Word Reordering using Rule based Parser in Telugu to English Machine Translation

[1]KanusuSrinivas Rao
Assistant Professor
Dept. of Computer Science
Yogi VemanaUniversity,Kadapa
AndhraPradesh, India.
kanususrinivas@yahoo.co.in

[2]Ratnakumarichalla
Assistant Professor
Dept. of Computer Science &Engg.
JNTUK, Kakinada
AndhraPradesh, India.
ratnamala3784@gmail.com

Abstract:- In Machine translation Word Reordering is an important intermediary stage, where the source language sentences with disordered words are oriented into the grammatical structure of target language. The main focus is imposed on the reordering of given words into the basic structure of target language using parser. A parser is employed on the POS tagged words, which uses the rule bank to orient the words into correct form. The rule bank is a collection of rules that consists of the syntax of target language. The parser with the set of rules, reorders the source language structured sentence into target language structure. The paper mainly focuses on the implementation of reordering for Telugu to English Translation where the SOV (Subject-Object-Verb) order of source language Telugu is converted into SVO (Subject-Verb-Object) order of target language English. The output of Reorder stage is presented in the format of a parse tree that forms the basis for the further post processing required.

Keywords:- *parsering, source sentence, target sentence, post processing, word reordering*

*****

1. INTRODUCTION

Machine translation is a process of translating the text in one language to another language. The process involves a series of steps such as tokenisation, tagging, translation, reordering and post processing. Word reordering is an important step and penultimate step in Machine Translation. A good word reordering mechanism produces an effective output. Word reordering can be done at source side, called pre-ordering or at target side, called post-ordering. In pre-ordering, the words are arranged in target language structure and are then translated into target language. The post-ordering employs an opposite mechanism where the words are first translated into target language and are then ordered as per the structure of target language. Both mechanisms have proven to be effective and the post-ordering is recently gaining importance.

This is an intermediary phase before post-processing where words are oriented in target structure i.e. in the basic structure of the target language. The reordered sentences can be used for training and testing. A good reordering mechanism improves the performance and efficiency of machine translation system.

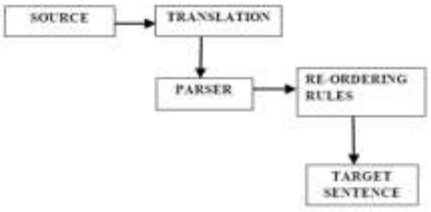


Fig1: Reordering in Machine Translation System

Reordering in Telugu to English Machine Translation System

This section provides similarity and differences between the universal language English and the Dravidian language Telugu. The following are the divergences:

- English is a highly positional language with rudimentary morphology, and default sentence structure as SVO (Subject + Verb + Object).
- Telugu is highly inflectional, with a rich morphology, and default sentence structure as SOV (Subject + Object +Verb).
- English uses prepositions while Telugu uses post-positions.
- Dravidian languages such as Telugu allow greater word order freedom.
- Telugu is relatively richer case-marking system.

In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses. Such constructions are not natural in Indian languages, and present major difficulties in producing good translations.

With the current state of art in MT, it is not possible to have Fully Automatic, High Quality, and General-Purpose Machine Translation. Practical systems need to handle ambiguity and the other complexities of natural language processing.

Therefore, to have a good translation system, reordering the source sentence in accordance to target sentence is needed. So reordering system for source sentences can make significant improvements over Indian languages.

English (target language) is a Subject-Verb-Object patterned language whereas Telugu (source 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.

2. PROPOSED SYSTEM

Initially, tagged input is given in the form of a sentence in target language (English) with the syntactic structure of source language (Telugu). This work aims at reordering the sentence into the syntactic structure of English with the help of reordering rules. It helps in reordering a wide variety of sentences in English language.

The different kinds of sentences that system can process are

- Assertive sentences
- Negative Sentences
- Imperative Sentences
- Interrogative Sentences
- Exclamatory Sentences
- Conjunctive Sentences

2.1. TAGGING

Part-of-Speechtagging (POS tagging or POST),also called grammatical tagging, is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

The different POS tags are subject noun(Subj-noun), verb(V), preposition(P), adverb(Adv), adjective(A), noun(N), determiner(Det), conjunction(Con), degree-adverbs (Deg), question tags, interjection words.

