Volume 2, Issue 2 February 2016

ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal Of Science, Engineering and Information Technology

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi, Urdu all research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: - North Asian International Research Journals - 221 Gangoo, Pulwama, Jammu and Kashmir, India -192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

MACHINE TRANSLATION SYSTEM FOR ENGLISH TO TELUGU LANGUAGE:A RULE BASED COMPLEX SENTENCE SIMPLIFICATION

A.RAGINI[1] & DR.M.HUMERA KHANAM[2]

[1](M.Tech, Student, Dept. of Computer Science and Engineering, SVU College of Engineering, Tirupati, Andhra Pradesh, India

[2](Associate Professor, (M.Tech, PhD), Dept. Of Computer Science and Engineering, SVU College of Engineering, Tirupati, Andhra Pradesh, India.

ABSTRACT

Machine translation is the process by which computer software is used to translate a text from one natural language to another but handling complex sentences by any machine translation system is generally considered to be difficult. A rule based technique is used to translate a sentence. User gives an input, which is an English sentence. The given input sentence is then tokenized into individual words. Each single word is tagged with its respective part of speech in english, the words which are not in the pre-defined database are tagged using grammatical rules. These individual words are then concatenated to form a sentence into Telugu. To simplify the complex sentences based on connectives like relative pronouns, coordinating and subordinating conjunctions, a rule based approach is used.

***INDEX TERMS**: Machine translation, POS tag, Sentence simplification, Rule Based Approach, English to Telugu.*

INTRODUCTION

In a large multilingual society like India, there is great demand for translation of documents from one language to another language. There are 22 constitutionally approved languages, which are officially used in different states, such as Hindi, Bengali, Guajarati, Oriya, Punjabi, Telugu, Kannada, Tamil, Malayalam, etc. In India Hindi is the national language and English is the common language for all states. Even though Hindi is our national language, Hindi is spoken only in northern states but in the southern region especially in Andhra pradesh most of the people speak only in their regional language (i.e. Telugu). So, for better communication English to Indian language machine translation is necessary. The ultimate goal of research on Natural language processing (NLP) is to understand human or natural languages and to facilitate human-machine interaction through human language or natural language. To achieve such research goal, NLP people has focused on different sub tasks. Machine translation(MT) is one of such sub-task.

Machine translation is the process of translating a text from source language into a target language with the help of computers. The translation process converts a text in one human language to another which preserves not only the meaning, but also the form, effect and style. Nowadays most of the online information is available in English. In a multi-lingual society different languages are spoken in different regions. So, for this purpose machine translation is required. Hence to build a MT system, one needs to have a clear view of the rules and grammar of the source language as well as the target language. English is a rich language and in this paper only Nouns, Verbs, Prepositions, Phrases and Inflections are considered.

This paper focused process of the MT system and the performance of it. Sentence simplification and segmentation can be performed by two approaches, either rule based or corpus based. In our paper rule based technique is used for simplification.

APPROACHES TO TAGGING

There are mainly two approaches for POS tagging 1) Linguistic or Rule-based approach 2) Machine learning or stochastic approach.

1. Rule-based tagging:

There are two stages. In first stage, use a dictionary to assign all possible grammatical categories to each word. In second stage, use a large list of hand crafted rules to identify correct single tag for each ambiguous word. Disambiguation is done by analyzing the linguistic features of the word, its previous word, its following word and other aspects. For example, if the previous word is article then the next word must be noun. This information is coded in the form of rules.

2. Stochastic tagging:

Stochastic tagging techniques make use of corpus. The most common stochastic tagging uses a HMM(Hidden Markov Model). Stochastic tagging techniques can be either supervised/unsupervised/hybrid. In HMM the states usually denote the POS tags. The probabilities are estimated from a tagged training corpus or untagged training corpus in order to compute the most likely POS tags for the word of an input sentence.

