ARTICLE

A HEURISTIC APPROACH FOR GRAPH BASED MACHINE TRANSLATION

Priyanka Malviya*, Gauri Rao, Rohini B. Jadhav, Mayuri H. Molawade

Dept of Computer Engineering, Bharati Vidyapeeth Deemed University, College of Engineering, Pune, INDIA

ISSN 0976-3104

ABSTRACT

Background: *Due to the drastic increase in electronic gadgets and widespread use of the computers two-third of the world population is using one or another type of software's. So this insists software industry develop software's based on linguistic approach. This makes machine translation as one of the most important key research topics in the current era. Many automatic machine translation methodologies exist like linear String matching, Pre context matching and much more. But every system has one or more exceptions that may yield some wrong outputs on some selected occasions. So as a tiny step towards this,* ***Methods:*** *this paper deals with the machine translation technique using the graph based approach where for the given phrases in English as respective Hindi phrases are identified. This technique explicitly uses the matrix space translation method for word identification which is catalyzed by the process of similarity index of Jaccard distance. The relevant phrases are formed for the given phrases in English based on the sub-graph matching and correlation technique.* ***Results:*** *Hence when the English phrases are given to the system as input, the output is Hindi phrase sentences.* ***Conclusions:*** *So, The proposed system using graph based approach having an advantage of poly-directional traversing ability to identify larger semantics between the two vocabularies. This is an added advantage to increase the performance of the system to give best results.*

KEY WORDS
Sub graph matching; Correlation; Pruning; Similarity index; Pattern identification.

Received: 23 June 2017
Accepted: 14 Aug 2017
Published: 8 Sept 2017

***Corresponding Author**
Email: priyankamalviya1006@gmail.com
Tel.: +91 9175935295

INTRODUCTION

Machines producing the translation from natural language to other without human assistance is termed as Machine Translation [1]. Computational linguistics is Major Domain research domain facilitating text to speech based translation from one language to other. A major challenge is design and development of complete automated Machine Translation System with limited language modeling and other limitations. A Large training dataset of the parallel corpus has been required to achieve reasonable translation. Currently, multilingual machine translation is necessitated to overcome language gap. [2]Target language lexicons with syntactical modeling are basically done in deriving translated text. Future more semantics have been considered inaccurate translation. In commonly observed that target language might have SOV(Sentence Object Verb) pattern whereas source might be in SVO(Sentence Verb Object) pattern. Modeling of this divergence has been done using rules [1]. In context to India 18 constitutional languages with 10 scripts exists and a large number of sub- local languages have been derived. Commonly Adopted Approach is Rule-based Statistical and Hybrid Translation.

Rule-based Translation: Translation system build on language protocol has been termed as a rule based system. Rule-based Translation is based on word to lexicon translation, Expert crafted rules have been stored in a database [3],[16],[28].

Statistical translation: Phrase to phrase Translation [30] is termed as statistical translation overlooking language format. Fluency has been achieved with statistical translation; as such most industry products like Google translator are based on statistical translation.

Graph database: Graph database seems to be the promising solution to model large and huge data. Currently, social networking and routing data are been modeled using graph database [4],[30].
Correlation: In most data mining application confidence is a major factor affecting the relevance of output. Correlation is a mathematical value that assists in computing strong relationship with two features. [30]
This research work is been organized in six sections as shown below Section I Introduction Section II Related Work, Section III proposed Methodology Section IV results and Discussion Section V Conclusion and Future Scope

RELATED WORK

To achieve optimal translation results, pattern based Ambiguity is to be modeled. Similarity and dissimilarity measures assist in evaluating pattern based similarity. Solving Classification and clustering problem [7].

System [8] methods focuses on "language divergence" challenge in translation. Interlingua and transfer based Approach for English to Hindi Translation. Work under Research [9] presents Angla Hindi an enhanced version of ANGLABHARATI system. Core Methodology is rule based Interlingua translation. Proposed System uplifts performance with example and statistical analysis.

System [3] has innovated simplified Methodology based on Integration twin techniques firstly ordering English sentences To Hindi Syntax, secondly applying Hindi Suffix. The system has been developed based

ISSN: 0976-3104

smaller dataset. System methods [10] have been delivered at IIT Delhi and address pattern ambiguity in English to Hindi translations. Word net has been found to be the solution to eliminate pattern divergence. The [11] system has based on the monolingual corpus. Corpus has been developed consisting of 44 million sentences and 787 million tokens. Corpus-based approach enhances statistical translation. The corpus is freely accessible. Scope remains to clean and make better corpus dataset.

