Retrieval of Multiple Language Text from Webpages using Big Data

[1] Charu Jain, [2] Aarti Chugh
[1,2] Assistant Professor, Department of Computer Science, Amity University Haryana

Abstract— With an extensive improvement in network infrastructure where information is flowing across national boundaries there is a need to handle language barriers. Many algorithms are designed for translation of the information on a webpage into different languages. This paper lays out the concept to improve the translation process of the available webpages information for multiple languages by the means of application of data mining and big-data technology.

Keywords—Big Data; Web Search Engine; Crawlers; Searching; Machine Translation; Phrase-Based Statistical Machine Translation; Noise Channel model; Bilingual Evaluation Understudy; Finite Automata

I. INTRODUCTION

The paper emphasis on the efficient use of parallel text and big data technology to translate the available web information for the multiple languages. For example translating from English to Hindi sometimes turned to be the incorrect translation which may affect the official government works. The need of effective translation can be fulfilled by scanning and computing the smart algorithms. These algorithms definitely comprise the volume, velocity and variety of Big Data. Traditionally search engine analyses the patterns from the large amount of text. Then these patterns are processed traditionally with the help of three techniques namely 1. Web crawling 2. Indexing 3. Searching. But for an efficient translation Bilingual Evaluation Understudy (BLEU) algorithm will be followed through a fuzzy approach where possibilities other than 0 and 1 can be considered. This can be implemented by applying BLEU algorithm on the set of English sentences E to convert it into another set of language Hindi sentences H. This will give all the possible correct translation from the data sources.

II. METHODOLOGY

A. Scanning Data

The first step involves the data pruning using the Noisy Channel Method. Traditionally the set of strings are extracted from the set of $\sum$. Here the set of incorrect strings will not be accepted by the $\sum^*$. Translation brings the number of permutation and combination of the strings. The incorrect translated text can be pruned using noisy channel method.

B. Knowledge Discovery using Big Data

Since the data is increasing exponentially and thereby also increasing the complexity to search the patterns. Hence making more difficult to design the algorithms. But there are examples like LinkedIn wherein the Big Data technology is creatively implement to generate the recommendations on the right side of the web pages. Similarly a semantic rule can be designed by learning a standardized Decision Tree. The decision tree induction will provide the platform to translate the multiple languages. The data preparation step is a crucial step as it forms the basis of the translation. The particular Grammar of that language can be included as an algorithm and this set of rules will help to clean and prepare data for the translation. For example applying algorithm on one set of data English, E = {E1, E2, E3,…….,En} will convert it into another set of data Hindi H = {H1,H2,H3,…….,Hn}.

C. WEB SEARCH ENGINE

The mechanisms under which the web search engines work should be modified for the translation. The key changes that can be outlined in the mechanism includes the crawling of the speeches with addition to the text from the various Big Data technologies like Hadoop so that an improved meaning can be derived for the particular strings of data. The next step during the knowledge discovery is indexing where the modelling should be done through analyzing only that content which are meaningful.

This should be automatically saved for the future reference. Then the best matching keywords – the set of correct strings should be indexed and searched simultaneously so that it can display the desired outputs. These outputs will certainly be the outputs of a human intelligence.

D. Association Rule

Normally crawlers are designed to search the content through spider web approach. This approach can be boosted by having the support and confidence of the translated text. It will add the human intelligence since it follows what the public or the language interpreters may predict. The computed support and confidence should be of the pruned data.

E. Models

- To compute the correct support and confidence it is required to have a proper set of data with minimum cost. For this the data warehouse model Fact constellation should be used. Since it will share the common semantic rules for the two languages.

- Once the fact constellation is achieved, meta-data can be derived out of that. This will show the transformations, relationship history of the grammars of the two languages. Once the meta-data shows the grammar, the best possible algorithms can be designed for the effective translation.

- Flexibility should be allowed in the designing of the fact constellation since data increases exponentially. It enables the correct meta-data repository from the huge amount of data.

- The private files of the web pages should not be crawled. It is recommended to use robot.txt to protect the data from indexing.

F. Mathematical Interpretation

Errors can automatically be corrected using decision function tree. These decision tree provides the direction for the tree traversal. Assuming that P is the probability of the occurrence of errors.