EXISTING MT SYSTEMS

MT systems are classified into various categories like Rule based, example based, statistical based, hybrid based, knowledge based, principle based and online interactive based methods . Rule based and statistical based methods are the earliest methods and most widely used. These approaches were used to translate the text from English to Indian languages and vice versa.

Rule Based MT Systems:

Rule based MT systems were the first commercial MT systems that work on linguistic rules of source and target languages. These rules will help in arranging the translated words correctly based on the context of the sentence. Rules are applied during analysis phase, transfer phase and generation phase. This rule based system consists of various steps like syntax analysis, semantic analysis, morphological analysis, syntax generation and semantic generation. Rule based MT systems are less robust and gives good grammatical results if it finds an appropriate parse else it fails.

Statistical Based MT Systems:
Statistical based systems are kind of empirical MT systems which uses huge amount of information that consists of text and its translations. This approach is predicated on parallel corpora. The three key components of any statistical MT systems are language model, translation model and search algorithm.

STRUCTURE ANALYSIS OF ENGLISH AND TELUGU LANGUAGES

Comparative analysis of the sentence structures in English and Telugu languages is important for efficient translation. English sentences are of various types: complex sentence, compound sentence and simple sentence. Compound sentences is a combination of two or more sentences.
The language pattern for simple sentence in English is as follows :
Subject + Verb + Object (SVO).
For eg: Raghu plays volleyball (Raghu+ plays + volleyball).
In Telugu the pattern for simple sentence is as follows :
Subject + Object + Verb.
The Telugu translation for the above sentence is as follows
రఘు+వాలిబాల్ +ఆడతాడు (Raghu+volleyball+aadathaadu)

Examples for grammar rules:
To produce the rules for translation grammatical analysis of both the languages should be done which is similar to sentence analysis. English and Telugu languages are based on independent grammar and they need to properly mapped.
Consider the example sentence :
1.“she was writing then”
Grammar pattern for this English sentence is given as
p + v + adv
where p:pronoun; v: verb: adv:adverb.
Corresponding Telugu sentence would be:
“ఆమె + అప్పుడు + రాసింది ”

Grammar pattern for this Telugu sentence is given as
p + adv + v
2. “ we+visited+(tajmahal+last+year) ”
Pos tags are:
p + v + (n + d + n’)
Translated telugu sentence:
మెము +(తాజ్ మహల్ + గత +సంవత్సరం)+చూసాము

Telugu patterns are:
p + (n +d + n’) +v

RULE BASED SENTENCE SIMPLIFICATION

For complex sentences a rule based simplification algorithm is implemented. Based on rules, the sentences are simplified in order to get exact translation. When a clause stands on its own and is independent, it is called main clause. Subordinate clauses are those clauses which cannot stand alone but depend on main clause for their meaning. Most of the sentences contain conjunctions and sentences are split based on conjunctions. Independent clauses can be joined by a coordinating conjunction to form complex or compound sentences. Dependent clauses often begin with a subordinating conjunction or relative pronoun. Our system handles coordinating conjunctions, subordinating conjunctions and relative pronouns. Coordinating conjunction includes for, and, not, but, or, yet and so. Subordinating conjunction includes after, although, because, before, if, since, that, though, unless, where, wherever, when, whenever, whereas, while, why. Relative pronoun includes who, which, whose, whom.
The proposed approach follows in following steps:
1. Split the sentences from the paragraph based on delimiters such as “.” and “?”
2. Delimiters such as (comma, {,}, [,],) are ignored from the sentences.
3. Individual sentences are split based on coordinating and subordinating conjunction.
Splitting is used to break the sentences which contain coordinating and subordinating conjunction whereas sentence simplification is to simplify the sentences which contain relative pronoun. The simplifying sentences will work for any translation system with English as source language. Here English to Telugu language is considered.

There are several “wh” connectives available out of which “who, whom, which, whose” are dealt. In this case, the relative clause(RC) can occur either in between the main clause(MC), or after the main clause. In both the cases, the connective words contain two possible dependency tags i.e. either “subject” or “object”.

