ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 7, July 2016

Analysis on the Translation of Data While Storing In Big Data Systems

[1] Gouthami Kumari G S [2] Saritha M V [3] Suneetha K R
Department of Computer Science and Engineering,
Bangalore Institute of Technology, VV Pura, Bangalore 560004, India

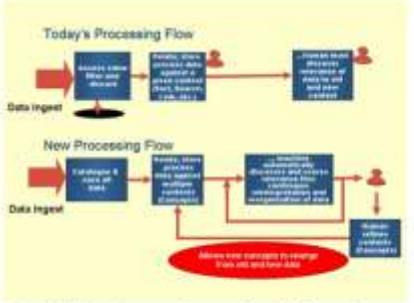
Abstract- **Big Data is a term that satiates the inadequacy of traditional data processing of large or complex data applications. Big Data processing systems must deal with very high data ingest rates (velocity) and massive volumes of variety of data. Most data translation approaches are based primarily on the replacement or elimination of human interference with automation in key areas of information processing on the voluminous, varied variety of data arriving at high velocity. While applying Knowledge Engineering (KE) to Big Data applications, the major challenge is that the data is either structured, unstructured or semi- structured. Data is written in multiple languages using characters, each character has different encodings when text is exchanged between the customers and the Enterprise. Several linguistics structures of the text, keyword taxonomies, content categorization, language translation, context and temporally-based information retrieval areas are being considered in Big Data management. The challenges include how to capture, transfer, store, clean, analyze, share, secure and visualize the data. This paper provides survey on the pre-processing methods used in the data translation such as various Statistical Machine Translations.**

Key Words**: Statistical Machine Translation, Rule based and Topic Based SMT, Cross Language Information Retrieval, Word Mapping**

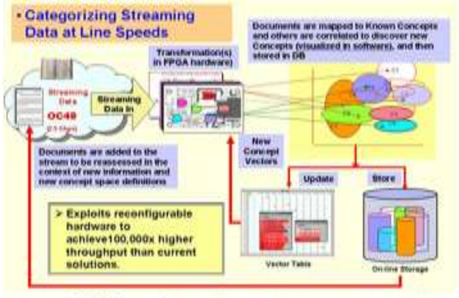
I. INTRODUCTION

Big Data can be processed as and when the live real- time data arrives. As new contexts emerge by using mathematical transformation algorithms on the data, the data streams can be re-processed. Two methods for processing data flows analysed are “Track before Detect” Information Management and “Stream data in real-time into machines that self-organize concepts from input semantics”. [1]

In the “Track before Detect” methodology, the data being gathered is filtered and only the information pertaining to the immediate contention is related, queried, tracked and saved accordingly as illustrated below. The Data is processed to find the relation between the information as per the human recognized concepts.



Track before Detect Approach to Information Management


• Categorizing Streaming Data at Line Speeds
Documents are mapped to Known Concepts and others are correlated to discover new Concepts (visualized in software), and then stored in DB
Transformation(s) in FPGA hardware)
Streaming Data
OC48
Streaming Data In
Documents are added to the stream to be reassessed in the context of new information and new concept space definitions
New Concept Vectors
Update
Store
➢ Exploits reconfigurable hardware to achieve100,000x higher throughput than current solutions.
Vector Table
On-line Storage

Self Organize concept-based streaming data processing

In order to reduce the intellectual mismatch between Human analysts and machines, special-purpose computing machines are designed to categorize the extremely large volumes of unstructured multi-lingual text data into corresponding concepts, using the “Stream data in *real-time* into machines that self- organize conceptsfrom input semantics”innovation. This leads to the improvement in the information management of Big Data.

Here, the Big Data Streams are transformed by high-speed semantic processing circuits into points in a multi- dimensional space, containing similar information. The documents with similar concepts are clustered into similar regions of the multi-dimensional space. Clustering the similar concepts self organizes the data documents (as viewed in above right figure).

Word Frequency is determined using Multidimensional Word Mapping (WM) Table, Dictionary Based WM, Pair-Wise WM. Proprietary Mathematical Transformation Algorithms like K –Means, AGS and Order is implemented to determine the concept categories.

This paper discusses on SMT of Big Data where *Section 2* presents the Literature Survey, *Section 3* discusses various frameworks of SMT, *Section 4* concludes the survey and References are found in *Section 5*

II. LITERATURE SURVEY

This section reviews a set of papers related to Statistical Machine Translations. Topic-based coherence models are implemented by authors Deyi Xiong and et al [7] to produce coherence for document translation, in terms of the continuity of sentence topics in a text. Adoption of a maximum entropy classifier predicts the target coherence chain coherencechainforeachsourcetexttobetranslated

Domain adaptation in phrase- based SMT via topic modeling is considered as per authors Gong Zhengxian and et al[8]. The proposed topic modeling approach employs one additional feature function to capture the topic inherent in the source phrase and help the decoder dynamically choose related target phrases according to the specific topic of the source phrase. Integrating a topic model into SMT, not only enhances the ability for domain adaptation but also avoids the linear growth of parameter numbers for interpolation of sub-models.

