An n-gram Approach Model for Language Detection & Translation System (LDTS)

[1]Ganesh Yenurkar, [2]Rajesh Nasare, [3]Sushil Chavhan, [4]Roshan Bhanuse,
[1,2,3,4]Assistant Professor
[1,4]Department of Computer Technology
[2]Department of Computer Science and Engineering
[3]Department of Information Technology
[1,3,4]Yeshwantrao Chavan College of Engineering, Nagpur, India

Abstract: Languages change over time, through the succession of generations, new words are used or developed, pronunciations vary, and morphology develops. Developing a system that finds phrases that have been replaced over time in a specific language is an area of interest because it can allow new generations to understand the old texts easily and without the need to neither experience old terminology nor search in old-fashioned words dictionaries. With the increasingly widespread use of computers And the Internet in India, on web there are large amount of information available and it is less. So that the translation becomes essential in India. Therefore, we will try to present some kind of improvement in the information system. Implementation of efficient language detection and translation application will solve this language barrier problem. Using this Language detection and translations system (LDTS) application, we can identify the language of the text which provide information to potential readers and therefore improve the flow of ideas. This translator provides instant translations between different languages. Such a translator system will be universally helpful, regardless of the language in which it's written. So, in this paper we consider the statistical machine for our n-gram approach.

IndexTerms **- n-gram approach, LDTS**

I. INTRODUCTION

Language is usually considered as an important asset of communication in every organization till date. Why is language so important? This question has multiple answers. Information gets extracted from from one's personal decisions taken on the basis from data. And the data is considered an important aspect in the individuals life. Languages change over time, through the succession of generations, new words are used or developed, pronunciations vary, and morphology develops. Developing a system that finds phrases that have been replaced over time in a specific language is an area of interest because it can allow new generations to understand the old texts easily and without the need to neither experience old terminology nor search in old-fashioned word dictionaries.

Communication is most important advances of our time is that we experienced. The most important thing of communication is the language that is considered to remain away from these advances. The language problem is one of the most important problems to be solved on each passing day in the globalized world. Despite increase in the number of available documents, unfortunately, there is no opportunity to use the language of unknown documents. In order to make available sources of information more useful, language identification and language comprehension have significant duty. Therefore, for everyone's language identification becomes an important task communication purpose.

As the web grows, language identification and its translation in general is becoming an important issue. Web environment provides enormous number of documents, often in the other languages which is not known to the user, which makes the task of identification uneasy. N-gram based approach, which is the basic method for text categorization, could be used with slight modifications to perform language detection. The main task of our LDTS system is to identify the language based on the text uploaded by the user. It will help the researchers, student, and teachers to study the topics thoroughly. Translation becomes successful only if the communication of the meaning of a source-language text by means of an equivalent target-language text.

Here we focus on the identification of document-based language in our study. Language identification approaches are divided into two methods: linguistic methods and statistical methods. Linguistic method which is approach in language identification estimates the language in the documents according to the grammar rules belonging to languages. The document-based approach in language identification, one of the linguistic method estimates the language according to the rules of grammar in language documentation. It makes searching according to the frequency of searches for words in the document and makes scoring them. The availability of the automated translation system makes it possible to translate an entire corpus into a new language. This paper shows that it is effectively feasible if input text is having maximum characters. The quality of the translation depends on the amount of manually transcribed data used for training the automatic machine translation (MT) component.

In this work, we propose a statistical machine translation model that automatically converts the original form of text to the desired language form, which in turn are used for building more efficient translation application.

II. LITURATURE SURVEY

The only difference we noticed, if any, is that our translator seems to be slightly better at certain expressions, providing user with the correct translation instead of a meaningless word-by-word equivalent. Our aim is to aid the users with translations when working on a document. [1]

Our translation tools have a special feature that is auto-detection. It enables you to auto-detect the language more precisely that needs to be translated in case user cannot identify the language. [1]

In our LDTS system some errors may be expected if and only if the sentences or phrases that are being translated are complex. But the accuracy and speed will be a big plus point to look out for LDTS.
The biggest advantage is that it will be an offline translation and can be accessed anywhere without the internet. Anyone can access it anywhere as they need [1].

