Implementation of Enhanced Parts-of-Speech Based Rules for English to Telugu Machine Translation

A. P. Siva Kumar *sivakumar.ap@gmail.com*
Assistant Professor,
Department of Computer Science and Engineering
JNTUA College of Engineering, Anantapur-516390, India.

Dr. P. Premchand *p.premchand@uceou.edu*
Professor,Department of Computer Science and Engineering
Osmania University,Hyderabad, India.

Dr. A. Govardhan *govardhan_cse@yahoo.co.in*
Principal & Professor,
Department of Computer Science and Engineering
JNTUH College of Engineering,Nachupalli, India.

Abstract

Words of a sentence will not follow same ordering in different languages. This paper proposes certain Parts-of-Speech (POS) based rules for reordering the given English sentence to get translation in Telugu. The added rules for adverbs, exceptional conjunctions in addition to improved handling of inflections enable the system to achieve more accurate translation. The proposed rules along with existing system gave a score of 0.6190 with BLEU evaluation metric while translating sentences from English to Telugu. This paper deals with simple form of sentences in a better way.

Keywords: POS-based Reordering, English to Telugu CLIR, BLEU

1. INTRODUCTION

Information Retrieval (IR) refers to the extraction of required information with a user query (formal statement of information need) written in one language (source language), from a large repository of documents that may be written in the same or some other language (target language). Getting only relevant data from the existing literature is made easy and faster by IR systems. The ever increasing requirement for multi-lingual information access along with the lack of technical support for multi-lingual processing bring about a new branch in research of Information Retrieval named Cross Language Information Retrieval (CLIR). It makes use of user queries written in one language to retrieve the relevant documents written in some other language. For example, a user may pose their query in English but retrieve relevant documents written in French.

English (source language) is a Subject-Verb-Object patterned language whereas Telugu (target language) is a Subject-Object-Verb patterned language that is the order of words with different parts-of-speech (POS) is not same in source and target languages. So, when a sentence is translated from source language to target language using word to word translation, the meaning of the sentence might be lost. This problem can be solved by reordering the words in the sentence based on some POS based rules.
POS tagger tool is used to identify the parts-of-speech of each word in the sentence. Then certain rules proposed in this paper, can be applied on the source sentence followed by word to word dictionary based translation. Gender based inflections are also handled. The added features enhance the quality of translated sentence by giving more accurate meaning.

The paper is organized as follows. Section 2 outlines the previous work on the translation by various organizations. Section 3 explains about the proposed system in detail. Section 4 contains the experimental results obtained by using this system and Section 5 concludes the paper.

2. PREVIOUS WORK

CLIR for Indian languages is undergoing considerable amount of research in various universities herein like Indian Institute of India (IIT), Bombay; National Centre or Software Technology (NCST) Mumbai (now, Centre for Development of Advanced Computing (CDAC), Bombay; International Institute of Information Technology (IIIT), Hyderabad. There are many machine translator systems still under production in India such as Anusaaraka project being done by IIIT, Hyderabad; Mantra (MAchine assisted TRAnslation tool) that converts English text into Hindi in a precise domain of personal administration, office orders, etc.; AnglaBharti project that is based on Pseudo Lingua for Indian Languages (PLIL). Reference [2] proposes several linguistic rules that could be incorporated in Generalized Example Based Machine Translation (G-EBMT) system for translation of English to any of the Indian languages like Telugu, Kannada, Malayalam and Tamil. The concept of word reordering of the source language sentence based on parts-of-speech tags is used also in Reference [4] for the languages Spanish, German and English.

The existing system uses generalized example based machine translation along with some linguistic rules that guide reordering of words present in a source language sentence. The dictionary based word to word translation will be the next step after reordering to achieve desired target language sentence.

3. PROPOSED SYSTEM

The design of the proposed system is an extension to the existing systems for reordering. Various stages are followed while translating a sentence from source language to target language. In each stage various reordering rules are applied to get a target sentence with correct meaning.

This system reorders the given sentence by first dividing it into words and attaching tags by using the POS tagger mentioned in [11]. Then the rules mentioned below will be applied to reorder the sentence.

3.1 Existing Rules

3.1.1 Verb Rule

This rule deals with the sentences consisting of a verb. If verb is present in the sentence, it should be moved to the end.