Input sentence: The people who live in Scotland are called Scots.
Splitting the above sentence as
The people live in Scotland. (RC)
The people are called Scots. (MC)
POS tag of Input sentence:
The/ DT, people/ NNS, who/ WP, live/ VBP, in/ IN, Scotland/NNP, are/ VBP, called/ VBN, Scots/ NNS.

Some more Examples for sentence simplification:
Input Sentence: The shoes which I bought yesterday are very comfortable.
Splitting the above sentence will result as shown below.
I bought the shoes yesterday(MC)
The shoes are very comfortable(RC).

Some more examples are
1.The book which is on the table belongs to Gowtham.
Splitting the above sentence will result as follows
The book belongs to Gowtham.
The book is on the table.
2. The students are studying because they have a test tomorrow.
The students are studying
because they have a test tomorrow.
3. Unless the coffee is hot I will not drink it.
Unless the coffee is hot.
I will not drink it.
the sentences are split based on the conjunctions.Coordinating conjunction includes (for, and, not, but, or, yet, so)and POS tag for coordinating conjunction is "CC" and the dependency tag is "cc". Subordinating conjunction includes (when, whenever, where, wherever, if, because, unless, though, etc.). Here, the relative clause can occur before the main clause, or after the main clause.
Consider an example,
Input Sentence: Ravi/NNP, waited/VBD, for/IN, the/DT, train/NN, but/CC, the/DT, train/NN, was/VBD, late/JJ.
In the above example relative clause is present after the mainclause. Here, "but‟ is the coordinating conjunction and it is thesplitter word. Here the sentences will be split into two simplesentences based on the splitter word. The connective word is always present in the relative clause.

Ravi waited for the train
But the train was late.

Thus the given complex sentences will be splitted based on the rules and segmented with POS tags for translation. After translation, the above two sentences are combined to get a meaningful sentence.

IMPLEMENTATION

An input is accepted from user, which is an English sentence. The given input sentence is then tokenized into individual words. These words are tagged with their respective parts of speech. All other words that are not found in the database are tagged using grammatical rules that we formulated. Using these POS tags, their respective word translations are retrieved from the database. These individual words are then concatenated to form a sentence which is the result of user's input.For this, HTML(Front-end), PHP(Middleware), MySQL(Back-end) technologies are implemented.

Sample code for database creation:

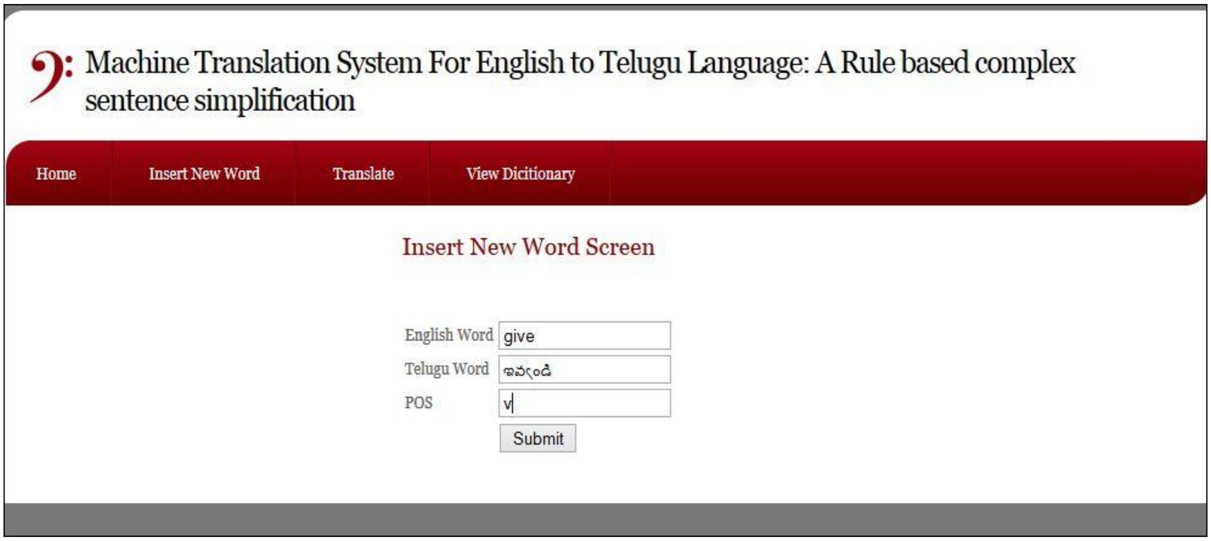
```
create table dict (eword nchar(255), tword nchar(255), pos nchar(20), past nchar(255), present nchar(255) ) engine=innodb defalut charset=utf8;
```