System [12] addresses a new challenge in translation. Code-mixing frequently encountered today, is mixing of two languages and using them. In the case of English and Hindi, Hinglish is the outcome. System address to identify foreign clauses in translation and perform the pure translation are used in several reports. System have specialized analyzers for Hindi-English summing unknown words. Overall perspective on Indian language has been addressed [13]. Large divergence remains in Indian languages, official reports have been built in local language building language gap. System [14] is designed on joint channel model with the alternative hypothesis. NEW 2009 dataset has been used to train system. The accuracy of 0.47 and F-score of 0.86 has been observed. Limitation observed is for nonstandard runs system performance falls. The scope of work is the enhancement of ranking algorithm token set modification. Matra fully automated system for English to Hindi translation has been described by [15]. The methodology of Matra is robust parsing, Incremental progress has been observed with enhancement in linguistic capabilities. System [16] introduces English to Hindi (EHMT) Translation system. The methodology is analyzing sentence structure of input English and generating Hindi Output based on recursive Tree writing algorithm. Word-Sense Ambiguity is majorly observed Challenge in language translation. [17] This system addresses word sense ambiguity and proposes a framework to eliminate Ambiguity.

These methodology adopted is the rule based correction with statistical input. Ambiguity has been handled effectively with word net implementation. In this System work author [18] marks that language divergence is the major challenge in English to Hindi Translation. Example based Translation system commonly face this challenge. Proposed framework identifies Divergence. Level based Translation system has been implemented by [13]. In the actual scenario no fully automated system exists for any language translation. Challenge observed in the context of translation and audience view.

Statistical translation corrective methodology has been adopted [19]. Effective corrections are done on statistical translation result in better output. A source parser and rule learner model are been presented. The author summarizes the overall status of Machine translation systems [20]. Standalone system, web-based software has been developed in the form of plug-ins for translation. Elaborated survey has been presented on every Indian translation system recognized.

System [21] addresses English to Marathi translation, Adopted methodology is Hybrid Translation. Proposed System implements Statistical translation and then applies rule-based translation for effective translation output. Hybrid Methodology [22] is adapted in English to Marathi translation produces better output. Observations marked by this method are rule based System are costlier and lack fluency. On another hand statistical system lack accuracy but have higher fluency.

Proposed statistical based translation correction with rule based enhances the system. This system [23, 24] available online feature graph based search and effectiveness of graph model. Graph model has been used in the scenario where large data exists to be modeled like social networks, routing paths. Graph based approach assist in mining patterns between objects. Comparative analysis of graph-based search presents Neo4j (Graph database to be better than RDBMS data schema. Graph data model is the scope of data mining and surely would assist large data handling effectively.

This system [26] has been designed on Moses and deep learning framework. System work [27] focuses on combining rule-based and statistical-based approaches. Each methodology [27], [29], [30] has got own benefits. The fusing of two techniques has been done to achieve the better system.

This system [25] first attempts to find research problem in machine translation. Proposed methodology is statistical machine translation based on graph based Approach. Above article is an extension [20] with complete research evaluation and algorithmic strategy. Graph-based Approach is ultimately the best approach only first of kind innovation presented by above system methods.

MATERIALS AND METHODS

The proposed system of machine translation from English to Hindi phrases can be detailed in the above overview [Fig. 1]. The Methodologies of this approach is deeply discussed in the below-mentioned steps.

IIOAB JOURNAL
ISSN: 0976-3104

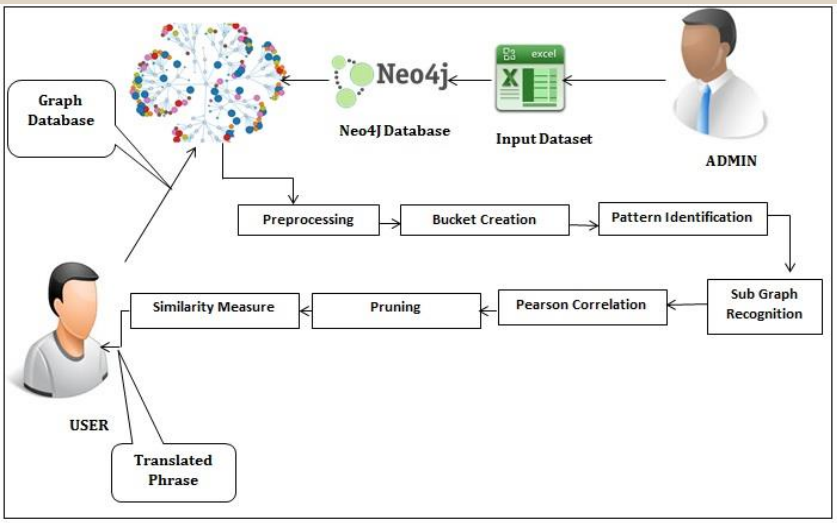

Fig. 1: System model.

