RESEARCH ARTICLE OPEN ACCESS

Amplifying Statistical Machine Translation (SMT) Model Using Graph-Based Approach: A Novel Idea

Priyanka Malviya [1], Prof. Gauri Rao [2]
PG Scholar [1]
Department Of Computer Engineering [1] & [2]
Bharati Vidyapeeth Deemed University College of Engineering
Pune - India

ABSTRACT
The Statistical machine translation system arises in the India to translate the source language text to target language text for the better communication. In the previous system, Rule-based English to Hindi Translator translates English affirmative sentences in Hindi. During this analysis, a system has a tendency to convert the simple English affirmative sentences to Hindi sentences. The Hindi language is still unfocused. In this system, English sentences will be taken as an input and output will be given as Hindi Sentences. To enhance the process of Machine Translation graph based pruning approach is introduced to achieve optimized results, better accuracy and efficient translation for English to Hindi Language Translation.

Keywords :— Accuracy, Communication, graph-based, statistical machine translation, pruning, Translator.

I. INTRODUCTION

In the present time, language services industries growing by taking trends as Machine Translation (MT).It offers a solution for the small budget, high volume, and short timeline projects. Because of Machine Translation, the benefit of translations automation may exceed. Machine translation can match word to word but unlike human translation does not contain the capacity to convert text like Machine Translation. The benefit of machine translation is to decrease turnaround time for translation projects and elimination of human errors such as typing mistake or spelling error and provide clarity and accuracy in sentences. Then with the Machine Translation, you can experience the benefit like:-

1. By proving the quality test.
2. Compliance increase readability level.
3. High caliber translator –which increase 50000 professional translators in the top 50 global languages are presented.
4. Machine translator having personalized customer care.
5. It is cost effective, user-friendly translation to reduce budget solution (problems).

Rule-based English to Hindi Translator converts English affirmative sentences to Hindi [1].During this analysis work; we have a tendency to convert the simple English affirmative sentences to Hindi sentences. That is often mainly a Machine Translation. We have chosen the transfer-based approach that is the thin line between the linguistics and also the direct approach. For that have designed the computer program or can say parser that helps us to map English sentence combining to the rules and then obtaining rules regenerate into a target language.

At this stage, humans, as well as NLP systems, interpret the meaning of individual words [2]. Many types of language processing contribute to word-level understanding – the first step to being a post or assign of a single part-of-speech tag to each word. Also at the lexical level, a semantic representation of that meaning can be replaced by those words that have only one possible meaning or sense. The nature of the representation changes due to the semantic theory utilized in the NLP system. This state focuses on understanding and finding the words in a sentence so as to uncover the grammatical rule, structure of the sentence and phrase. It needs both a grammar and a parser [1], [2].

Machine translation has different architectures such as Direct, Transfer Based, Interlingua, Statistical, Example-Based, and Hybrid. Each of them has its advantages and disadvantages and selection of the approach can be made based on the domain of the application. Proposed research work is an innovation and presents a programmable Machine Translation system. There are many problems arises in the machine translation many approaches are present to solve the machine translation problems of phrases from one language to another by phrase-based translation. This Proposed system to enhance the process of machine translation system introduced

graph based pruning approach achieves the optimized result and accurate translation for English to Hindi Translation.

These paper contain the following terms are as follows first features used in this proposed system, Second Related Work, Third proposed Methodologies, and the conclusion.

II. FEATURES USED IN THIS PROPOSED SYSTEM

A. *NoSQL Database:-*

NoSQL is abbreviated as "Non-relational or not only SQL" database which contains the mechanism of storage data and retrieval of data ,which is not arranged in the tabular relation format (or modules) as used in the relational database. The relational database is an older database system, but it didn't obtain (or work) on the NoSQL database. But as we see that use of the internet is an increase since the twenty century because of google,facebook,Amazon.com etc. uses big data, and real-time web application ,the use of NoSQL system are increased to support SQL-like query language to manage a large amount of data. The motivation of NoSQL includes ssimplicity of design and manage data, simpler scaling, and clustering, of machines, finer control over the availability of data.

B. *Database:-*

Neo4j is used to create the graph database in this proposed system. The Neo4j is taking the advantages of the strength of multiple database technologies. Example: -social databases application might use a document database type of data like Excel file (such as MongoDB) to search a word along in a graph database (such as neo4j) to provides real-time personalized product recommendation. To plot a graph in the application needs to store data in multiple databases, each having its data model (graph vs. document) with the connection of MongoDB to Neo4j and synchronized data in document and graph type automatically makes the process simple.