Let E be the set of English data set where E = {E1, E2, E3,…….,En}.

Let H be the set of data Hindi where H = {H1, H2, H3,…….,Hn}.

Now the probability for this matrix will be
Probability(G|I) = ∑P(H1, H2, H3,…….,Hn) | ∑P (E1, E2, E3,…….,En) (1)

Where G is the goal state of the language and I is the initial state of the language.

Here the Goal state language is Hindi and the Initial state language is English

Let us extend the equation so as to understand the efficient translation from English to Hindi.

G. Error Correction

- Let the set E comprises of both capital and small letters i.e. E = { A,B,C,D,……Z,a,b,c,d,……z }
- And the new set H comprises of all the possible letters, symbols H= { H1,H2,H3,….,Hn,h1,h2,h3,……hn }
- We know that G is the Goal State but while conversion or translation there may be some challenges. The first issue is missing data. Missing data usually occurs if no information is provided for several item sets.
- Duplication of data can also occur. To avoid these issues a special test case should be generated and these test case should be updated in the meta-data repository so that required translation can be formulated next time.
- Data entry errors like misspellings, overlapping, and inconsistent data are the problems which are not available at schema level. The big challenge is Data Quality. The solution is to divide it into two types of problems that is Single Source Problem and Multi Source Problem where schema and instance level problem are defined at individual level.
- Sometimes algorithms too have some duplications, to remove that it is recommended to merge together the sorted data. The database of matched variable should be counted by a COUNT variable and these value can be stored in COUNT so that total number of duplicated algorithms (codes) can be computed.

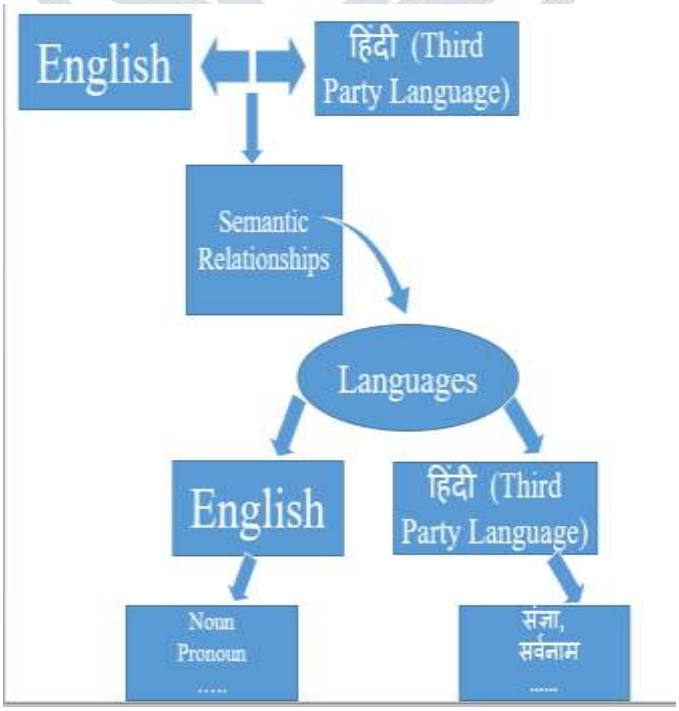

Figure 1. Illustration to map two languages using semantic relationships

- The human error is also predictable, the best way to correct is to fix it as soon as possible by the spell checking technology based on the database available for a stand dictionary.
- Many popular translators use the pivot concept where the database of three languages are compared. English acts as an intermediator while translating the two languages like Hindi and German.
- Pivoting improves the Statistic Machine Translation. Later the system was termed as MERT (minimum error rate training) which also increased the BLEU points.
- All the above suggested measure are not enough for a complete perfect translation. Because it does not solves the problem. To give a boost to the problem or to give a concrete solution for the effective translation, it is required to maintain a new standards in the database. Thus giving the programmers to minimize the time and complexity to design such solutions or software for the translation.
- The main reason to prefer English and to be dependent on it for the programming lies in the root of it. The reliability of developers on English is due to the fact that all the English alphabet along with the special symbols are included in the ANSI code.

III. CHALLENGES AND SCOPE

Unlike English the concept of formulizing the ANSI Code for Hindi or any other language will initially not acceptable. There must be different views and opinions regarding the time and space complexity and cost to design the complete framework.