The subject noun differs from other nouns in the case that it forms the main subject or actor of the sentence and degree adverbs are the words that qualify adverbs. Determiners are a set of grammatical words that are somewhat like modifiers, but actually serve the function of specifier making more precise or definite the phrase that follows. It includes a diverse set of grammatical words: demonstratives, consisting of this, these, that, those; articles, consisting of a/an, the; wh-words, consisting of which, what, whose; possessives, consisting of possessive adjectives such as my, your, or his and possessive nouns such as John’s or Ram’s; and quantifiers such as some, any, every, each,more, or neither and numerals.

These POS tags are manually assigned to each word while giving the input. This system can be further developed to assign auto-tagging.

2.2 Tokenization

Tokenization is the process of demarcating and possibly classifying sections of a string of input characters. The resulting tokens are then passed on to some other form of processing. The process can be considered a sub-task of parsing input.

Take, for example,

The quick brown fox jumps over the lazy dog

The string isn't implicitly segmented on spaces, as an English speaker would do. The raw input, the 43 characters, must be explicitly split into the 9 tokens with a given space delimiter (i.e. matching the string““). It is identified by the ASCII character 32.

2.3 Dividing into phrases

The tokens formed after tokenization are divided into phrases basing on their POS tags. Phrase is a syntactic unit headed by a lexical category. Phrases are defined as sequence of words or a single word, having syntactic significance.

The following rule will account for declarative sentences, but not for imperatives, which have no subject.

S → NP + VP

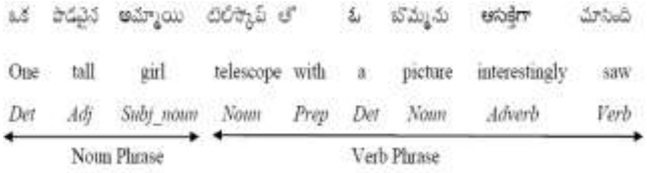
3. PHRASES

The main phrases are Noun Phrase (NP), Verb Phrase (VP), and Prepositional Phrase (PP). The sub phrases include Adverb Phrase (AdvP) and Adjective Phrase (AP). The main component of the phrase can be understood by the name itself.

3.1. Noun Phrase

Initially, a search is carried out to identify the Subject Noun in the given input and basing on it, the phrases are further divided. If it is found, all the words till subject noun are considered as Noun Phrase and the remaining words are included in Verb Phrase. If the subject noun is not found, then the sentence is considered to be in imperative mode and is handled in a different way.

For example, the following sentence is translated into English by word to word mapping and are then basically divided into 2 phrases basing on the POS tags i.e. all the words upto subject noun are taken into NP and remaining as VP. The POS tags are represented in italic letters in the below example.


ఒక పొడవైన అమ్మాయి టెలిస్కోప్ తో ఓ బొమ్మను ఆసక్తిగా చూసింది
One tall girl telescope with a picture interestingly saw
Det Adj Subj_noun Noun Prep Det Noun Adverb Verb
Noun Phrase
Verb Phrase

Now these phrases are individually dealt to be reordered. The words in Noun Phrase are ordered as per the following rules.

NP → (Det) NN

NN → (AP) NN (PP) | N

AP → (AdvP) A

AdvP→ (Deg) Adv

PP → P (NP)

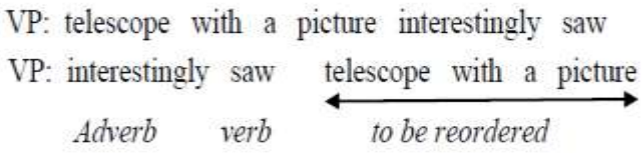
The parenthesis ( ) indicate the optional choices i.e. they may or may not present. It implies the noun phrase to be reordered such that, the subject noun is to be preceeded by an Adjective phrase, which is preceeded by a determiner. The adjective phrase consists of an adjective which is preceeded by an Adverb phrase. The adverb phrase, in turn, consists of adverb preceeded by a degree adverb.