ALGORITHM FOR RULE BASED MACHINE TRANSLATION:

```
begin
EnglishWord[k] := Parsing(I);
l:= Sizeof(D);
for j; =0 to k do
        if token is a preposition set PREP=1
        else
                PREP=0
        End if
        If (PREP=1) compare the rule and extract meaning for prepositional phrase
//Comparing sentence with rules provided
for i:= 0 to r do
        for j:= 0 to k do
                S:=CompareRule(EnglishWord[j]);
        endfor
endfor
//finding word to word to meaning from English to Telugu
for i:= 0 to k do
        for j:= 0 to l do
        if
        (EnglishWord[i]==EnglishMeaning[j])
        then
        TeluguWord[i] =TeluguMeaning[j]);
        endif
        endfor
endfor
O:=TeluguSentenceConstruct(TeluguWord[k],S);
return O;
end
```

EXPERIMENTAL RESULTS

The first phase in machine translation system is giving text input, which is shown in Fig 2. The parts of speech tagging is shown in Fig 5. This system also allows the user to enter the new words into the dictionary shown in Fig 1.Users can have a brief glance at the dictionary to obtain meanings of simple words.

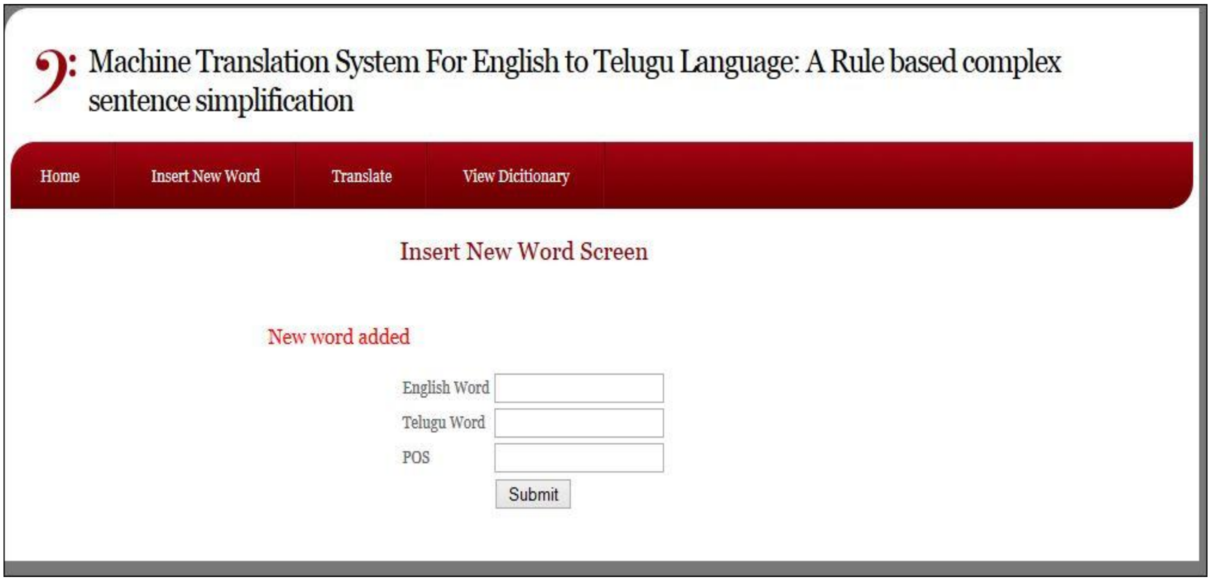


Machine Translation System For English to Telugu Language: A Rule based complex sentence simplification

Insert New Word Screen

English Word: give
Telugu Word: ఇవ్వండి
POS: v

Fig 1: Snapshot for Inserting words into dictionary



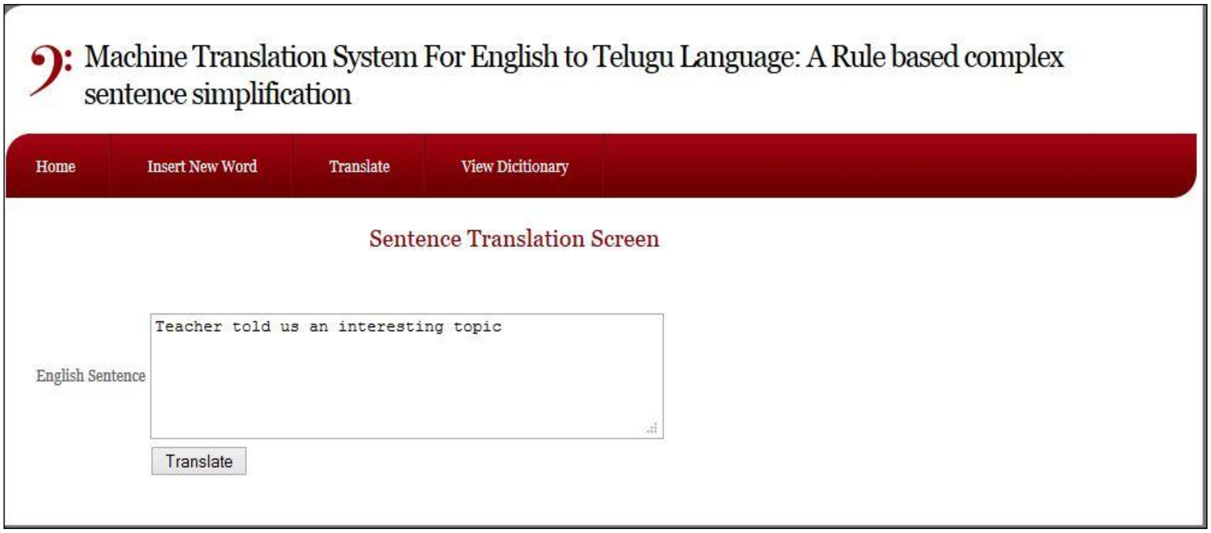
Machine Translation System For English to Telugu Language: A Rule based complex sentence simplification

Insert New Word Screen

New word added

English Word:
Telugu Word:
POS:

Fig 2:Snapshot of accepting new word

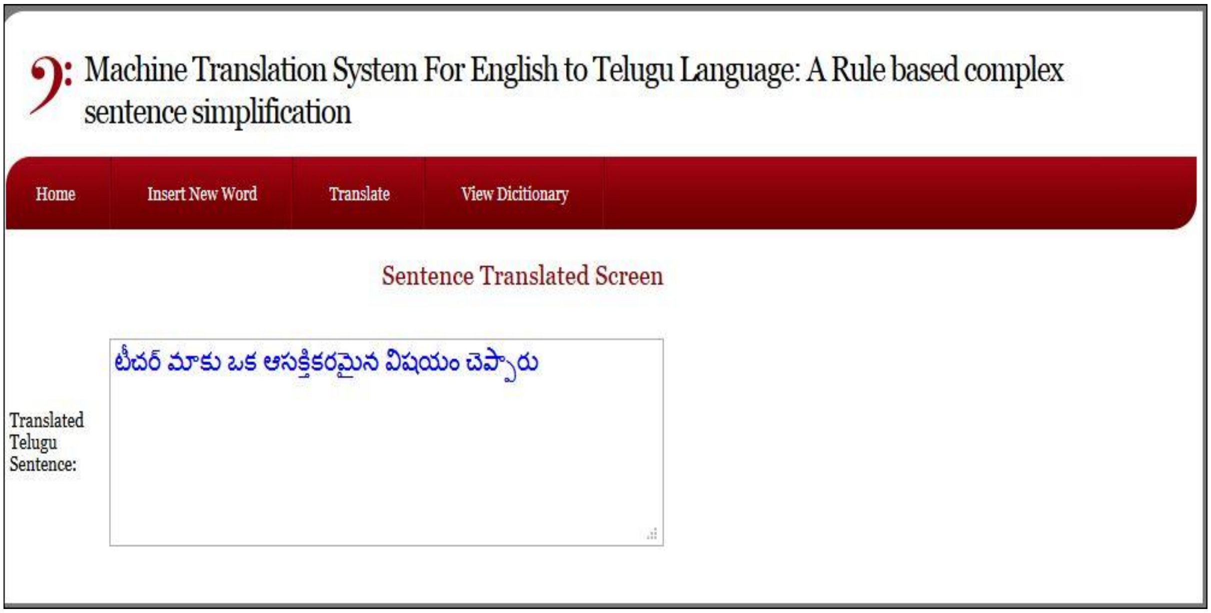


Machine Translation System For English to Telugu Language: A Rule based complex sentence simplification