Consider “I eat mango” (English). This will be reordered as “I mango eat” as “eat” is a verb. Its translation will be “nenu maamidipandu tintaanu” (Telugu).

3.1.2 Conjunction Rule

It can handle sentences with one conjunction which may be present at the beginning or in the middle of the sentence. The parts of the sentence before and after the conjunction are treated as separate phrases which are translated separately and joined at the end in the same order. Consider “I studied well but the results are poor” (English). Here “I studied well” and “the results are poor” are considered as two phrases separated by the conjunction “but”. So, the two phrases are translated separately and joined at the end as “Nenu baaga chadivaanu kani manchi phalitalu raledu” (Telugu).

3.2 Proposed Rules

3.2.1 Proper Noun Rule

This rule deals with proper noun that refers to name of a company, organization, institute, person etc. which cannot be translated. In such case we use transliteration directly.

Consider ramu, john, jntu, IBM etc. Here, these words will be transliterated as they cannot be translated using dictionary. Other types of nouns (e.g. cow, chair, banana) can be translated directly using dictionary.

3.2.2 Adverb Rule
Sentences having verb and adverb should be reordered in such a way that verb is placed at the end of phrase immediately preceded by adverb.

Consider “He walks faster than Rajesh” (English). Here “walks” is verb and “faster” is adverb. Its translation will be “Atadu Rajesh kanna tvaraga nadustadu” (Telugu). Here “nadustadu” (verb) is placed at the end immediately preceded by “tvaraga” (adverb).

3.2.3 Dative Rule
This rule deals with a noun or pronoun when it is the indirect object (refers to the person or thing that an action is done to or for) of a verb. Indirect object is appended with either “ku” or “kosam” accordingly while translation.

Consider “He gave her a gift” (in English). This should be translated as “Ameku athadu oka bahumanam ichadu”. Here “her” is an indirect object. When word to word translation is performed, “her” is translated to “ame”. But, it does not give correct meaning. So, by applying this rule, we get translation as “ameku”.

3.2.4 Conjunction Exception Rule
This handles exceptional cases of conjunction rule. It says that the phrases of a sentence having conjunctions like “if”, “though” and “although” should be swapped as they will take different ordering in English and Telugu.

Consider “You will pass the exam if you study well” (English). Here the phrases are “you will pass the exam” and “you study well” should be swapped and translated as “nuvvu baaga chadivithe nuvvu pareekshalu paasavuthaavu” (Telugu).

3.3 Stages of Translation
The above mentioned rules for translation can be performed by applying them in a specific order as explained below (as shown in Figure1)

3.3.1 Stage 1
Initially, a POS tagger tool is used to associate each word in the sentence with the corresponding parts-of-speech tags. Based on the tag linked with each word the reordering is performed. For much better translation a better tagger can be used.

For example, “Rajesh walks fast but he failed in the competition.” is tagged by the POS tagger as:
Rajesh_NNP walks_VBZ fast_RB but_CC he_PRP failed_VBD in_IN the_DT competition_NN.
Here, NNP-Singular or mass noun,
VBZ - verb, 3rd. singular present,
RB-Adverb,
CC- Coordinating Conjunction,
PRP- singular nominative pronoun,
VBD-past tense verb,
IN - Preposition or subordinating conjunction,
DT- singular determiner/quantifier and
NN - Noun, singular or mass

3.3.2 Stage 2
In this stage, the presence of conjunction is checked. If it is not present then the flow is directly transferred to stage3. Else, the conjunction rule is applied. The exception with the conjunctions is

also handled in this stage. If the exception case occurs with the conjunction, the sentence is reordered accordingly by applying conjunction exception rule.
For example, “Rajesh walks fast but he failed in the competition”. Here firstly the presence of conjunction is checked. The conjunction “but” is present, so the sentence is divided into three phrases