No proper rules or Grammar had been defined to convert the data set {E} into data set {H} or any data set of foreign language. A prototype for this language should be formulated as early as possible. An attempt ANSI Code

	A	B
1	**Machine Code**	**Hindi Alphabet**
2	0001 -	अ
3	0010 –	आ
4	0011 –	इ
5	0100 –	ई
6	.....	.....
7	.....	.....
8	.....	.....
9	.....	.....
10	.....	.....

Figure 2 Prototype to implement other language (Hindi) in the memory

The prototype is designed to give the rough estimate of the time and space required for the accurate translation. The emphasis on the consideration of Hindi alphabets and its symbol is also important for a new revolution in programming. The new database will create the precise grammar and the corresponding syntactic pattern.

A. A platform to accommodate Sanskrit as a new programming language.:- The concept of the revolution in the field of programming and the summarized challenge and scope are explained in the following manner :-

- *LOC :* The more number of LOC (Lines of Source Code), the more complex the program becomes. Definitely the research & development on implementing Hindi in the ANSI code will open the gates for Sanskrit. Thereby reducing the LOC's of many complex and big algorithms.

- *Sanskrit as a programming language, Understanding Sanskrit from Hindi Deletion:* The revolution is to bring a proper formulated grammar which can reduce the time and space complexity at a very large extent. This is possible with the use of Sanskrit. Reducing the time and space complexity of big algorithms will save the huge costs in designing software.

- *Need of Artificial Intelligence:* Select The converted set of data (knowledge) from English {E} to Hindi {H} This comverted data using Channel Noisy method with a fuzzy approach will help to process a new output that is information. The example may also include the speech recognition feature.

- *Digital India:* Simultaneously the vision for India of Digital India is only and only possible if all the web information is available in Hindi. This will also improve the big data analysis for Indian Buisness Industry. According to the 2001 census for the population of 1,028,610,328 around 422,048,642 are the Hindi speakers which is 41.03. The Digital India supporting organization like Google, Facebook can implement it in the similar approach to reach out to 40% user using Hindi.

IV. CONCLUSION

The methodology described above illustrates how big data can be used for multiple language text retrieval from web pages. Although, there are numerous challenges but the technique can be further combined with neural networks or other existing solutions to get an efficient translation algorithm.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jiane "Data Mining Concepts and Techniques" Elsevier, Third Edition.

[2] Alex Berson, Stephen J. Smith 2004 "Data Warehousing, Data Mining and OLAP", The McGraw-Hill .

[3] Takako Aikawa, Maite Melero, Lee Schwartz, Andi Wu "Generation for Multilingual MT" Microsoft Research Paper, One Microsoft Way.

[4] Tomas Mikolov, Quoc V. Le, Ilya Sutskever 2013,"Exploiting Similarities among Languages for Machine Translation" Journal of Artificial Intelligence Research, 44 (2012) 179-222 .

[5] Preslav Nakov Hwee Tou Ng 2014 "Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages" IEEE transactions on knowledge and data engineering, vol. 26(1).

[6] Germán Aquino, Waldo Hasperué, César Estrebou and Laura Lanzarini 2013 "A Novel, Language-Independent Keyword Extraction Method" in XIX Congreso Argentino de Ciencias de la Computación, 978-987-23963-1-2.

[7] Karthik Visweswariah, Jiri Navratil, Jeffrey Sorensen, Vijil Chenthamarakshan, Nanda Kambhatla "Syntax Based Reordering with Automatically Derived Rules for Improved Statistical Machine Translation 2010", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1119–1127, Beijing.

[8] Anthony Aue, Arul Menezes, Bob Moore, Chris Quirk, Eric Ringger 2004 "Statistical Machine Translation Using Labeled Semantic Dependency Graphs", Microsoft Research, 1 Microsoft Way, Publisher ACL/SIGPARSE.

[9] S. M. Fakhrahmad, A.R. Rezapour, M.H. Sadreddini, M. Zolghadri Jahromi, 2012 "Machine Translation Based on Data Mining and Deductive Schemes", Proceedings of the World Congress on Engineering 2012 Vol II, WCE 2012, July 4 - 6, London, U.K..