C. *Graph-based database approach:-*

It is the type of NoSQL database which use graph theory to store, evaluate, map and find query relationship between data. The graph is the collection of nodes and edges. The graph is having lots of interesting terms by using graph database to mine data from social media, big database warehouse, etc. To manage the graph database the tools Neo4j and Hadoop are used.

D. *Subgraph matching :-*

In the computer field, the Subgraph isomorphism contain Computational mechanism as G is the graph which includes H as a subgraph which is given as the inputs and one of the graphs must determine whether G contain a subgraph i.e. isomorphic to H. Sometimes subgraph isomorphism may be solved in polynomial time and graph contain a Hamiltonian Cycle and is therefore NP-Complete. In these G= (V, E), v is the vertices/node, e is edged H= (V', E') is the graph. The equation for subgraph matching G_0= (V_0, E_0): V_0 subset of V and E_0 is a subset of E, n ($V_0 \times V_0$) is the vector of vertices. Mongoose is used to create and store a graph in MongoDB database in Neo4j.

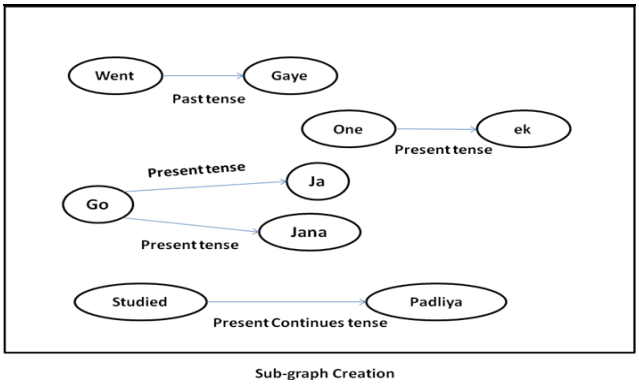

Fig.1. Sub-graph creation

Example:-'went, go, studied' is the node and which is connected to the other node through edges (which contains some value like; Gaye, ja, jana, padliya).

E. *Bucket creation:-*

A Bucket is a mechanism which is a logical unit of storage are used to store objects, which consist of data and data about data (metadata) that explain the data.

F. *Pattern matching:-*

It is the mechanism to check the given input sequence of token for some existing patterns. The pattern matches each in sequence structure or tree structures.

III. RELATED WORK\

In these proposed system Author Present a statistical machine translation model that uses hierarchical phrases that contain sub-phrases. Using Bilingual (BLEU) as a metric of translation accuracy, Author finds that system performance increases the accuracy using tools. The limitation of this methodology contain Phrase-based Translation suffers from

accuracy, and presented framework is developed for flat structures, not for hierarchical complex structures. Syntactically informed statistical machine translation is the scope of this proposed method.[3]

In this system Author introduces a word alignment framework that facilitates incorporation of syntax encoded in bilingual dependency tree pairs, Generate shorter outputs On baseline word alignment the limitation of this methodology is Comparative evaluation is not presented. Future work Bootstrap alignment using simple heuristics without relying on complex IBM models.[4]

This proposed system, describe an approach for filtering phrase tables in a Statistical Machine Translation system. This proposed system is implemented only one way of stating a null hypothesis that neutralizes translation relation in a bilingual corpus this is the limitation. The future system can be extended on false discovery rate present on the language conversion model. [5]

This proposed system in 2010 presents typical phrase-based machine translation (PBMT) system which works on the phrase conversion from a source language to target language. The limitation of this method is rich syntactic and structural needs to be incorporated for better translation scope is to Improving LCS decoding algorithm using syntactic clues.[6]

The author presents a novel translation quality informed procedure for both extraction and scoring of phrase pairs in PBSMT systems. The supervised approach is adopted can be further extended to unsupervised and finally best semi-supervised. Multiple combinations of the methodology can be used to improve a system in the future. [7]

This system proposes and focuses on a method for classifying bilingual entries that were automatically extracted from aligned parallel corpora as correct or incorrect, by using a Support Vector Machine based classifier. Comparison with the usual approach using just Moses and parallel corpora aligned at sentence level are to be reported is the limitation of this proposed system. Extend this technique of machine learning, to classify the translation equivalents distant language pairs such as English Hindi. [8]

Author introduce a novel pruning criterion that places phrase table pruning on a sound theoretical foundation only the model p (˜e| ˜f) has been taken into account for segmentation. A better estimate by also considering the distortion costs can consider in future work. [9]

In this system, author describe phrase-based and hierarchical phrase-based statistical machine translation systems which are uses to increase the translation accuracy by phrases and word matching using the Moses tool. This method can select the most useful phrase pairs and rules, including those that are low in frequency but still very useful. Pruning framework could be embedded in other translation models as a syntax-based model is a future work. [10]

IV. PROPOSED METHODOLOGY

Features of the proposed methodology are depicted in figure 1. And it can be explain with the following steps.

