ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Literature Survey on "Statistical Translation System for English to Marathi Using KNN and N-Gram Method"

Shilpa Modani[1], Priyanka More[2] , Harshal Shedge[3], Pramod Salunkhe[4]

B.E. Student, Department of Computer Engineering, BVCOEL Engineering College, Pune, Maharashtra, India[1]

B.E. Student, Department of Computer Engineering, BVCOEL Engineering College, Pune, Maharashtra, India[2]

B.E. Student, Department of Computer Engineering, BVCOEL Engineering College, Pune, Maharashtra, India[3]

Professor, Department of Computer Engineering, BVCOEL Engineering College, Pune, Maharashtra, India[4]

ABSTRACT: Statistical Machine Translation is used to translate one language to another language. It is one of the part of natural language processing. This system having three important models that are Language Model (LM), Translation Model(TM), Decoder. In this system, we will generate an output as decoder and that uses language model and translation model. For this we will uses two languages such as English and Marathi. In this model translation done is English language to Marathi language. For this we will use bigram, trigram, unigram and n-gram methods to find the probability. In this we generate the three results that are with KNN, without KNN, using both methods.

KEYWORDS: Machine translation, Corpus based machine translation, Machine learning, Information retrieval, N-gram, Predictive Analytic, Multilingual chatter, Statistical machine translation, KNN

I. INTRODUCTION

Statistical machine translation is a machine translation model in which the translation is done on the basis of statistical paradigm . The statistical paradigm are divided from the analysis of text corpora. In machine translation the translation is done by the computer system automatically without humans interaction. In the statistical machine translation the translation is done whit the help of dictionary of word, set of words so it is called as dictionary based translation. The translation of one language to another language is done by the machine translation , in which we use many algorithms to rule our system. The translation can be done by Word to Word, Sentence to Sentence and Paragraph to Paragraph. The translation of these all type needs dictionary of words of both the languages.

The statistical machine translation consist of translation model and Language model. The language model is used to store the dictionary of the two languages which we are using for translation. The translation model translate the information with the help of language model. The translation is done by many algorithms , we are using N-gram algorithm for the translation.

In the present days the information searching is done widely but some information are known to us and some are unknown. So to understand the unknown information we need translator which can translate the language of unknown information to our known language so that we can understand it proper. So for this purpose we are creating a language translator which can translate English language to Marathi language and vice versa.

The performance of translation system is very important so to check the performance we are comparing our system with one another algorithm, k- nearest algorithm. We calculate our performance with k nearest algorithm, without k nearest algorithm and with both the k nearest and n- gram algorithm. In this we propose a statistical machine translation(SMT) system for translating English to Marathi language.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

II. LITERATURE SURVEY

The MT is divided by translation into four types as shown in the fig. 1. Namely Literal, Rule based, Knowledge based And corpus based. In this the corpus base is further divided into two types namely statistical translation and ex based translation, the rule based is divided into again two types namely transfer based transformation and interlingua translation.