3.2. Verb Phrase

The Verb Phrase, not only consists of the verb, but also followed by other phrases such as noun phrase or a prepositional phrase. While reordering the verb phrase, the verb is placed such that it is preceeded by an adverb phrase (if any) and followed by either NP or PP. It can be understood by the following rule.

VP → (AdvP) V (NP) | (AdvP) V (PP)

In the above example, the remaining words which are taken into VP are reordered in the following manner.


VP: telescope with a picture interestingly saw
VP: interestingly saw telescope with a picture
Adverb
verb
to be reordered

The remaining noun phrase also consists of a prepositional phrase which is to be reordered.

3.2.1. Prepositional Phrase

The main component is preposition and a preposition phrase always consists of a Noun Phrase followed by the preposition. It can be specified by the following syntax.

PP → P (NP)

To reorder prepositional phrase, we apply preposition rule. It states that a preposition and the preceding noun phrase in the input is to be swapped. After that the noun phrases are added in the reverse order to the output to form the syntactical structure of English.

PP: telescope with → with telescope

NP: with telescope a picture

NP: a picture with telescope

The reordered sentence is obtained by combining the individual phrases. The final output is the basic syntactic structure of the target language. The final reordered sentence for the above example is as below

"One tall girl interestingly saw a picture with telescope"

This can be syntactically represented by the following parse tree

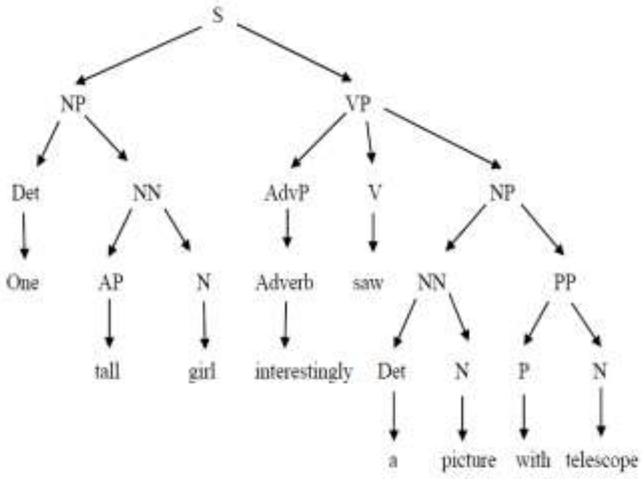
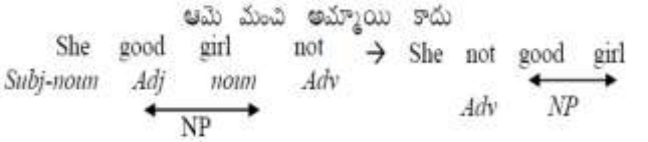


Fig 2: Parse tree generated for a given example

3.3 Handling Negative Sentences

A negative sentence is identified by the word 'Not'. It usually occurs at the end of sentence in Telugu language. Here the word before 'Not' is to be identified. If it is a noun (N), then Not is placed at the starting of that particular Noun Phrase. If the preceeding word is a verb (V), then Not is placed before the verb phrase (VP).

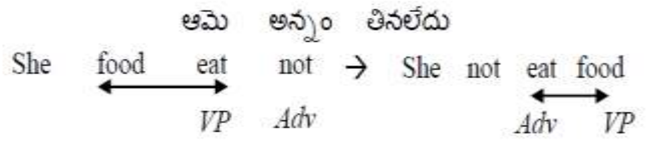
Example1:



Here, as ‘Not’ is after a NP, it switched to a place before the corresponding NP. The output “She good girl not” does not serves as the final output to machine translation It needs POST-PROCESSING. The output produced at the end of reordering serves as input to post-processing. In this phase, the auxiliary verb ‘is’ will be added to make the sentence grammatically correct with respect to the target language.

After post-processing, the reordered sentence will be modified as “She is not good girl”. In this paper, we are only focusing on the reordering phase.

