, Austro-Asiatic, Tai-Kadai and the Sino-Tibetan language families (Microsoft Encarta Online Encyclopedia, 1997). The languages that stem from the Dravidian family, are - *Tamil, Kannada, Malayalam* and *Telugu*, spoken in the South Indian states- Tamilnadu, Karnataka, Kerala and Andhra Pradesh. Most modern languages in North India, such as *Hindi, Urdu, Punjabi, Gujarati, Bengali, Marathi, Kashmir, Sindhi, Konkani, Rajasthani, Assamese and Oriya,* stem from Sanskrit and Pali.

The paper uses examples mostly in Kannada, but the method can be extended to other Indian Languages. Most Indian Languages can take any of the following six forms: SOV, OSV, OVS, VOS, VSO, SVO where, S is Subject, V is Verb, O is Object. However, as argued by many linguists (Mohanan, K. P., 1982), verbs behave differently from other constituents of S when they are not placed in the final position and hence, Indian languages are treated as verb final languages[1]. Hence, in this paper we treat Indian Languages in general, and Kannada in particular as verb final languages, with freedom of word order applied only to the NP's (Noun Phrases) and PP's (Prepositional Phrases) that are direct daughters of an S node in a constituent structure. Consider as an example, the word order for the sentence "she is going home" in Kannada (the sentences in Kannada have been transliterated using the "Om" Transliterator (Ganapathiraju Madhavi, et. all, 2005)),

aval'u(S)	manege(O)	hoguti*daal'e* (V)
she	home	going+is
manege(O)	aval'u(S)	hoguti*daal'e* (V)
home	she	going+is

This paper shows how the common features seen in linguistic constructs of all Indian Languages help in the process of Machine Translation. The process has been explained in detail and evaluated for Kannada.

The paper is organized as follows. Section 2 outlines the proposed method and explains it in detail. Section 3 reports the evaluation results obtained using this method. Section 4 concludes and suggests possible improvements.

2 Proposed Method

The flow of the proposed method is shown in Figure 1. The rest of the paper explains this method in detail only for Kannada.

As the meanings of idioms cannot be inferred from the meanings of the words that make it up, idioms are stored separately in a file in the following format:

bite the hand that feeds → un'd'a manege erad'u bage

When the input sentence is found in the file containing the idioms, the sentence after the arrow "→" is returned as its translation. If the input sentence is not present in the idioms file, the input sentence enters the G-EBMT module containing generalized sentence rules (explained in section 2.1). If a rule for the input sentence is not present, then, language specific rules, explained in section 2.2, are applied.

2.1 A Simple G-EBMT

EBMT has been proved to be successful, but requires large amount of pre-translated text. G-EBMT (Ralf D. Brown, 1999) has linguistically tagged entries in its database and reduces the amount of pre-translated text required. Similar examples are tokenized to show equivalence classes and stored as a generalized example.

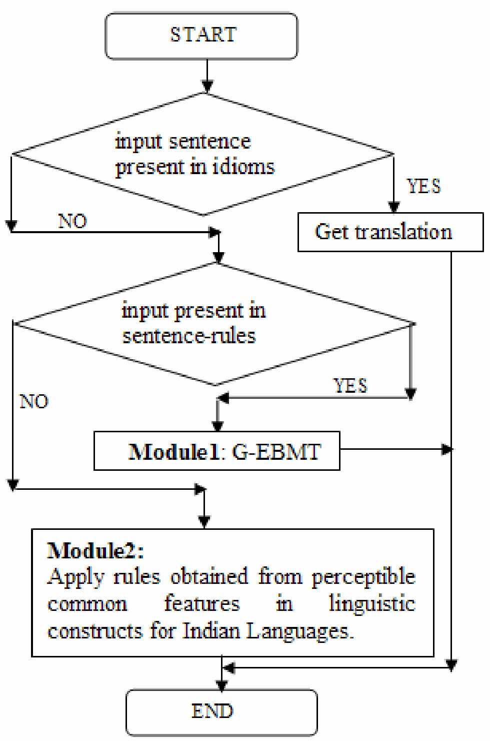
START
input sentence present in idioms
YES
NO
Get translation
input present in sentence-rules
YES
NO
Module1: G-EBMT
Module2: Apply rules obtained from perceptible common features in linguistic constructs for Indian Languages.
END

Figure1: Proposed method

Source sentence rule	→ Target sentence rule
she brought a <noun>	→ aval'u <noun> tandal'u

Input sentence	:	she brought a *dog*
Tokenized	:	she brought a <noun>
Translated	:	aval'u <noun> tandal'u
Output sentence	:	aval'u *naayi* tandal'u

In the above example, the input sentence (sentence to be translated) matches with one of the source sentence rules. The target sentence rule is obtained and the translation of the token (dog → naayi) is plugged in as shown.