A. Step *1:*

The proposed system is based on the graph based approach for identification of the proper phrases for the Hindi language for the given phrase in English as input.

As the very initial step of the system we need to train the proposed model with different words in English and their respective alternatives in the Hindi language. So the system needs to feed by this database to train itself by the creation of hyper-graph for this process. A hyper-graph is crafted based on the identification of proper nodes which are unique in their category. This identification of unique elements will be repeated to all categories like English words, Hindi translated words, phrase adjectives, etc.

Once the unique nodes are identified, then these nodes are bounded by the respective edges which indicate the relationship between the two nodes in the form of closest semantics.

Then by using these nodes and edges a hyper-graph is created where common features found clustered on traversing with inner edge formation in between the tiny clusters. This graph is stored in most advanced graph database like neo4j.which will be used further when query phrases are given to the system as input.

ALGORITHM 1: Hyper graph Creation

//Input: Data collection Set S = {s_i,h_i,s_m}
//Output: Hyper Graph **G** (s_i,h_i,s_m)
Where
S_i – English Words (node)
h_i – Hindi words(node)
s_m- Semantic

Step 0: Start
Step 1: Get the Set **S**
Step 2: FOR **i=0** to Size of **S**
Step 3: Separate s_i, h_i into List L_s,L_h
Step 4: **END FOR**
Step 5: Get unique elements form L_s andL_h
Step 6: N_s=Size of L_s(Number of nodes for English words)
Step 7: N_h=Size of L_h(Number of nodes for Hindi words)
Step 8: FOR **i=0** to Size of N_s
Step 9: FOR **j=0** to Size of N_h
Step 10: Identify the relational Edges **E** using s_m
Step 11: Form Graph **G**
Step 12: **END FOR**
Step 13: **END FOR**
Step 14: return **G**
Step 15: Stop

B. *Step2*

Here in this step user will feed the phrases in the English language which are preprocessing to remove any special symbols and then these words are tokenized in a vector for the further processing.

ISSN: 0976-3104

C. Step 3:

Here in this step, the internally hidden words from a word are identified by using matrix space translation process. Where all the word combinations are evaluated by combining with the next character of the word. By doing this system derives other existing words from the given words. This can be easily denoted with the following model.

If the user is given the word called "Going" then our system identifies all its combination words like go, goi, goin, going. Then these four combinations are checked in the dictionary for their existence in English language and then compute a relationship between the words. Like in the considered example system identifies “go" and "going" as the two words from the given word “going” and identifies "present continuous “as the relationship between the two words. This can be represented by the algorithm 2.

ALGORITHM 2: MATRIX SPACE TRANSLATION

//Input: Data collection Set **W** = {wi}
//Output: Matrix space Set **M**

Step 0: Start
Step 1: Get the Set **W**
Step 2: FOR **i=0** to Size of **W**
Step 3: get S_i of Wi
Step 4: FOR **j=0** to length of si
Step 5: sb_i=substring (s_i, 2→j)
Step 6: Add sb_i to **M**
Step 7: **END FOR**
Step 8: **END FOR**
Step 9: return **M**
Step 10: Stop

D. Step 4:

After the step of matrix space translation with derived words from this is used to create multiple patterns of resulted language words. These patterns are derived with the well-organized combination of words in the entire possible manner. This pattern creation is achieved by using power set generation techniques which yield true combination of words which are represented in the form of vector.