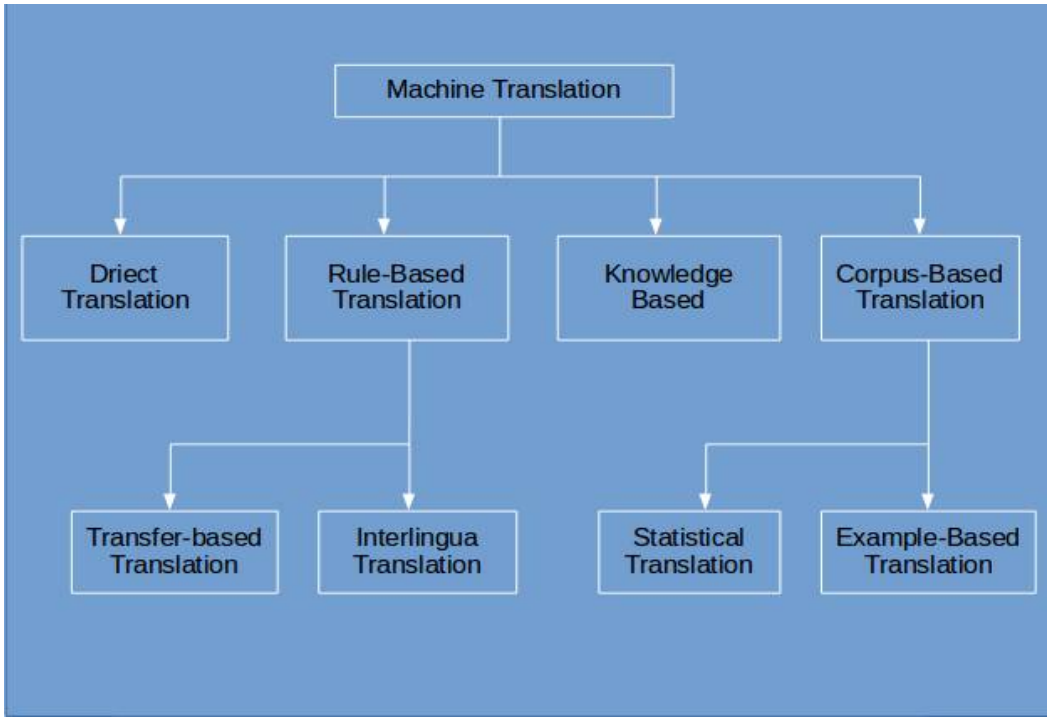

Fig.1

The Literal translation is used to translate the one language into another language one word at a time. So by translating word by word or the lexeme by lexeme of nontechnical type direct translation has meaning of mistranslating words. This translation is used in earlier days for translation. The advantages of this is it is easy to understand and have widen vocabulary. The disadvantage is inaccurate translation and also it is time consuming.
Rule based translation is a system which is linguistic based system. In this system the linguistic information about the Language to be translated and the original language i.e. source and target are retrieved from each language. The form of the dictionary may be unilingual, bilingual or multilingual. It also checks the grammar and the semantics of both the languages. The advantages of rule based translation is mistranslation and the disambiguation problems can be resolve by having full information about the languages. The disadvantage is the ambiguity is avoided by this method so that produces poor results.
The transfer based translation is work in three phases firstly it analyses the source code to check grammatical structure, then translated to the intermediate representation of the target code and then finally generate the target code. In this we use analysis, bilingual dictionary and the target language. In the interlingua translation the source code is translated to many target code that is the transferring the meaning of verbal to nonverbal system.
The corpus based translation is now use widely , it gives translation with the complete knowledge of source and target languages. It is classified into two types statistical and example based translation.
The example based translation the previously computed translation is used as the example for the new translation so that in less time it will give accurate translated result. The system produces new translation with the help of old examples . The main difference between statistical and example based translation is statistical model uses the statistics

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

for the text alignment and the translation is done by the statistical approach, but in example based translation the system uses examples to predict the translation.
The knowledge based translation uses both ontology and dictionary based knowledge. In artificial intelligence, research of language analysis has use large knowledge base with the help of semantic based approach.

III. DESIGN AND IMPLEMENTATION

In this we develop corpus for statistical machine translation by using parallel text of English and Marathi languages. In this the three models are used namely language model, translation model, translation machine.
In this the language model consist of the directories and the related language information. The translation model is used to check the grammatical structure of the source language. The translation machine is used to translate or decode the source text into the target text with the help of language model and the translation model. The fig 2. Shows the design of the translation model.