Evaluating different architectures based on Moses in the web environment for handling the translation between different language pairs is proposed by Francisco Oliveira and et al[9]. Translation requests can be passed to different MT systems and the best result based on some evaluation algorithms need to be obtained. An extra module can be added in classifying the domain of the source sentence and sending the server with appropriate resources targeted for that area. As a result, more than one MT system targeted for the same type of translation can be used using different resources.

The proposed lemma translation with generated surface form and additional post-process is implemented by authors Mohammad Anugrah Sulaeman and et al[10]. Lemma translation uses lemma and POSTAG of word in its translation process. There are many methods that is used for decoding process, e.g., heuristic search, A* search, beam search. These methods show an improvement over the existing system with a 116% increased BLEU score on Japanese to Indonesian translation and 26% increased BLEU score on Indonesian to Japanese translation.

Based on HNC (Hierarchical Network of Concepts) theory, Runxiang Zhang and et al [11] study the method based on rules for identification and transformation of comparative sentences in patent texts leading to improved performance of patent machine translation effectually. Data is categorized into comparative sentences of different types, and then pattern discovery is made. Classification algorithms such as SVM and CRF are used. Evaluation using the closed test showed its effectiveness.

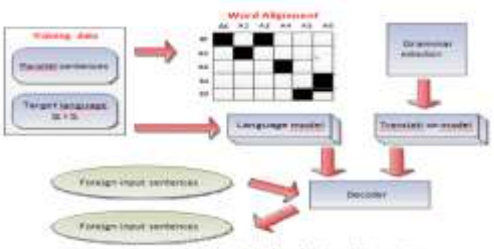
Authors Lei Chen and et al [12] Combining a rule-based reordering model along with the conventional dependency parsing on the text leads to the asymmetry of morphological structures and word ordering alleviation between source language and target language. Experiments show that this method combination can improve the performance of the statistical machine translation system, and by bettering the accuracy of dependency parsing will affect the final translation directly.

Rahul.C and et al [13] rule based reordering approach is incorporated and morphological information for English to Malayalam statistical machine translation. Their approach avoids the use of parsing for the target language (Malayalam), making it suitable for statistical machine translation from English to Malayalam, since parsing tools for Malayalam are currently unavailable They (i) reorder the English source sentence according to Malayalam syntax, and (ii) use the root suffix separation on both English and Malayalam words. By applying simple modified transformation rules on the English parse tree(given by Stanford Dependency Parser) reordering is performed and is succeeded by a morph analyzer.

Amin Mansouri and et al [14] have innovated a noisy filtering system based on MaxEnt classifier to distinguish between correct and incorrect sentence pairs. Several variations on SMT, such as hybrid MT or statistical post editing MT, baseline SMT, Factored SMT, Verb-aware SMT and statistical post editing of an existing rule-based MT has been proposed in this paper.

III. FRAMEWORKS OF SMT

In the process of statistical machine translation (SMT) of Big Data, two modules, that is, module 1: constructing the target- language sentence, and module 2: translating from target texts to source texts language are considered. The SMT ideology is in that the sentence with highest probability is selected by optimizing an objective function, which describes the relationships between source language sentences and target language sentences. The introduction of the SVM classifier [2] to solve the statistical machine translation (SMT) for English language has been proven to outperform Topic-Based SMT and Rule- Based SMT. Different from the traditional methods, the proposed Classification algorithm is implemented in SMT.



Framework of the statistical machine translation system

SMT lies in that the sentence with highest probability is selected by the following equation:

$\widetilde{E}=\operatorname{argmax}P(EC)$ E: English C: Chinese

E

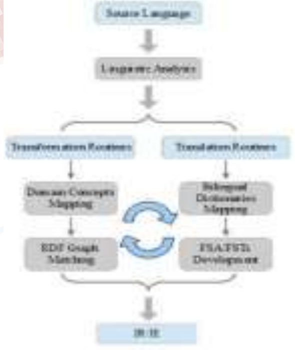
To solve the Big Data classification problem, SVM is exploited, and SVM aims to solve the classification issue by the following equation.

$$\{w x_i + b >= 1, y = 1 \; \{w x_i + b <= 1, y = -1$$

In further study by Silviu Paun [3], an analysis is performed on useful correlations between event-related data in Microblogs and on the data coming from various sensors, to identify data patterns. The data correlations is observed from Microblog platforms and IoT sensors in a large scale system, a scale which simulates the dimensions (3Vs: volume, velocity, variety) of Big Data.