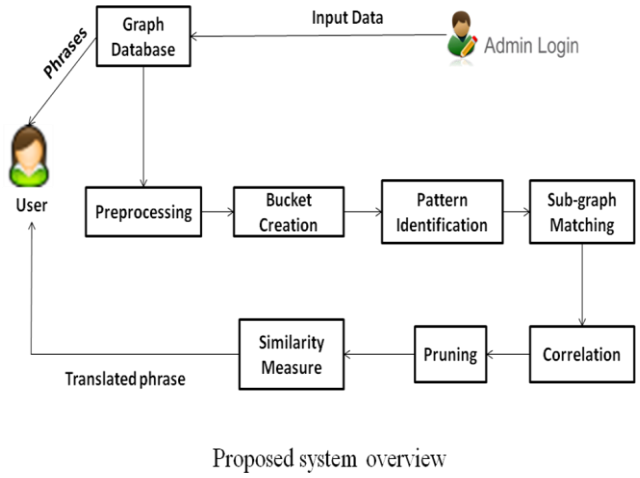

Fig.2. Proposed System Architectural Overview

Mathematical Model for Machine translation of proposed system

1. Let M={ } be a Machine Translation
2. Identify input as Q={q1,q2,q3,q4……qn}
   Where qn is the phrase data element
   M= {qn} of English language
3. Identify Td as an output of translated Hindi language
   M={qn,Td}
4. Identify process P={Pp, Bc , Pi, Sm ,Pc ,Jc ,Pr}

Where,
Pp=Preprocessing,
Bc=Bucket Creation,
Pi=Pattern Identification,
Sm=Subgraph Matching,
Jc= Jaccard Coefficient
Pr=Pruning
M= {qn,Pp, Bc, Pi, Sm, Pc, Jc, Pr,Td}
Union of a subset of M gives the final result.

A. *Graph creation:*

This is the most important part of the proposed methodology, where phrases are feed in the worksheet to the model. By using these phrases, a proposed system identifies the unique phases which are considered as the nodes and their meanings are treated as the edges. After

that, a relationship matrix is evaluated to determine the desired nodes and edges. Once this process is thoroughly finished, then they are represented by way of graph data structures and stored in a neo4j graph database.

Graph Creation

Algorithm: - Create_Graph (N1, N2, e)
Input: - Phrases contain English Word
Output: - Combine matching of English sentence with Hindi as graph

1. Set_Graph
2. **for** all elements in N1, N2 ,e **do**
   Recursively analyzing immediately meaning of English sentence to Hindi, node N1, connect with node N2,
   all possible checking and matching create feasible subgraph
3. **for** all possible subgraph connect through edges which contain English grammar tenses (Present, Past, Future)
4. **do** insert or update or modified as nodes are proper matched of meaning of every phrases.
5. Finish graph creation

Created graph stored in the database neo4j

B. ***Pattern identification:***

Once the graph database is ready then a user can feed the input to get its desired translation terms. Initially given sentences are preprocessed by trimming and tokenization and then they are subjected to pattern identification via bucket forming process where all possible meaningful combination are evolved from the given input words and tagged with its base words. These tagged words are coarsened to get the proper enhanced meaningful alternative phrases and then they are clustered in the desired manner for the further use.

C. ***Subgraph Identification through correlation:***

Here in this step of our model based on the clustered phrases along with their patterns a graph data structure has been created and this data structure is matched with the main graph database stored in neo4j. Then by using correlation technique which involves Pearson correlation rule finally yields the most correlated part of the graph that we call as a subgraph. Pearson correlation yields the correlation between two vectors in the form of a decimal Value which deviates between 0 to 1. Any value of correlation which deviates towards 1 indicates higher correlation.

D. ***Similarity index for translation:***

This is the last step of translation where all the correlated phrases are gathered, and they are checked for similarity index by using Jaccard coefficients that can be represented by the equation 1.

$$Jc = P / (P+q+r) \quad \text{(1)}$$

Where,
Jc – Jaccard Coefficient
P= for both objects Number of variables are positive
Q= Number of variables that are positive for ith object and negative for jth objects
Q= Number of variables that are negative for ith object and positive for jth objects
Q= Number of variables that are negative for both objects
Then the results of similarity indices are optimized for providing the best phrase translation in Hindi for the given input.