Sentence Translation Screen

English Sentence: Teacher told us an interesting topic

Fig 3: Snapshot of giving an input English Sentence

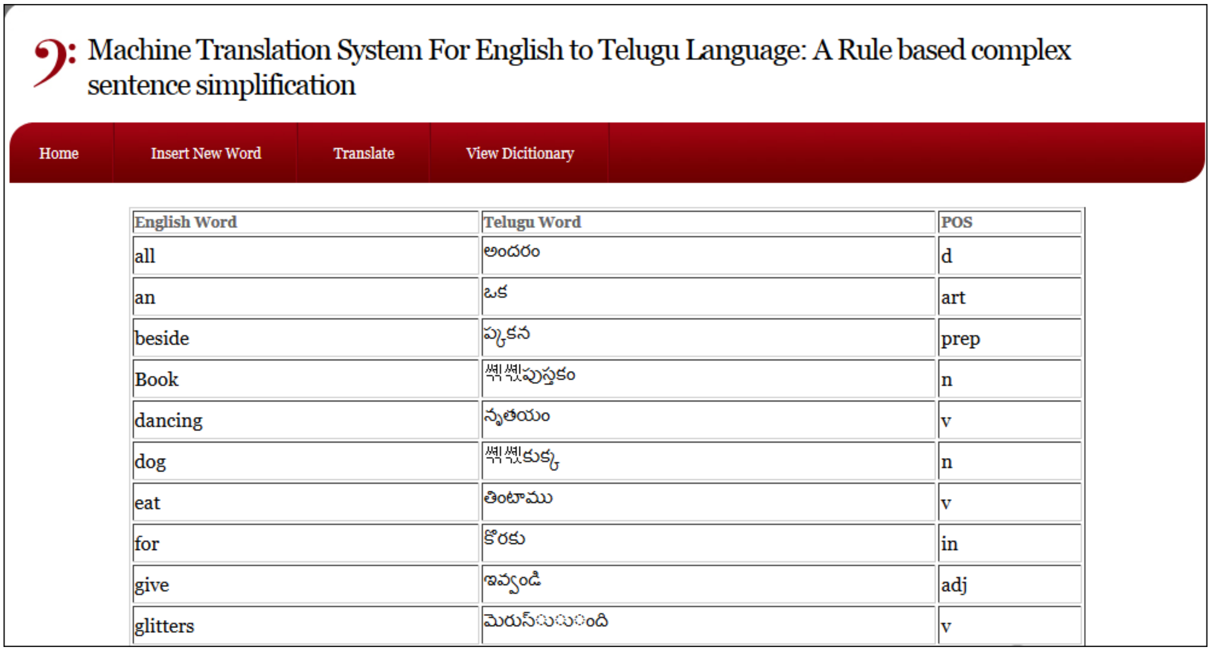


Machine Translation System For English to Telugu Language: A Rule based complex sentence simplification

Sentence Translated Screen

Translated Telugu Sentence: టీచర్ మాకు ఒక ఆసక్తికరమైన విషయం చెప్పారు

Fig 4: Snapshot of translated Telugu Sentence


Machine Translation System For English to Telugu Language: A Rule based complex sentence simplification
Home
Insert New Word
Translate
View Dicitionary

English Word	Telugu Word	POS
all	అందరం	d
an	ఒక	art
beside	ప్కకన	prep
Book	[illegible]పుస్తకం	n
dancing	నృతయం	v
dog	[illegible]కుక్క	n
eat	తింటాము	v
for	కొరకు	in
give	ఇవ్వండి	adj
glitters	మెరుస్ుుంది	v

Fig 5:Snapshot of Words with POS Tags in predefined Dictionary

CONCLUSION

Many of the Indian languages such as Tamil,Kannada, etc. are syntactically similar to Telugu.So many more rules can be developed to decrease ambiguities and achieves better results.This technique will also help for the development of machine translation systems from English to other Indian Languages.This simplification technique gives better results for complex sentences.This system is tested on different data sets, each for training and testing the efficiency and achieved more than 93% accuracy.

REFERENCES

- Akshar Bharati, Vineet Chaitanya, Amba P. Kulkarni and Rajeev Sangal, " ANUSAARAKA: Machine Translation in Stages", A Quaterly in Artificial Intelligence, Vol. 10, No. 3, July 1997
- Sanjay Kumar Dwivedi and Pramod Premdas Sukhdev, "Machine Translation in Indian Perspective", Journal of Computer Science, june 2010.
- Latha R Nair and David Peter S, "MAchien Translation system for Indian Languages", IJCA, Vol 39, No. 12012
- Sugata Sanyal and Rajdeep Borgohain, "Machine Translation Systems in India", arXiv, April 2010.
- Judith Francisca and Md Mamun Mia, "Adapting Rule Based Machine Translation From English To Bangla", IJCSE, Vol 2. No. 3 Jun-Jul 2011.
- Sitender and Seema Bawa, "Survey of Indian Machine Translation Systems," IJCST, Vol 3, Issue 1, Jan-Mar 2012
- Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

- Lewis, M. Paul, Gary F. Simons, and Charles D.Fennig(eds.). 2014. Ethnolouge: Languages of the world, seventeenth edition, Dallas Texas: Sil International. http://www.ethnologue.com.
- Antony, P. J. "Machine Translation Approaches and Survey for Indian Languages." Computational Linguistics and Chinese Language Processing Vol 18 (2013): 47-78.
- Takao Doi and EiichiroSumita. 2003. “Input sentence splitting and translation”, Proc. of Workshop on Building and using parallel Texts, HLT-NAACL 2003.
- Katsuhito Sudoh et al. 2010. “Divide and Translate: Improving Long Distance Reordering in Statistical Machine translation”.
- R. Chandrasekar R, B. Srinivas. 1997. Automatic induction of rules for text simplification.

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- North Asian International Research Journals-221, Gangoo Pulwama - 192301
Jammu & Kashmir, India
Cell: 07298754556, 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com