In this module, the translation is done using example texts and equivalence classes only at the sentence level. Doing this at the phrase-level by combining all the partial matches gives a bad translation (shown in the example below). This is mainly because the freedom of word order is ap-

[1] As explained by K.P. Mohanan for Malayalam (Mohanan, K. P., 1982) putting the verb in a non-final position,
(i) Places a heavy nuclear pitch accent on the verb removing all the word melodies that occur after the nuclear accent.
(ii) Lengthens the final vowel of the verb.
This leads to a preference among speakers to place the verb in final position. Non-final verbs have heavy contrastive meanings not necessarily associated with shifting of other constituents. Also, there exist certain kinds of embedded clauses in which the verb cannot be scrambled away from the final position.

plied only to direct NP daughters of S in Indian Languages. Hence arranging these partial matches as a sequence of partial phrases leads to an ungrammatical sentence with wrong placement of verbs (maadalilla, paalisi). We evaluate our methods using the BLEU metric (Kishore Papineni et. al, 2002). The BLEU score computes the geometric average of modified *n*-gram precisions, using *n*-grams upto length N and positive weights which sum to one. The BLEU score (with 3(N)-gram precision) for both the phrase-level generalization and sentences-level generalization is shown in brackets next to the output sentences.

Input: they did not follow the rules

Phrase-level Generalization

1) <pron> did not → <pron> maadalilla
"they did not" → "avaru maadalilla"

2) "follow the <noun>" → "<noun>annu paalisi"
"follow the rules" → "niyamagal'l'annu paalisi"

Output: "avaru maadalilla niyamagal'annu paalisi" (0)

Sentence-level Generalization:

"<pron>did not follow the <noun>"→
"<noun> annu <pron> paalisalilla"

Output:
"niyamagal'l'annu avaru paalisalilla" (1)

As a result of this difficulty in combining partial matches, we have not used phrase level rules in our system.

2.2 Apply Language Specific rules

The various stages in this module are shown in Figure 2.

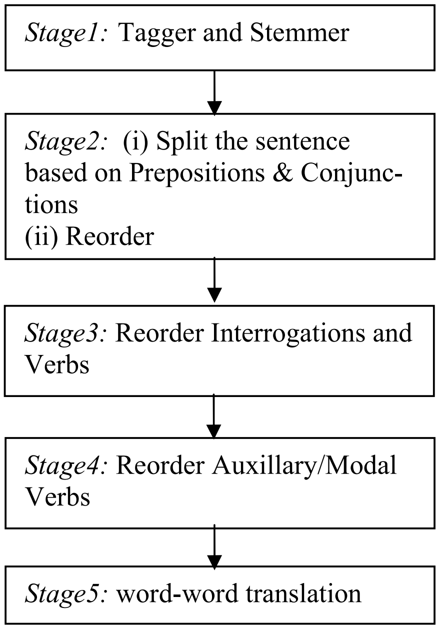
Stage1: Tagger and Stemmer
Stage2: (i) Split the sentence based on Prepositions & Conjunctions
(ii) Reorder
Stage3: Reorder Interrogations and Verbs
Stage4: Reorder Auxillary/Modal Verbs
Stage5: word-word translation

Figure 2*:* five stages in applying the rules

The procedure is discussed in detail for the sentence "he is playing in my house" in each of the stages mentioned below and the output of each stage is summarized in Appendix A. In each stage, for general applicability we have also explained with examples any special cases.

Stage1:

The input sentence is preprocessed to convert from uppercase to lower case letters. Part-of-Speech tagging using decision trees (Helmut Schmid, 1994) is performed on the input sentence as shown in the example.

he	is	playing	in	my	house
he_PP	is_VBZ	play_VBG	in_IN	my_PP$	house_NN

PP-Personal pronoun
VBZ - Verb, 3rd person singular present
VBG - Verb, gerund or present participle
IN - Preposition or subordinating conjunction
PP$ - Possessive pronoun
NN - Noun, singular or mass

The output of the tagger is later modified using the conversion Table 1. This was done in order to explicitly mention the auxiliary verbs that play an important role in Indian languages.

he_PP is_auxv play_VBG in_IN my_PP$ house_NN

words to be re-tagged	to auxiliary verb
does, is, has	VBZ to auxv
do, are, have, am	VBP to auxv
did, was, had, were	VBD to auxv

Table1: Conversion table

Stage2:

The translation of the example is shown in Figure 3 for twelve Indian languages. It can be observed that the arrangement of words for all Indian languages is the same. They only differ in the way the auxiliary /modal verbs and the prepositions attach themselves to the verbs and nouns respectively.

