International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 6 1730 - 1733

Machine Translation Using Open NLP and Rules Based System "English to Marathi Translator"

Mr. S. B. Chaudhari
JJTU Research Scholar (JhunJhunu Rajasthan)
sbchaudhari@yahoo.*com*

Abstract: This paper presents a proposed system for machine translation of English Interrogative and Assertive sentences to their Marathi counterpart. The system takes simple all English sentences as an input and performs its lexical analysis using parser**.** Every token produced by parser is searched in the English lexicon using Lexical analysis. If the token is found in then lexicon, its morphological information is preserved. Here we broadly use Open NLP and Rule Based System. Machine Translation is main areas which focusing to Natural Language Processing where translation is done from One Language to Another Language preserving the meaning of the sentence. Big amount of research is being done in this Machine Translation. However, research in Natural Language processing remains highly centralized to the particular source and due to the large variations in the syntactical building of languages.

Index Terms *- Language Translation, Lexical Analysis, Machine Translation, Natural Language Processing, Rule Based Translation, POS tagging.*

*****

I. INTRODUCTION

Machine translation, is a Heart of Natural Language Processing, is important for dividing and separating the language obstacles and facilitating for bi-lingual translation. Marathi, is a language derived from Sanskrit, is spoken by 80 million people in India. The script currently used in Marathi is called Devnagri Script [1]. While translating source language to target language changing of the word order and its form according to the Marathi grammar of the target language is very important. For the scope of this paper the English is the Source Language and Target Language is Marathi.

Marathi is the one of popular language in India, Basically from Maharashtra i.e. Mother tongue of state Maharashtra. More than 80% peoples speak this language as their mother tongue. This Language is written from left to right, top to bottom of page. The Marathi words id akin to Sanskrit like „mahina‟ as a „maas‟ and „navin‟ as a „nava‟.

The different linguistic people could not able to interact with other language but they will not able to understand. This concept of translation will helps people to communicate. Also help to fill gap between communications of different linguistic people. It will also helpful who have taken education in English but poor knowledge of Marathi.

II. ACTUAL IMPLEMENTATION

In the implementation of this system, it necessary to have vocabulary dictionary. Because with help of dictionary we organizing corresponding Marathi words. Marathi words plays very important role of translation. Dictionary database is endless.

Table 1: Production Rule.

TABLE 2 PRODUCTION RULES			
English Pattern(r)		Marathi Pattern(r')	
r1	S→ n + v + n'	r1'	S →n + n' + v
	Seema + was peeling + potatoes		Seema batate Solat hoti.
r2	S→ n+v+n'+n''	r2'	S→ n+n''+n'+v
	Knowledge + lights + the way + to heaven		Dnyan Swargakade janyacha rasta ahe.
r3	S→ p+art+adj+n	r3'	S→ p+n+adj+art
	It + is a + costly + pen		Te pen mahag ahe
r4	S→ P+v	r4'	S→ P+v
	We+were playing		Amhi khelat hoto
r5	S→ n+v	r5'	S→ n+v
	The moon+shines		Chndra chamakto
r6	S→ P+d+v	r6'	S→ P+d+v
	We+all+breathe		Apan sagle shwas gheto
r7	S→ d+art+n	r7'	S→ d+n+v
	This+is a+picture		He chitra aahe
r8a	S→ n+v+(p+n')	r8a'	S→ n+(p+n')+v
	Karim+cut+(his+finger)		Karim ne tyache bot kapale
r8b	S→ n+v+(p+art+adj+n')	r8b'	S→ n+(p+art+adj+n')+v
	Grandfather+told+(us+a+funny+story)		ajobane gamtidar gosht sangitli
r8c	S→ n+v+(n'+adv')	r8c'	S→ n+(n'+adv')+v
	Habib+goes+(to college+regularly)		Habib collegala roz jato
r8d	S→ n+v+(d+n'+c+n'')	r8d'	S→ n+(d+n'+c+n'')+v
	I+washed+(my+hands+and+face)		mi maze hat ani chehra dhutla
r9a	S→ p+v+n	r9a'	S→ P+n+v
	I+eat+rice		Mi bhat khato'khate
r9b	S→ p+v+(n+p')	r9b'	S→ p+(p'+n)+v
	He+told+the news+to everyone		tyane sagljyana batmi sangitli
r9c	S→ p+v+(n+d+n')	r9c'	S→ p+(d+n'+n)+v
	We+visited+(Diynamoti+last+year)		amhi magil varshi mayanamotila gelo hoto
r9d	S→ p+v+adv	r9d'	S→ p+v+adv
	He+was sleeping+then		to zopla nantar
r10	S→ p+n+v+pre+d+p'+n'	r10'	S→ p+n+p'+n'+pre+v
	each+Muslim+is brother+of+every+other+Muslim		ek muslim dusrya muslim cha bhau asato
r11	S→ v+p+d+adj+n		S→ p+d+adj+n+v
	Give+him+some+crisp+potao-chips		tyala thode kurkurit batate chips de