Example 2:



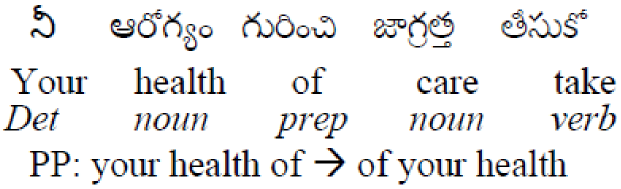
Similarly, as not is placed after a verb, it will be moved to the beginning of that Verb phrase after reordering it.

3.4 Handling Imperative Sentences

These sentences don’t have any subject. They usually comprise a command, advice or request. The verb forms the soul of the sentence. When a subject noun is not found in the given input, the sentence is considered to be in imperative mood.

These sentences usually start with Verb Phrase which in turn consists of NP and PP and is handled as mentioned above as in declarative sentences. The only difference is that it starts with noun phrase.

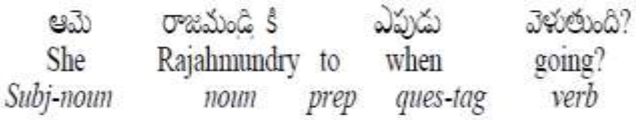
For instance,


నీ ఆరోగ్యం గురించి జాగ్రత్త తీసుకో
Your health of care take
Det noun prep noun verb
PP: your health of → of your health

As there is no noun subject, it outputs: “take care of your health”

3.5. Handling Interrogative Sentences

These sentences use the questions tags such as which, when, what, where, why, who and how. The question tag usually comes at the start of the sentences in English. Initially, after giving the input, the system searches for any question tags as specified in POS tags and sets them aside. After the remaining words are reordered as per the declarative or imperative sentences, the question tag is placed at start of the sentence. This works well with WHO, WHEN, WHERE, WHY but an additional rule is to be followed for accommodating which, what and how.


ఆమె రాజమండ్రి కి ఎప్పుడు వెళుతుంది?
She Rajahmundry to when going?
Subj-noun noun prep ques-tag verb

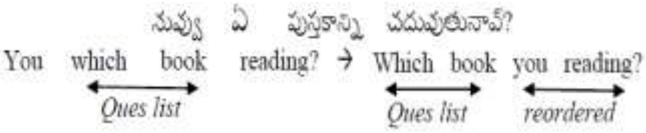
Here, the question tag will be set aside and the remaining words: “She Rajahmundry to going” will be reordered to yield: “she going to Rajahmundry” as per the rules mentioned above. Now the question tag is placed at the start of the sentence which yields the final output as:

When she going to Rajahmundry?

Thus the basic structure of target language will be formed and can be sent to the next stage of machine translation to include auxiliary verbs as well. This serves as an input to the next stage.

If the question tag is WHICH or WHAT, then the next word in the input is checked. If the next word is a noun, then the question tag along with this noun are to be set aside. The remaining word are reordered as usually and then both the question tag and noun are placed before this reordered sentence.

For instance,


నువ్వు ఏ పుస్తకాన్ని చదువుతున్నావ్?
You which book reading? → Which book you reading?
Ques list
Ques list reordered

Here the question tag which will be searched first, and then the next word if found to be a noun, are added to the Question List. The remaining words will be ordered as per the rules. Now, the question list will be added to the beginning of the sentence.

The same method is followed for the question tag HOW, except for the change that the next word is checked whether an adjective or not. If found to be an adjective, it is added to the question list and placed at start of the sentence same as the above case.

3.6. Handling Exclamatory Sentences

To handle exclamatory sentences, initially we identify the interjection words and then the remaining words are employed to be reordered. Then the interjection word is placed at the beginning.

It is somewhat similar to the method implemented in interrogative sentences. Here, instead of the question tags, we consider the interjections.

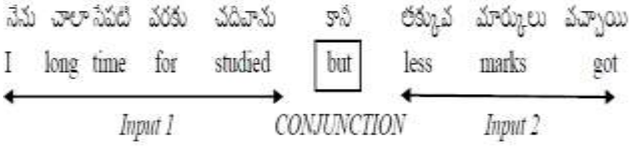
3.7. Handling Conjunctions

Conjunction is a word that joins together words, phrases, or sentences. Thus, a conjunctive sentence is the combination of two different sentences which can be dealt separately. When an input is given, the presence of a Conjunction is checked. If it is found, each phrase separated by the conjunction is treated as a separate input to the parser and after reordering, they are again combined by means of a conjunction.