The freedom of word order here can be applied to "he" and the Prepositional Phrase, "in my house". Hence, we consider the words after the preposition including the preposition to form the first part/phrase, P1, and the rest of the sentence to form the next part/phrase, P2. It can also be observed that the preposition is placed at the end of P1 (post-position). Due to the free order of these parts, we can also have P1 as “he is play” and P2 as “in my house”, but for the sake of simplicity we stick to the case shown in the example below (as the goal of this paper is to get one good enough translation of the input).

Example:
English:
he is playing *in my house*

Equivalent in Indian Languages:

Kannada:
nanna maneyalli avanu aad’uttiddaane
word order: my house in he playing is

Tamil:
Enooda viddil avan vil’aiyaadikkonddirukkir’aan
word order: my house in he playing is

Telugu:
naa int'ilu vaad'u aad'ukun't'unaad’u
word order: my house in he playing is

Malayalam
ent*e viitt'il avan kal'ichchukondt'irikkukayaand-~
word order: my house in he playing is

Oriya:
aama ghare se kheluchii
word order: my house in he playing is

Hindi:
mere ghar mein vah khel rahaa hai
word order: my house in he playing is

Gujarati:
maaraa gharma ii ramechhe
word order: my house in he playing is

Punjabi:
Sadda ghar oo khelrayaasi
word order: my house in he playing is

Sanskrit:
mamah gruhe sahaa kriidati
word order: my house in he playing is

Bengali:
amaar baditaake chele khelche
word order: my house in he playing is

Assamese:
mur ghorat he kheliise
word order: my house in he playing is

Marathi:
maaza gharat to khedat aahei
word order: my house in he playing is

Figure3: Translation in 12 Indian languages

Original sentence:
he is playing in my house.

P1: *in* my house
P2: he is play

Reordering the preposition:

P1: my house *in*
P2: he is play

For cases with two or more prepositions in a sentence, the splitting of the sentence and the placement of prepositions are done as shown below,

Special case:
I am going *to school* *with Mary*
Mary jote *shaalege* naanu hooguttidene
P1: Mary *with* (Mary *jote*)
P2: school *to* (shaale *ge*)
P3: I am go (naanu hooguttidene)

Sentences with conjunctions are split as shown below (E1,E2…). The splitting is not applied if the conjunction conjoins nouns.
The coordinating conjunction AND connecting two PP’s is shown below:

Special case:
She is free to develop her ideas and to distribute it
E1: She is free
E2: develop her ideas *to*
E3: *and*
E3: distribute it *to*

Stage3:

If a verb is present in any of the parts, (here, P1, P2) then, place the verb at the end of that part.

P1: my house in
P2: he is *play* (*play* is placed at the end of P2)

If a verb and a particle are present together in any of the parts, place the verb along with the particle at the end of that part.

Stage4:

If an auxiliary/modal verb is present in any of the parts, then, place the auxiliary/modal verb at the end of each part.

P1: my house in
P2: he play *is* (*is* is placed at the end of P2)

For sentences with negation (with: “not”), Perfect forms (with: have, has, will have), Perfect Progressive forms (with: have/has/had been, will have been) the “not”, “be” and “have” etc. occurring after auxiliary/modal verbs are also placed at the end:

Special case:
He is not playing in my house
E1: my house in
E2: he play *is not*

Stage5:

Join the parts obtained from Stage4 and translate word by word. The English-Kannada dictionary is prepared in the following format:
english-word category kannada-word.

in **IN** alli
in **RB** o’lage
to **IN** ge (if “to” is followed by a noun)
to **IN** alu
put on **V-P** haakiko
(where, **RB**- adverb, **V-P**-verb particle)

Join the parts obtained from Stage 4, and translate word to word.

my	house	in	he	play	is
nanna	mane	alli	avanu	aad’u	ide

Actual Kannada translation:

nanna mane alli avanu aad’uttiddaane

The “to (IN)” in the sentences is translated based on the category of the word following it. If a noun follows a “to”, then the “to” is translated as “ge” or “kke” (which depends on the noun being used) or “lu”(if a verb follows the preposition).