E. Step 5:

This is the part where already created graph database has been queried to get the best possible subgraphs for the created patterns. And these subgraphs are analyzed for their correlation with the Hindi words based on the semantics. While performing this correlation, two lists have been created for both the language words for all the possible subgraphs lists to measure the correlation between them using Pearson correlation, which can be represented using equation 1

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(x^2 - \frac{\sum x^2}{n})}\sqrt{(y^2 - \frac{\sum y^2}{n})}}$$

Where
x is the entities of query phrase
y is the entities of resulted phrase
n is the size List

On evaluating this correlation value will be yielded by r in between 0 and 1. So any value which is nearer to 1 always indicates better correlation.
So the system prunes the values which are less than mid probability, i.e. 0.5 and then all other remained phrases are collected in a list for further investigation of similarity measure.

F. Step 6:

This is the last step of the proposed methodology were finally extracted phrases of Hindi language are measured for the similarity index using Jaccard distance. All the characters of extracted phrase words are exchanged with the proper ASCII values of Hindi language to display the results title.

IIOAB JOURNAL
ISSN: 0976-3104

RESULTS AND DISCUSSION

To deploy the proposed model system uses Java technology with Net-beans as IDE and neo4j as the graph database.The system is put under the hammer for vigorous tests to evaluate its performance by conducting several tests as discussed below.

Mean Reciprocal Ratio (MRR)

After translation of the phrases from English to Hindi system is to rank the answers in between 0 to 6 as per the satisfaction. For all the ranks from 0 to 6 Reciprocal ratios are assigned (RR) ***as 1/1, 1/2, 1/3, 1/4, 1/5, 0.***
For example if ranked the system as 2 for a result, then reciprocal ratios will 0.5. If it is 3 then reciprocal ratios will 0.33. For any value that is 6 then RR is 0.
The mean reciprocal rank (MRR) is the average score over all yielded answers.

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{Rank_i}}{N}$$

Where,

$Rank_i$ is the rank of the first correct occurrence in the top five ranks for yielded results
N is the number of tested results.
The system performed an experiment to evaluate the rank retrieval using the MRR, and the results are summarized as shown in [Table 1].

Table 1: MRR for different Runs

No of Phrase	MRR
10	0.98
20	0.88
30	0.79
40	0.89
50	0.88

On plotting graph for the values tabled in [Table 1], observe that proposed model yields average MRR of 0.822 that is 82.2 % and this shows the good sign of any translation system in heuristic approach.

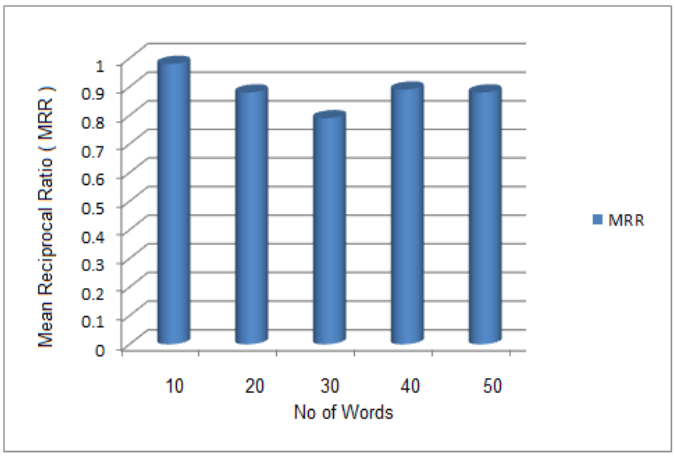

Fig. 2: Performance evaluation through.

Performance evaluation

Headings, The performance of the system are measured using most common and effective features like precision and recall. Precision is used to evaluate the relative preciseness of the system which always exposes the fact of Positive-ness of the system. Whereas Recall is using to measure the relevance of the system and it shows the sensitivity of the system through which it has been evaluated.

Precision can be stated as ratios of relevant phrases are translated to the sum of Relevant and irrelevant phrases are translated.
Which can be more deeply explain with the below equation

$$P(x) = (R/(R + IR)) * 100$$

IIOAB JOURNAL
ISSN: 0976-3104

Where,

P(x): Precision Function
R: Number of Relevant phrases is translated
IR: Number of Irrelevant Phrases is translated

Recall can be stated as ratios of Relevant phrases are translated to the sum of relevant phrases are translated and relevant phrases are not translated. Which can be more deeply explain with the below equation

$$R(x) = (R/(R+!R))*100$$

Where,

P(x): Recall Function
R: Number of Relevant phrases is translated
! R: Number of Relevant Phrases is not translated

System conducted the experiment based on precision and recall parameter and gathered information is tabulated in the below [Table 2].