To combine the reordered phrases, the conjunction rule and the conjunction-exception rule are to be employed. Generally, the conjunction rule is applied. If the exception case occurs with the conjunction, the sentence is reordered accordingly by applying the conjunction exception rule.

3.7. 1 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 processed separately and joined at the end in the same order.
Consider,


నేను చాలా సేపటి వరకు చదివాను కాని తక్కువ మార్కులు వచ్చాయి
I long time for studied but less marks got
Input 1
CONJUNCTION
Input 2

In the above example, input 1 and input 2 are processed separately and they yield the following intermediary result.
Input 1: I studied for long time (declarative sentence)
Input 2: got less marks (imperative sentence)

Now, these two are connected in the SAME ORDER by the conjunction to give the output as

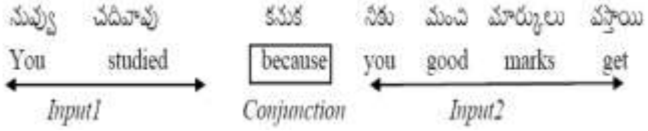
I studiedfor long time but got less marks

↓

Conjunction

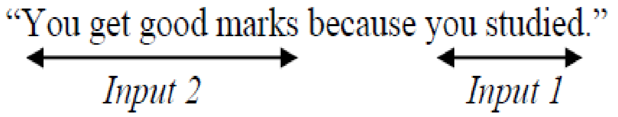
3.7.2 Conjunction Exception Rule

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 the sentence,


నువ్వు చదివావు కనుక నీకు మంచి మార్కులు వస్తాయి
You studied because you good marks get
Input1
Conjunction
Input2

In the similar manner, input1 and input2 will be separately processed and they yield the intermediary results as
Input1: You studied
Input2: You get good marks

These two phrases will be joined by the conjunction in REVERSE ORDER to give the final output as:



All the reordering rules are applied separately for the two phrases. While checking the presence of a 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.

4. CONCLUSION

The word reordering system gives efficient results as a part of Machine Translation. Reordering is done through the implementation of a rule based parser. To improve efficiency, the reordering is performed at the lowest level of phrase and these constituent phrases of a main phrase are logically combined through bottom-up approach. The output is clearly displayed in the form of parse tree so that a naive user can also understand and get the clear picture of how parsing is done. The manual tagging can be replaced with auto tagging to reduce the burden of user. It can be improved to deal with other kinds of sentences such as infinitives, comparatives etc.

REFERENCES

[1] Laurel J. Brinton, “The Structure of Modern English”

[2] 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

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

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

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

[6] Isao Goto, Masao, Utiyama, “Post-ordering by Parsing for Japanese-English StatisticalMachine”.

[7] Brown, C.P., “The Grammar of the Telugu Language”. New Delhi: Laurier Books Ltd, 2001.

[8] Gwynn and Krishnamurti: "A Grammar of Modern Telugu", volume 11, Oxford University Press, Delhi, 1987.

[9] W.John Hutchins and Halord L. Somers, "An ntroduction To Machine Translation", Academic Press Ltd.,1992.

[10] Jurafsky, Daniel and Martin, James.H, "Speech and Language Processing-An Introduction to Natural Language processing, Computational Linguistics and Speech Recognition", 2002.

[11] http://en.wikipedia.org/wiki/Google_Translate

[12] http://nlp.stanford.edu:8080/parser

Authors Biography

[1]K.Srinivasa Rao, did his M.Tech(CSE) from JNTUCE, JNTU, Hyderabad in the year 2009. He has total 7 years of experience in teaching. Currently he is working as Assistant Professor at Yogi Vemana University, Kadapa.

[2]RatnaKumariChalla, did her M.Tech (CSE) from HCU, Hyderabad in the year 2009. She has total 6 years of experience in teaching. Currently she is working as Assistant Professor at JNTUK, Kakinada. AP, India.