The a priori modeling of the engineering problem of interest is crucial for both (1) efficient (rapid) collection, representation, and structuring of domain knowledge; and (2) the proper integration of domain knowledge with analytical KE methods in order to facilitate the extraction of useful knowledge. In order to convert the free-form text provided by Enterprise applications, with respect to failure modes, two methods from engineering design, the Function Analysis System Technique (FAST) and Failure Modes and Effects Analysis (FMEA) are implemented, to provide the necessary domain knowledge model. This model then drives the collection, representation, and structuring of the failure modes for the product of interest as the class labels when applying data mining classification techniques. The use of these domain models using failure model to drive the knowledge engineering process, enables causality to be naturally incorporated and once the structured data is derived by mentioned failure modes, it is relatively straight-forward to process the data by Enterprises.

The Lexicon-Grammar (LG) methodology for the development of an ontology-based Cross-Language Information Retrieval (CLIR) application, achieves the translation of Natural Language (NL) queries in any language by means of a knowledge-driven approach allowing to semi- automatically map natural language to formal language(FSA), for reducing the human- computer interaction and interference by humans. The human interference dilutes the scope of proper data translation. CLIR allows mapping of data and meta-data to implement specific domain prototype system. CLIR application translates the queries across the languages desired and searches for the required information on the web. Subsequently it translates again all the information and the documents detected in this way into the user's preferred language. Bidirectional Mapping of data and query is done. The Electronic dictionaries (simple word or compound word dictionary types) are used to develop local grammars in the form of FSA/ FSTs(Finite State Automata/Finite State Transducers). A pre-processing phase converting natural language strings into reusable linguistic resources is performed, to extract information from free-form user queries, which is used to match with already available ontological domain conceptualizations. (Can be viewed below)



CLIR System Workflow

The large quantities of knowledge coverage and deep semantic relations are preserved which may interlink different languages during CLIR.

In the usage of OODB(Object Oriented Database Model)model in Big Data systems, data is represented at different levels of vagueness of uncertainty of the source providing the data and imprecision of the data. The Uncertainty during class definition consisting of uncertain attribute values and fuzzy types with intervals are dealt with. The Imprecisions are handled at first and second level of precisions. [6]

IV. CONCLUSION

As per the analysis conducted on the various proposals and implementations, it is observed that the translation of data streams in Big Data is governed by the human interaction and extent of human interference. By automating the translation mechanisms to detect the similarity in data by clustering them into concept groups, it is possible to query the Big Data in multilingual forms and at a faster rate using CLIR. The scope for improvement in faster processing, accuracy and security of the data translation of Big Data is open.

REFERENCES

[1] Stephen G. Eick, John W. Lockwood, Ron Loui, James Moscola, Doyle J. Weishar "Transformation Algorithms for Data Streams", IEEEAC paper #1633, Version 5, Updated December 5, 2004

[2] Jia Tan,Wang Chao, " Data-English Language Statistical Machine Translation Oriented Classification Algorithm" , International Conference on Intelligent Transportation, Big Data & Smart City, 2015

[3] Silviu Paun , "Pattern Discovery in Big Data Streams"

[4] T Munger S Desa, C. Wong , "The Use of Domain Knowledge Models for Effective Data Mining of Unstructured Customer Service Data in Engineering Applications", IEEE First International Conference on Big Data Computing Service and Applications, 2015

[5] Johanna Monti Mario Monteleone, Maria Pia di Buono, Federica Marano, "Natural Language Processing and Big Data An Ontology-Based Approach for Cross-

Lingual Information Retrieval" , SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013

[6] J Blanco, N Marin, O Pons, MA Vila , "Softening the Object Oriented Database Model : Imprecision, Uncertainty and Fuzzy Types", IEEE Conference, 2001

[7] Deyi Xiong, Min Zhang, Member, IEEE, and Xing Wang, "Topic-Based Coherence Modeling for Statistical Machine Translation", ACM Transactions on Audio, Speech March 2015

[8] Gong Zhengxian, Zhou Guodong, "Employing Topic Modeling for Statistical Machine Translation", IEEE 2011

[9] Francisco Oliveira, Fai Wong, Sam Chao, Pui-Chi Fong, "Design of Web based Machine Translation Environment for Multi-languages based on Moses", International Conference on System Science and Engineering, Macau, China - June 2011

[10] Mohammad Anugrah Sulaeman, Ayu Purwarianti, "Development of Indonesian- Japanese SMT Using Lemma Translation and Additional Post – Process", The 5th International Conference on Electrical Engineering and Informatics August 10-11, 2015, Bali, Indonesia

[11] Runxiang Zhang, YaoHong Jin, "Identification and Transformation of Comparative SentencesinPatentChinese-EnglishMachineTranslation" International Conference on Asian Language Processing, IEEE 2012

[12] Lei Chen, Miao Li, Maintao He, Hui Lui, "Dependency Parsing on Source Language with Reordering Information in SMT", International Conference on Asian Language Processing, IEEE 2012

[13]Rahul.C, Dinunath.K, Remya Ravindran, K.P.Soman, "Processing For English-Malayalam Statistical Machine Translation "

[14] Amin Mansouri, Heshaam Faili, "State-of-the-art English to Persian Statistical Machine Translation System ", The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), IEEE 2012