Table 2: Precision and Recall Performance

No. of Phrases	R	!R	IR	Precision	Recall
10	10	0	0	100	100
20	16	1	3	84.21053	94.11765
30	25	0	5	83.33333	100
40	38	1	1	97.4359	97.4359
50	47	1	2	95.91837	97.91667

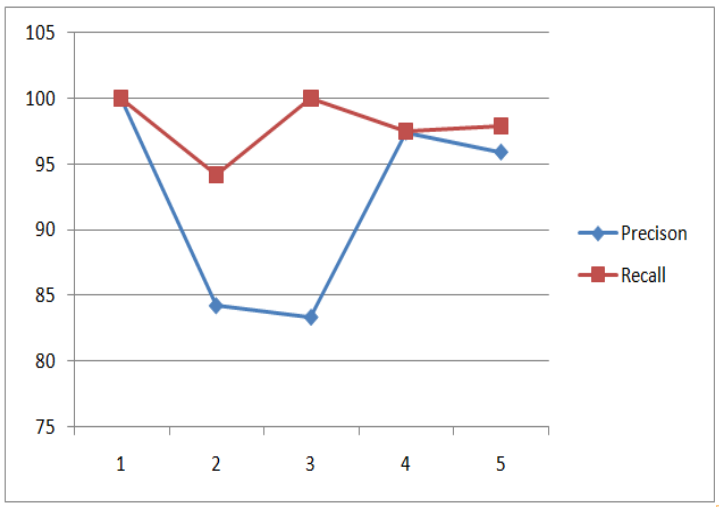

Fig. 3: Performance comparisons of precision and recall evaluation.

By observing the above plot, observe that both precision and recall are achieved approximately 90% of performance accuracy which is the good sign of the proposed system using graph-based approach for machine translation process.

CONCLUSION

This Research work addresses English to Hindi translation problem. System accepts English phrases generating Hindi translation on higher accuracy and fluency. The quality of translation depends on size of input of existing system, lack word ambiguity and lack rule knowledge. Search space for translating is large in existing system, which are major challenges addressed by our research work. Most of the rule based machine translation systems are having larger probability of yielding non semantic results. This is all due to missing of very much needed semantic relationship between the two vocabularies. So graph based approach of the proposed system is having an advantage of poly-directional traversing ability to identify larger semantics between the two vocabularies. This is an added advantage to increase performance of

IIOAB JOURNAL
ISSN: 0976-3104

the system to its best. And this can be analyzed in the prior section with linear rule based system of machine translation. The analysis of the system performance can be done by the Precision and recall method which observe that both precision and recall are achieved approximately 90% of performance accuracy which is the good sign of the proposed system using graph based approach for machine translation process.

Graph based machine translation system can be enhance in the future by adding large and very large phrase conversion techniques using huge databases in distributed systems and cloud computing paradigm. The work can be extended to include multilingual corpus of different languages in the source-target pair. The target and source languages can be increased from present one language. The system can also be put in the web-based portal to translate content of one web page in English to Hindi. A mobile application can also be developed in which message containing English text is sent to the client in Hindi language. The translated text can be reordered and processed to overcome grammatical mistakes which will be part of post-processing. This will improve score of human evaluation. What has been achieved by above work:-

- Accuracy and fluency in balance way is been achieved which existing system lack.
- Graph based approach enhances search process.
- First of kind work to model Machine translation in graph scenario.
- As system data increases precision enhances.

CONFLICT OF INTEREST

None

FINANCIAL DISCLOSORE

None

ACKNOWLEDGEMENTS

To prepare this system of-" A Heuristic approach for Graph-based Machine Translation" has been prepared by Miss. Priyanka Malviya and Prof.Gauri Rao. Priyanka would like to thank her faculty as well as the department, parents, and friends for their support and confidence in her.