In Indian Languages, there are a few words that are joined together while speaking and writing. This process of joining two words is called as Sandhi.

mane(house) + alli(in) = maneyalli
mane(house) + inda(from) = maneyinda

Before applying sandhi:

nanna mane alli avanu aad’uttiddaane *(form1)*
nanna maneyalli avanu aad’uttiddaane *(form2)*

However, the sentence before applying *Sandhi*, still conveys the same meaning and is grammatically correct. Any form of the sentence, *form1* or *form2* can be used. So, Sandhi is not applied in this case (Noun+Preposition).

Special care needs to be taken with sentences containing particles. The verb and the particle have to be translated together. If translated separately, the translated sentence gives a different meaning.

Special case:
Put on your shoes.

If particles are not taken care of,
your shoes on put

ninna chappali meile id’u
(meaning: put your shoes on top)

The dictionary should also contain the translation of these verb-particles together. The entry of such verb-particles is as shown in Stage5.

In Indian Languages, especially in Dravidian Languages, the tense and gender present in a sentence inflect the verbs as shown in Figure 5. In the first example in Figure 5, the verb “plays” generates the word, “aad’uttaane”, where “aad’u” is the translation of the verb “play”. The Sandhi and inflections caused due to the tense are shown in italics, “*ttaane*”. The inflections caused due to the gender are shown in italics+bold, “***ne***”. When the gender refers to a “male”, “***ne***” is used, and when it refers to a “female”, “***l’e***” is used.

English Sentence	Kannada Sentence
He plays in my house	nanna maneyalli avanu aad’u*ttaa****ne***
She plays in my house	nanna maneyalli aval’u aad’u*ttaa****l’e***
He is playing in my house	Nanna maneyalli avanu aad’u*ttiddaa****ne***
She is playing in my house	nanna maneyalli aval’u aad’u*ttiddaa****l’e***

Figure 4: Effect of tense and gender in Kannada

From the output of Stage 4:
P1: my house in
P2: he play *is*

In the above example, “play is” should be translated based on the pronouns (“he”, “she”, “it”…). The auxiliary/modal verb (“is”) gets attached to the verb (“play”). Hence the output from Stage4 requires some modification based on the gender and auxiliary/modal verbs. To solve this problem, the following is stored in a file,

he_is_verb	ttiddaane
she_is_verb	ttiddaal’e
he_is_adj	yavanu
she_is_adj	yaval’u
*_is_not_verb	ttilla
they_is_verb	ttidaare
he_will_verb	ttaane
she_will_verb	ttaal’e
they_will_verb	ttaare
he_will_be_verb	ttirutaane

*independent of gender

Figure 5: A part of the [auxiliary/modal]-[verb]-translation list

Input: he is playing in my house (a)

nanna maneyalli avanu aad'u *ide* (b)

The translation of any auxiliary verb, modal verb and "not" present in the sentence are removed from (b) to get (c). In the example (a), "he_is_playing" matches with the sequence, "he_is_verb" in the list shown in Figure 5.

nanna maneyalli avanu aad'u (c)

Finally the auxiliary verb form in kannada, "ttiddaane" which matches with, "he_is_verb", from Figure 6, is attached to the end of the sentence to get (d).

nanna maneyalli avanu aad'u*ttidaane* (d)

3 Experiments

A set of 100 sentences was picked up randomly from newspapers. The length (number of words) of these sentences varies from 3 to 13, giving an average length of 7 words per sentence. As the aim of the paper is to show how linguistic rules help in translation, we have assumed that all words that appear in the testing corpus are available in the dictionary. These sentences were translated by the system (proposed method). For evaluation purpose, these sentences were also translated by a human.

To check the quality of this method used for translation, the BLEU score evaluation technique (Kishore Papineni et. al, 2002) was applied. The BLEU score is given by,

$$BLEU=BP.\ \exp\left(\textstyle\sum_n w_n \log p_n\right) \qquad (i)$$

Where,

BP is the brevity penalty factor, given by,

$$BP = \begin{cases} 1, & \text{if } c>r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases}$$

w_n=positive weights = 1/N,

p_n=modified n gram precisions,

c = length of the translation obtained from the system.

r = length of the correct translation, translated by a human.

Applying log to (i),

$$\log_e BLEU=\min(1-r/c,0) + \textstyle\sum_n w_n \log p_n$$

Normally, N is taken as 4. As the length of the sentences taken for this system varies from 3 to 13, applying the BLEU score with N=4 for sentences of length 3, gives a score of 0 even if the translation is perfectly fine. Hence, N=3 was selected for this system.