There for we extend the database as per need.

2.1 ADDING PRODUCTION RULES

We have shown the production rules in fig.1. For both English and Marathi words side by side. In the table „r‟ represent the English rule and „ r‟ ‟ represent the Marathi rule. These rules are individual for each sentence. This rules are also explain in language translation system. The English rule pattern will change according to Marathi grammar rule. In this table indicates not all rules but indicates some rule related translation of sentences or passages/paragraphs.

2.2 PROCESS OF TRANSLATION

2.2.1 TOKENIZATION

The Tokenizer segments an input character sequence into tokens like words, punctuation and numbers. Open NLP has multiple Tokenizer implementations like Whitespace, Simple and Learnable Tokenizer. In this input is Sentence and output is word level token. The following fig: 2. shows the actual blocks of the system how system will work. All the phases in this system will pass through lexical parser. This parser will do lexical analysis as per input sentences and will give morphological structure. Using this structure I produce the rule for Marathi sentences and storing into the database. In this system English and Marathi Lexicons are much more important for word separating and mapping.

2.2.2 POS Tagging

In this part we do the identification of the part of speech such as a noun, verbs, adverb for each word of sentence helps in analyzing role of each rule in sentences. So here “tag” method is used for tagger class of Open NLP. Example: Input – Tokens and Output – tag to each token.

2.2.3 SEARCH THE TOKEN

English and Marathi bilingual vocabulary dictionary is maintain. When we provide some English input to system it will tokenize all words and search into dictionary and given to translator as following Input-Token
Output – Corresponding Marathi Word for Each token. After this we move towards the search rule in database.

1.1.1 SAERCH RULE FROM DATABASE

Here we already store number of rules which contain production rule for translation. So given sentences will be translated according to rule. After POS tagging, the appropriate Marathi word will be fetch from dictionary. Those Marathi words are arranged according to rule and corresponding English to Marathi Translation is shown to user. Input – English sentences
Output– Rule Matching and Corresponding Marathi sentences.

2. ACTUAL PROCESS WITH EXAMPLE

Let us take following example and see translation process:
E.g.: She likes book reading.

1. First this all words must be stored in the dictionary. If not present enter them to dictionary.

2. To add Marathi word also for each English word as pair in dictionary.

3. To add production rule for this sentences that we tokenize this sentence.

4. After tokenize I get 4 words a)She, b)likes, c)book, d)reading. Each word will get assigned one tag and index as follows

She : [0] PRB (means Pronoun)

Likes: [0] VBZ (means Verb)

Book: [0] DT (means determiner/ Article)

Reading: [0] NN (Means Noun)

In this index shows how many words in sentence is particular type. So here in this example one pronoun is present “she” and others are pronoun, verb and determiner.

5. Then we add corresponding rule structure of target language i.e. Marathi. If we translate this sentence in to Marathi then Marathi sentence is:” Tila pustake Vachayala Avadatat”. So here we need to add corresponding Marathi rule as “She books reading like”.