V. CONCLUSIONS

In modern statistical machine translation systems, accurate language translation is a complex task, for that various methodologies are present but still facing accuracy problems. Above system, this method requires different solutions to overcome accuracy problem. One of the solutions is proposed in this system on programmable machine translation in English to the Hindi language, which is based on Graph-based pruning approach which will eventually remove complexities of previous research with better accuracy.

ACKNOWLEDGMENT

To prepare the Paper of-" Amplify Statistical Machine Translation (SMT) using graph-based approach: A Novel Idea" has been prepared by Miss Priyanka Malviya and Prof. Gauri Rao.

Author to thank my faculty as well as my whole department, parents, and friends for their support and confidence and obtained a lot of knowledge during the preparation of this document.

REFERENCES

[1] G. V Garje and G. K. Kharate, "Survey of Machine Translation in India," *Int. J. Nat. Lang. Comput.*, vol. 2, no. 4, 2013.

[2] P. Dungarwal, R. Chatterjee, A. Mishra, and A. Kunchukuttan, "The IIT Bombay Hindi ⇔ English Translation System at WMT 2014," 2014.

[3] D. Chiang, "Hierarchical phrase-based translation," Comput. Linguist. vol. 33, pp. 201–228, 2007.

[4] Y. Ma, S. Ozdowska, Y. Sun, and A. Way, "Improving word alignment using syntactic dependencies," in *Proc. 2nd ACLWorkshop Syntax andStructure in Statist. Translat.*, Columbus, OH, USA, Jun. 2008, pp.69–77.

[5] N. Tomeh, N. Cancedda, and M. Dymetman, "Complexity-based phrase-table filtering for statistical machine translation," in *Proc. MT Summit XII*, Ottawa, ON, Canada, Aug. 2009.

[6] M. Posta, Y. Caob, and G. Kumarb, "Joshua 6: A phrase-based and hierarchical statistical machine translation system," *Prague Bull. Math. Linguist.*, no. 104, pp. 5–16, 2015.

[7] N. Tomeh, M. Turchi, G. Wisinewski, A. Allauzen, and F. Yvon, "How Good Are Your Phrases? Assessing Phrase Quality with Single Class Classification," in *Proc. Int. Workshop Spoken Lang. Translat.*, San Francisco, CA, USA, Dec. 2011, pp. 261–268.

[8] K. Kavitha, L. Gomes, and G. P. Lopes, "Using SVMs for Filtering Translation Tables for Parallel Corpora Alignment," in *Proc. EPIA*,2011

[9] R. Zens, D. Stanton, and P. Xu., "A systematic comparison of phrase table pruning techniques," in *Proc. Joint Conf. Empir.Meth. Nat. Lang. Process. Comput.Nat. Lang. Learn.*, Jeju Island, Korea, Jul. 2012, pp.972–983.

[10] M. Tu, Y. Zhou, C. Zong, and S. Member, "Translation Model Pruning," vol. 23, no. 11, pp. 1847–1857, 2015.

[11] G. Rao, C. Agarwal, S. Chaudhry, N. Kulkarni, and D. S. H. Patil, "Natural language query processing using semantic grammar," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 219–223, 2010.

[12] G. Rao and R. Based, "THREE DIMENSIONAL VIRTUAL ENVIRONMENT FOR Address for Correspondence," no. Ii, 2011.

[13] Rao, Gauri, et al. "Natural language query processing using semantic grammar." *International journal on computer science and engineering* 2.2 (2010): 219-223.

[14] P. T. Patil, "A Study on Evolution of Storage Infrastructure," *Patil Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 6, no. 7, pp. 501–506, 2016.

[15] Pawar, Supriya Haribhau. "A Study on Big Data Security and Data Storage Infrastructure." *International Journal* 6.7 (2016).

[16] P. Malviya, "A Study Paper on Storage Area Network Problem-Solving Issues," vol. 4, no. 4, pp. 151–156, 2016.

[17] Patil, Sheetal S., Sunita S. Dhotre, and Uday C. Patkar. "SMART UTILITY FOR MSAS 2005." *Journal of Engineering Research and Studies E-ISSN* 976: 7916.

[18] S. Patil and P. G. Rao, "Web Page Template Generation and Detection of Non- Informative Blocks Using Trinity," vol. 5, no. 3, pp. 70–72, 2016.