The BLEU score is harsh for Indian Languages that stem from Sanskrit and the Dravidian family, mainly because of the fact that these languages are rich in Sandhis. This can be observed in the following example:

Input sentence:

He is going home for lunch

Output sentence from the system:

Ootakke avanu mane hooguttidaane

Listed translation:

Ootakke avanu manege hooguttidaane

BLEU score for output (N=3) = 0

Although the translation implies its meaning, the BLEU score returns a score of 0 (absence of trigrams). This is mainly due to the absence of "to" before "home", which is ignored in english. To give a fair score, the words with sandhis were split before evaluating as shown below,

The correct translation, translated by a human was split as shown:

Oota kke avanu mane ge hoogu ttidaane

The output from the system was split as,

Oota kke avanu mane hoogu ttidaane

The BLEU score with the above sentences gives 0.6237. The score for a test set of a hundred sentences was found to be 0.7164. Out of the hundred sentences, 18 sentences entered the G-EBMT module and the rest, 72 sentences entered the second module.

Experiments on effect of the corpora size (for G-EBMT) on the score were also performed. The results obtained with different sizes of the corpora are shown in Table 2.

It can be seen that there is substantial improvement in the average score when Module1 (G-EBMT) and Module2 (language specific-linguistic rules) are combined. The errors were mainly due to ambiguities in the meaning of the words. A small contribution to the error was also made by the ambiguity that exists between Prepositions and Particles.

No. of sentence rules	Average BLEU Score
0	0.6646
1,000	0.6995
5,000	0.7056
18,000	0.7164

Table2: Effect of number of rules on accuracy

4 Conclusion and Future work

A method for translating Indian Languages with limited resources has been proposed. The method uses only three rules (stage2, stage3 and stage4) giving a BLEU score of 0.7164 (on an average). However, the algorithm requires linguistic expertise. The linguistic rules are common to all Indian languages as the structure of all these languages remains the same. Thus, the method described here can be applied to all Indian Languages.

There are certain other issues to be dealt with; more concentration needs to be given to the case markings on the subject and object of a sentence, especially with dative cases appearing on the subject. This requires some way of identifying the grammatical relations. As the translation was done from English, identifying the subject as the first NP under S and object as the NP under VP (by Chomsky) helps a lot, but to determine the different cases to be used in different situations requires attention.

For future work, we would like to increase the number of generalized rules present in the G-EBMT system and analyze the effect of increase in sentence rules on the BLEU score. We would also like to build tools to automatically generate the linguistic rules. Since a major contribution to the errors was due to word sense ambiguity, we will also work on resolving these ambiguities.

Acknowledgement

We would like to thank Lori Levin (at Carnegie Mellon University) for her help in bringing the paper to its final form.

References

Mohanan, K. P. 1982a. "Grammatical relations in Malayalam". *In J. Bresnan(Ed.), The Mental Representation of Grammatical Relations, 504-589. Cambridge, MA: MIT Press.*

Ganapathiraju Madhavi, Balakrishnan Mini, Balakrishnan N, Reddy Raj, "Om: One tool for many (Indian) languages", *Journal of Zhejiang University SCIENCE*, Vol 6A, No. 11, pp 1348-1353, Oct 2005.

Helmut Schmid, "Probabilistic Part-of-Speech Tagging using Decision Trees", *Proceedings of International Conference on New Methods in Language Processing*, September 1994.

"Indian Languages" *Microsoft Encarta Online Encyclopedia,*1997, http://uk.encarta.msn.com

Kishore Papineni, Salim Roukas, Todd ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", *Proceedings of 40th Annual Meeting of the association for Computational Linguistics (ACL)*, pp.311-318, July 2002.

Ralf D. Brown, "Adding Linguistic knowledge to a Lexical Example-Based Translation System", *Proceedings of the Eighth International Conference on Theoretical and Methodical Issues in Machine Translation*, p.22-32, August 1999.

Robert Frederking and Sergei Nirenburg, "Three Heads are Better than One", *Proceedings of the Fourth Conference on Applied Natural Language Processing, ANLP-94*, Stuttgart, Germany, 1994.

Appendix A. Summary of the Stages in Module 2

Input sentence:

he is playing in my house

Output of Stage1:

he_PP is_auxv play_VBG in_IN my_PP$ house_NN

Output of Stage2:

P1: my house *in*
P2: he is play

Output of Stage3:

P1: my house in
P2: he is *play*

Output of Stage4:

P1: my house in
P2: he play *is*

Output of Stage5:

nanna mane alli avanu aad'u*ttidaane*