p1: Rajesh walks fast
p2: but
p3: he failed in the competition

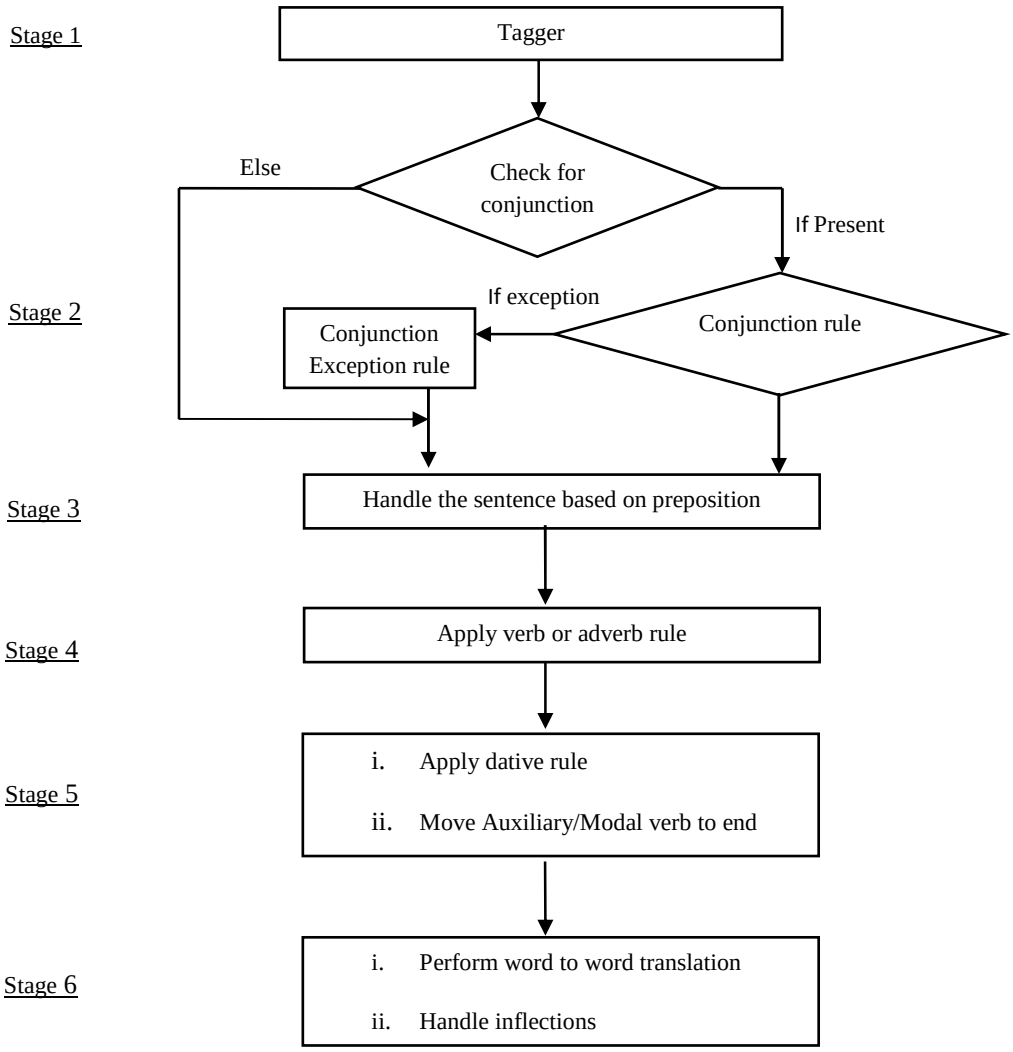

FIGURE 1: Flow of operations in translation from source language to target language.

All the reordering rules are applied separately for the two phrases (p1 and p3). While checking the presence of conjunction, it is also verified that whether it is an exceptional conjunction or not. If so, it is handled separately by swapping the phrases before and after the conjunction. Consider ‘You will pass the exam if you study well’. The sentence contains the ‘if’ exceptional conjunction. So the phrases should be reordered as shown in Table 1.