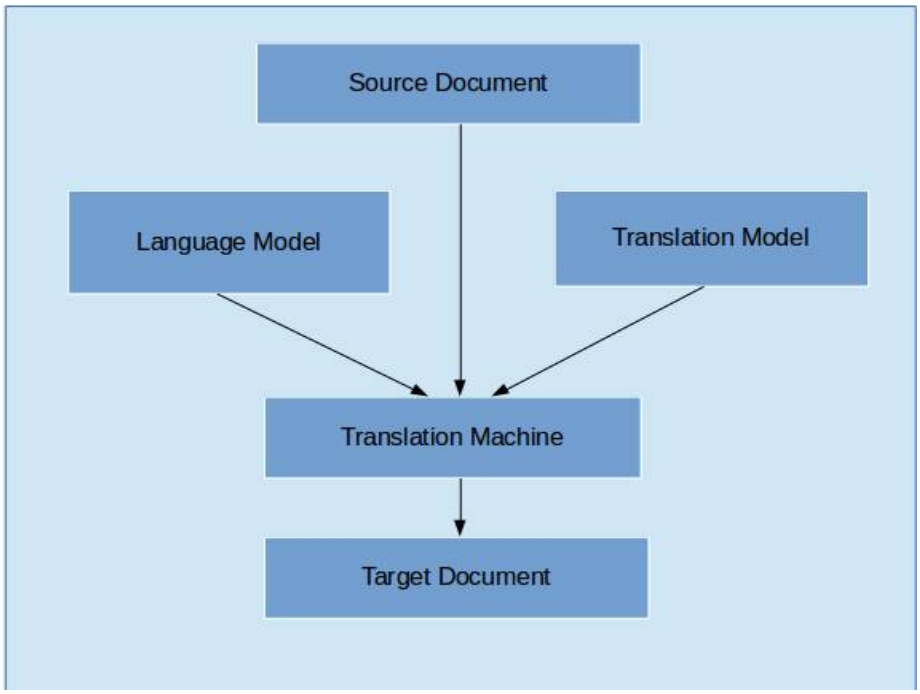

Fig. 2

The language model will determine the probability p(l) for the target language. The language model is dependent for the accuracy and the fluency if the translated text. In this the translation is done by two type by word to word or by sentence to sentence, for word to word we are using unigram algorithm and for sentence to sentence we are using n gram algorithm. Most probably the n gram model is being used. The translation model perceives the maximum probability of targeted language. In this e is considered as English and m is considered as Marathi language. So the English translation will be generated by eq.1.

$$m' = \max_m P(e/m) \quad \ldots(1)$$

Where,

P(e/m)= Probability in file translation.
e = Source language
m = Possible translation of targeted language

By the Bayes theorem we can represent P(m/e) as shown in eq. 2

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

$$P(e/m)=\frac{P(e)P(m/e)}{P(m)} \quad \ldots. \quad (2)$$

By using eq. (1) and (2) we get eq. (3) as follow,

$$m' = \max\_m \frac{P(e)P(m/e)}{P(e)} \quad \ldots(3)$$

In statistical translation, There are many ways to translate a sentence from one language to another. In this the SMT makes the use of translation model, language model and the translation machine.

IV. CONCLUSION

In this, we developed statistical machine translation that translates English language to Marathi language. It translate single English sentence or the group of English sentences into Marathi language. In this we compare the accuracy of translation with knn, without knn and with both the knn and n-gram algorithm. So that we will get probably accurate translated text.

REFERENCES

1. B.N.V Narasimha Raju, M S V S Bhadri Raju "Statistical Machine Translation System for Indian Languages."IEEE 2016.
2. Yulia Rossikova, J. Jenny Li, and Patricia Morreale"Intelligent Data Mining for Translator CorrectnessPrediction."IEEE 2016.
3. Pramod Salunkhe Aniket D. Kadam Prof. Shashank Joshi, Prof.Shuhas patil Dr.Devendrasingh Thakore, Shrikant Jadhav "Hybrid Machine
4. Translation for English to Marathi:A Research Evaluation in Machine Translation"ICEEOT 2016.
5. Pramod Salunkhe Aniket D. Kadam, Prof. Shashank Joshi, Prof.Shuhas patil, Dr.Devendrasingh Thakore, Shrikant Jadhav "A Research Work on English to Marathi Hybrid Translation System"IJCSIT 2015.
6. Pramod Salunkhe Aniket D. Kadam, Prof. Shashank Joshi, Prof.Shuhas patil, Dr. Devendrasingh Thakore, Shrikant Jadhav, "First Survey Effort on Translation Systems" IJETTCS 2016.
7. T Siddiqui, U. S. Tiwari, "Natural Language Processing and Information retrieval," Oxford Press, 2008.
8. Sanjay Kumar Dwivedi and P P Sukhadeve, "Machine Translation System in Indian Perspectives," J. Computer Sci., Vol. 6, Issue No.10, pp. 1111-1116, 2010.
9. D.D. Rao, "Machine Translation A Gentle Introduction", RESONANCE, July 1998.
10. Adam Lopez, "Statistical Machine Translation," ACM Computing Surveys, Vol. 40, Issue No. 3, Article 8, August 2008.
11. Philipp Koehn et.al, "Moses: Open Source Toolkit for Statistical Machine Translation," Proceedings of the ACL 2007 Demo and Poster Sessions, pp. 177–180, June 2007.
12. Resnik, P. et-al, "The Web as parallel corpus," Computational Linguistics, vol.29, Issue No. 3, pp. 349 – 380, July 2003.
13. Koehn, P. et-al, "Factored translation models," In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868 -876, June 2007.
14. Lopez, A., "Hierarchical phrase-based translation with suffix arrays," In Proceedings of the 2007 Joint Conference on Empirical Methods inNatural Language Processing and Computational Natural Language Learning, pp.976 - 985, June 2007.
15. Koehn, P., "Pharaoh: A beam search decoder for phrase-based statistical machine translation models," In Proceedings of the Conference of the Association for Machine Translation in theAmericas (AMTA), April 2004.
16. P.F. Brown et.al, "The Mathematics of Statistical Machine Translation: Parameter Estimation," Journal ComputationalLinguistics, Vol. 19, Issue No. 3, June 1993.