REFERENCES

[1] Antony PJ. [2013] Machine translation approaches and survey for Indian languages. Computational Linguistics and Chinese Language Processing 18(1): 47-78.
[2] An Effective Knowledge base system Architecture and issues in representation techniques, https://pdfs.semanticscholar.org/7951/1e9a147e44bef8ae615b2575790c283dba01.pdf[online], accessed Dec 2016.
[3] Stein A, http://stp.lingfil.uu.se/~joerg/mt09/f2_RBMT_eval-2x2.pdf[online], accessed Dec 2016
[4] Takashi W, Motoda H. [2003] State of the art of graph-based data mining. Acm Sigkdd Explorations Newsletter 5.1: 59-68.
[5] Hand DJ, Mannila H, Smyth P. [2001] Principles of data mining. MIT press,
[6] https://www.surveysystem.com/correlation.htm[online], accessed Jan 2017
[7] Introduction to data mining, https://onlinecourses.science.psu.edu/stat857/node/3, accessed Jan 2017
[8] Shachi D, Parikh J, Bhattacharyya P. [2001] Interlingua-based english–hindi machine translation and language divergence. Machine Translation 16(4): 251-304.
[9] Sinha RMK, Jain A. [2003] AnglaHindi: English to Hindi machine-aided translation system. MT Summit IX, New Orleans, USA ,494-497.
[10] Ananthakrishnan R, et al. [2008] Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. International Joint Conference on Natural Language Processing
[11] Chatterjee N, Goyal S, Naithani A. [2005] Resolving pattern ambiguity for english to hindi machine translation using WordNet. Workshop on Modern Approaches in Translation Technologies.
[12] Ondrej B, et al. [2014] HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation. LREC
[13] Sinha R. Mahesh K, Thakur A. [2005] Machine translation of bi-lingual hindi-english (hinglish) text. 10th Machine Translation summit (MT Summit X), Phuket, Thailand (149-156.
[14] Dwivedi SK, Sukhadeve PP. [2010] Machine translation system in Indian perspectives. Journal of computer science 6(10): 1111.
[15] Das A, et al. [2009] English to Hindi machine transliteration system at NEWS 2009. Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. Association for Computational Linguistics,
[16] Ananthakrishnan R, et al. [2006] MaTra A practical approach to fully-automatic indicative English-Hindi machine translation. Symposium on Modeling and Shallow Parsing of Indian Languages (MSPIL'06).
[17] Rao D, Bhattacharya P, Mamidi R. [1998] Natural language generation for English to Hindi human-aided machine translation. VIVEK-BOMBAY- 11: 32-39.
[18] Carpuat M, Wu D. [2007] Improving Statistical Machine Translation Using Word Sense Disambiguation. www.cs.ust.hk/~marine/papers/CarpuatWu_EMNLP2007.pdf
[19] Gupta D, Chatterjee N. [2003] Identification of divergence for English to Hindi EBMT. Proceeding of MT Summit-IX.
[20] Akshar B, et al. [1997] Anusaaraka: Machine translation in stages." VIVEK-BOMBAY- 10: 22-25.
[21] Visweswariah K, et al.[2010] Syntax based reordering with automatically derived rules for improved statistical machine translation. Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics,.
[22] Naskar S, Bandyopadhyay S. [2005] Use of machine translation in India: Current status. Phuket, Thailand: Proceedings of MT SUMMIT X 13-15.
[23] Salunkhe P, Bewoor M, Patil S. [2015] A Research Work on English to Marathi Hybrid Translation System. (IJCSIT) International Journal of Computer Science and Information Technologies 6(3): 2557-2560.
[24] Salunkhe P et al. [2016] Hybrid machine translation for English to Marathi: A research evaluation in Machine Translation :(Hybrid translator). Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on. IEEE.
[25] Graph Database, https://www.fbi.h-da.de/fileadmin/personal/u.stoerl/DBAkt-WS1415/Vorlesung/DBAkt-WS1415-GraphDB-10-Introduction.pdf[online], accessed Mar 2017.
[26] Graph database, http://mayor2.dia.fi.upm.es/oeg-upm/files/eswc2014/Tutorials/SDMinGraphDataBases/SlidesTutorialGraphDatabases2014.pdf[online], accessed Mar 2017.

ISSN: 0976-3104

[27] Malviya P, Rao G. [2017] Amplifying Statistical Machine Translation (SMT) Model Using Graphm Based Approach: A Novel Idea, International Journal of Computer Science Trends and Technology (IJCST) 5 (2): Mar –Apr 2017.

[28] Kasthuri M, Britto S, Kumar R. [2014] Rule Based Machine Translation System from English to Tamil, World Congress on Computing and Communication Technologies ,2014.

[29]Malviya P. [2016] A Study Paper on Storage Area Network Problem-Solving Issues. International Journal of Computer Science Trends and Technology, 4(4): 151–156.

[30]Malviya P, Rao G. [2016] A Model Literature Analysis on Machine Translation System Finding Research Problem in English to Hindi Translation Systems, International Journal of Control Theory and Applications, (9):43, 361–370.