Sentence	Phrases
You will pass the exam if you study well (Original sentence)	p1: You will pass the exam p2: if p3: you study well
You study well if you will pass the exam (After Reordering)	p1: you study well p2: if p3: you will pass the exam

TABLE 1: Sentence with and without conjunction rule

3.3.3 Stage 3

Here, the sentence is split based on the preposition present in it. Then the phrases before and after the preposition are swapped.
p1: Rajesh walks fast
p2: but
p3: he failed in the competition

For the above example, the preposition is present only in the p3 phrase. So p3 should be split as p3 and p4. After reordering the phrases are as follows:
p1: Rajesh walks fast
p2: but
p3: the competition in
p4: he failed

3.3.4 Stage 4

In this stage, the presence of verb or the combination of adverb and verb is checked and verb rule or adverb rule are applied accordingly. For the above example, p1 has the combination of verb and adverb and hence they are reordered as
p1: Rajesh fast walks
p2: but
p3: the competition in
p4: he failed

3.3.5 Stage 5

Here, the dative cases are checked and if present, dative rule is applied. And also in this stage, the auxiliary/modal verbs are identified. If an auxiliary/modal verb is present in any of the parts, it will be placed at the end of that phrase.

Consider an example “he is playing games”. After crossing the above stages the sentence will be “he is games playing”. Here “is” is an auxiliary verb, thus it should be moved to the end of the sentence as “he games playing is”.

3.3.6 Stage 6

After crossing all the above 5 stages the word to word translation is performed by using bilingual English to Telugu dictionary. Then Proper noun rule is applied for the words not found in dictionary. This stage also handles the inflections that are different forms of a verb based on the gender after translation into target language.

For the above example “Ramesh” is not found in the dictionary so the proper noun rule is applied and the translated phrases will be
p1: Ramesh veganga nadu
p2: kani
p3: poti lo
p4: athadu viphalam ayyenu

Also the inflections present in the sentence will be handled as given in [2]. Thus at the end, combining all the phrases with the inflection rule we get the translated sentence as:
'Ramesh veganga nadustadu kani poti lo athadu viphalam ayyenu'
In this way by following all the six stages an English sentence can be translated to Telugu appropriately giving a better quality translation.

4. EXPERIMENT

For the evaluation of the proposed system we have selected 100 simple English sentences from the daily newspaper in which the count of words varies from 3 to 12. For translation purpose, we have used a bilingual dictionary containing all the words used in testing corpus. To perform evaluation technique the sentences are translated by the proposed system and also by a human.

Quality can be treated as the agreement between the machine translation and the human translation. The system is said to be good if its translation is very close to that of the human translation. To determine this quality of the proposed system we used BLEU (Bilingual Evaluation Understudy) score evaluation technique referred in [3]. The BLEU score is given by,

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

Here,

$$p_n = \frac{count\ of\ correct\ n\text{-}gram\ match}{count\ of\ total\ n\text{-}gram}$$

where BP is the brevity penalty factor, given by,

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

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),

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

In the proposed system the length of the sentence starts from 3. Hence we use N=3 (that is trigram model) in the system. The trigram model consists of subsequence of 3 words to form trigrams. By examining how many standard deviations each 3-gram differs from its mean occurrence, the p_n value is determined. The evaluation technique when performed on proposed system with a set of 100 sentences gave a score of 0.6190.

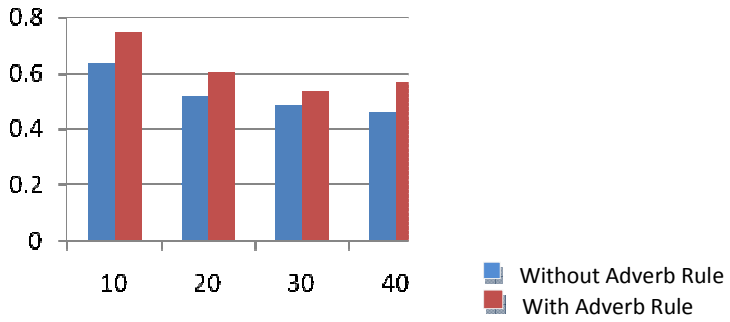

FIGURE 2: BLEU score without and with adverb rule

The values in the Table 2 can be represented as Figure 2, which shows the BLEU scores of sentences without adverb rule against with adverb rule for English to Telugu translation. The x-axis represents the number of sentences and the y-axis represents the BLEU score.

No. of sentences	Without adverb rule	With adverb rule
10	0.6364	0.7444
20	0.5161	0.6040
30	0.4830	0.5357
40	0.4637	0.5687

TABLE 2: BLEU score for without and with adverb rule

In the similar way, Figure 3 shows the BLEU scores of sentences without conjunction exception rule against with conjunction exception rule for English to Telugu translation, which are tabulated in Table 3. In this figure also the x-axis represents the number of sentences and the y-axis represents the BLEU score.