6. So we add this rule to database as follow.

PRB-VBZ-DT-NN | PRB-DT-NN-VBZ (Left part indicate English sentence and Right part indicate Marathi production rule).
After execution of all above steps we got the Marathi sentence as output.Finally, we are not concluded here, in this system we also provide the paragraph/passage translation facility which is not ever provided. Because all existing research are given only for single sentence translation process. After conclusion we also provided some

 ISSN: 2321-8169
1730 - 1733

snapshots of the system. With file upload and Translated file downloading facility.

III. FUTURE WORK

In the future we will do the next type of sentences i.e. Exclamatory and Imperative sentences. Because these sentences are very hard to tokenize which contains some special character like "!". Also like to resolve the ambiguity in the meaning of words in the sentences like "bank". E.g. "I am standing in front of bank". Here two possible context of word „bank" – bank of river or the money bank. Also Grammar of English language allows the change in sentence without changing their meaning to aloe such flexibility in future.

IV. EXPERIMENTAL RESULTS

In following figure i.e. fig: 3, will provide the facility of file unload. The contends of the file will be the number of English statements or passages/paragraphs. After uploading file the system will read all contends from file pass to the parser. Parser will parse all sentences and tokenize it simultaneously system check all Marathi words related to English if found then it will do next process if found then system immediately ask to add Marathi word to vocabulary. The next process is to find production rule from database.
In fig: 4. Shows actual translation system with Input and Output parameters. In this figure you will see that input is in the form of English and output will in Marathi with proper meaning.

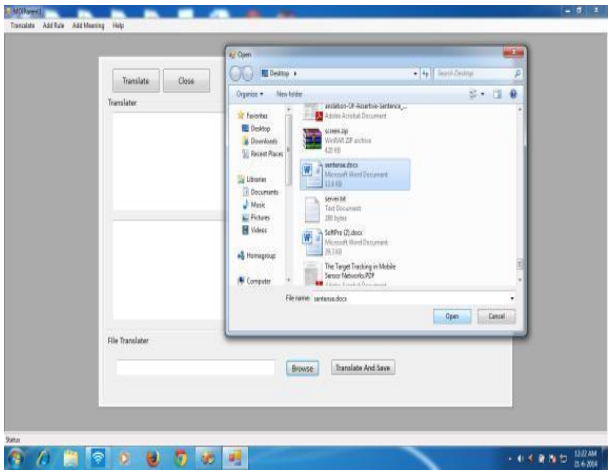

Fig: 3. File Upload To System

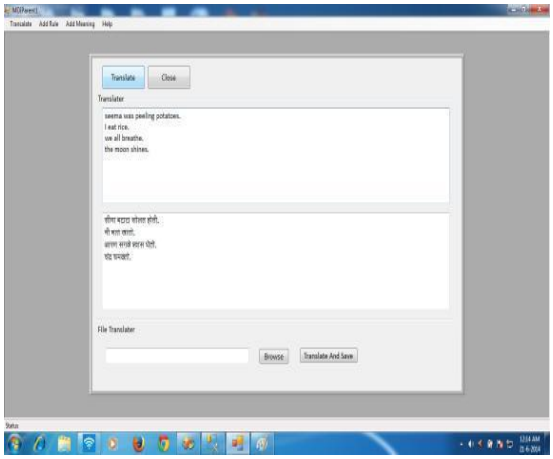

Fig: 4. Actual Translation.

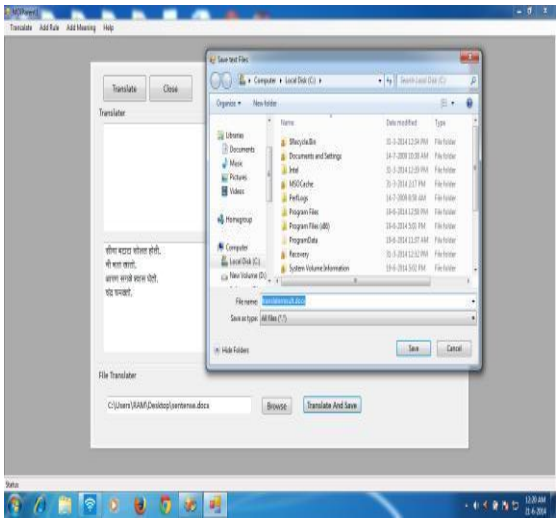

Fig: 5. Save Translated file.

V. CONCLUSION

In this paper, the system work is done as much as possible using self designed parser; in this we have shown totally different work as compared to existing research of language translation. At least in India there is very small work is done for English to Marathi translation. A lot of research is possible in this area. Anyone can do number of variation in this system in future. In this paper we worked only on Interrogative and Assertive sentences. There is unlimited opportunity to upgrade the current research. In Natural Language Processing the numbers of variations are almost unlimited because of its changeable according to the time. Human Language Technology (HTL) that people is making new words for their convenience. Thus the system will provide basic need of machine translation using Open NLP and Rule Based System for English to Marathi Translation.

REFERENCES

[1] Abhijeet R. Joshi, M. Sasikumar, "Constructive approach to teach inflections in Marathi language",www.cdacmumbai.in/design/corporate_site /.../pdf.../CATIML1.pdf

[2] Sangal, Rajeev,Dipti Misra Sharma, Lakshmi Bai, Karunesh Arora, Developing Indian languages corpora: Standards and practice, November

[3] Sangal, Rajeev, Shakti Standard Format: **SSF,** January 2007.

[4] Bonnie J. Dorr, Pamela W. Jordan, John W. Benoit, „A Survey of Cur-rent Paradigms in Machine Translation", LAMP TR-027, Dec. 1998.

[5] Bonnie J. Dorr, „Interlingual Machine Translation: A Parameterized Approach",IEEE transaction on Artificial Intelligence, Volume 63, Is-sue1-2 (October 1993).

[6] Dr. Shridhar Shanvare, „Abhinav Marathi Vyakaran, Marathi Lekhan", Vidya Vikas Mandal, Nagpur.

[7] D.I. De Silva, P.K.D.A. Alahakoon, P.V.I. Udayangani, D. Kolonnage, M.H.P. Perera, and S. Thelijjagoda, Application of Transfer based Machine Translations from Sinhala to English", 978-1-4244-2900-4/08 ©2008 IEEE

[8] Dr. Shridhar Shanvare, „Abhinav Marathi Vyakaran, Marathi Lekhan", Vidya Vikas Mandal, Nagpur.

[9] Naila Ata, Bushra Jawaid , Amir Kamarn, „Rule based English to Urdu Machine Translation", 2007.

[10] Rajiv Sangal, Vineet Chaitanya, „Natural Language Processing- a Paninian Perspective", Akshar Bharati Group,PHI publication.

[11] R. M. K. Sinha and Anil Thakur. 2005. Translation Divergence in English-Hindi MT. In the Proceeding of EAMT Xth Annual Conference, Budapest, Hungary, 30-31 May.

[12] GUPTA, Deepa, and Niladri Chatterjee (2003). Identification of Divergence for English to Hindi EBMT. In Proceeding of MT Summit-IX, pp. 141-148.

[13] Md. Abu Nisar Masud, Md. Munasir Mamun, 2003. A General Approach to Natural Language Generation. In Proceeding of IEEE, INMIC.

[14] S. Khan, Z. Parvez 2003. An Expert System Driven Approach to generating Natural Lnguage in Romanize d from English Documents. In Proceeding of IEEE, INMIC.

[15] R.M.K. Sinha and Anil Thakur. 2005b. Handling ki in Hindi for Hindi-English MT. In the Proceeding of MT Summit X, Bangkok, 12-16 September.

[16] Min Zang, Hongfei Jiang, 2008, Grammar comparison study for Translation Equivalence Modeling and Statistical Machine Translation. In the Proceeding of the 22nd International Conference of Computational Linguistics pages 1097-1104.

[17] T. Mark Ellison, Simon Kirby 2006.Measuring Language Divergence by Intra-Lexical Comparison, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 273–280.