No. of sentences	Without conjunction exception rule	With conjunction exception rule
10	0.2024	0.2393
20	0.2650	0.2864
30	0.2522	0.2666
40	0.2251	0.2619

TABLE 3: BLEU score for without and with conjunction exception rule

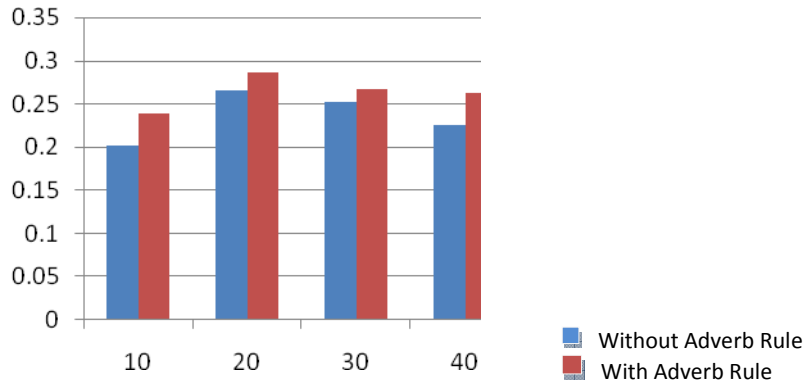

FIGURE 3: BLEU score without and with conjunction exception rule

5. CONCLUSION AND FUTURE WORK

This paper enhances POS based reordering rules that preprocess the user query for better translation in order to use it in searching relevant documents written in Telugu. The added rules enable the system to deal with adverbs and conjunctions in a better way. The proposed system gives a BLEU score of 0.6190 (on an average). The performance of the system highly depends on the POS tags attached to the given source sentence. Better the tagger, the more efficient the translation will be.

There is no perfect machine translator for Indian languages which stem from Sanskrit and Dravidian family, mainly because of the reason that they are rich in sandhis. More concentration should be given to handle this. We also would like to handle other type of sentences like interrogations and exclamations in future work.

6. REFERENCES

[1] R.Gangadharaiah & N. Balakrishnan, "*Application of Linguistic Rules to Generalized Example Based Machine Translation for Indian Languages*", Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages, India, 2006

[2] Mustafa Abusalah, John Tait & Michael Oakes, "*Literature Review of Cross Language Information Retrieval*", World Academy of Science, Engineering and Technology, 2005.

[3] P.Kishore, Salim Roukas, Todd ward & Wei-Jing Zhu, "*BLEU: a Method for Automatic Evaluation of Machine Translation*", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318, 2002.

[4] Maja Popovic & Hermann Ney, "*POS-based Word Reorderings for Statistical Machine Translation*", in Proceedings of the Fifth International conference on Language Resources and Evaluation, 2006.

[5] Anne R. Diekema, "*Translation Events in Cross-Language Information Retrieval: Lexical Ambiguity, Lexical Holes, Vocabulary Mismatch, and Correct Translation*", Dissertation at School of Information Studies, Syracuse University, 2003.

[6] Sethuramalingam S, "*Effective Query Translation Techniques for Cross-Language Information Retrieval*", MS Thesis submitted at IIIT Hyderabad, India, 2009.

[7] Sudip Naskar & Sivaji Bandyopadhyay, "*Use of Machine Translation in India: Current Status*", AAMT J., 36:25-31, 2004.

[8] Sanjay Kumar Dwivedi and Pramod Premdas Sukhdeve, "*Machine Translation System in Indian Perspectives*", Journal of Computer Science 6 (10): 1082-1087, 2010.

[9] Shu Cai, Yajuan L & Qun Liu, "*Improved Reordering Rules for Hierarchical Phrase-based Translation*", International Conference on Asian Language Processing, 2009.

[10] ZHANG Xiao-fei, HUANG He-yan & ZHANG Ke-liang, "*Cross-Language Information Retrieval Based on Weight Computation of Query Keywords Translation*", Intelligent Computing and Intelligent Systems, 2009 IEEE International Conference, 2009.

[11] Parts-Of-Speech tagger tool – http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/postagger.