Automatic Error Detection and Hint Generation methods described by Sychev O. A., Mamontov D. P. for detecting mistakes and generating feedback using edit distances, tree analysis and natural language generation and compares it with brute-force approach [18]. This method was implemented in the question type plug-in Correct Writing, developed by the authors of the article. To correctly divide the teacher's answer and student's response into lexemes, the teacher must select a language, in which the answer is expected. At the moment, C, C++, and English language are supported. The application of that method of text generation was also make possible to generate messages about misplacing or omitting of not individual lexemes only, but any non-terminal symbols of the grammar of the language being learned.

Meredita Susanty, Sahrul *et.al.* were developed an artificial neural network model for not only classifying the words as (non) offensive words but also considering the structure of the sentence to get its context. The challenges were informal grammar and word abbreviation used in social media. Hence, there were noise elimination and normalization processes to address these challenges. Their research shown a computer simulation results show excellence accuracy of 99.18% training, 94.28% validation, and 96.8% testing, only by utilizing the sigmoid activation function. This model could assist government enforcing the information and electronic transaction law and decreases the number of disputes due to aspiration freedom abuse in social media [19].

The issues on information assurance and security was discussed by Mrs. Thejaswini S. and Indupriya C [20]. They discovered detection of Advanced Persistent Threat (APT) in DNS and vulnerability analysis. The goal of this paper was to provide the overview of how natural language processing could be used to address cyber security issues.

Muhammed Enes Almahdi ; Yusuf Sinan Akgül *et.al.* were proposed a simple and effective dictionary-based validation method. They were applied their approach to the Turkish language. As a dataset, they used the Turkish parliamentary minutes from 1920 to 2014. They accomplish promising results; %54.76 accuracy measured by their validation method on words with at least %0.01 probability. In that work, they shown that the ability to detect over time words substitutions within the same language in an unsupervised way, i.e., without using any time-based parallel corpora, especially morphologically-rich languages. Using adversarial training, they could learn the mapping between time-based and time-independent word embedding spaces, which allow to apply an energy function to score the similarity of words over time[21]. Their proposed model was employed an adversarial training procedure that would learn how to align between time-based word embedding spaces and time-independent global word embedding space, within the same language.

TABLE 1:
Comparison between Google translator and our system same language.

Google Translator	**Proposed System**
Web based system	Desktop based
Needs internet access every time.	Internet access does not require.
Does not has document browsing function.	It has document browsing function
User must log in to internet connection every time when he wants to translate.	User must download it once only.

Suraj Sharma, Sabitra Sankalp Panigrahi et.al. proposed a design of a Topic Detector machine which combines the power of LDA and Word2Vec to detect topic from mixed text. The experiment as carried on a mixed text of English and Hindi to detect topics. The technique to kenizes the mixed text of Hindi and English and models the min to feature vector trough a process of Word2Vec. These vectors were clustered and the cluster centers are identified as the topic of the cluster of tokens[22]. A deep neural network based English Sign Language detection based on hand gestures images was designed and validated by Palani Thanaraj Krishnan; Parvathavarthini Balasubramanian et.al.[23]. The work analyzed a three-layer convolution network with batch normalization for hand Testing performance of the trained DNN gesture detection.

A peak accuracy of 100% was obtained for training process and 82% for validation process. Moreover, a test accuracy of 70% was obtained for the sample test images. A further improvement in ESL machine translation required sex peri mentation with more number of convolution blocks and using different optimizer function. Kyeong-Hwan Kim, Chang-Sung Jeong et.al. were proposed the Korean fake news detection system using fact DB which was built and updated by human's direct judgement. Their system received a proposition as an input to verify and search the related articles so that it verifies if the article found by the entered proposition can be semantically concurred with the proposition [24]. Our LDTS translator system is very easy to use and closer to human translation interface. User Interface will be more users friendly and it will be more attractive.

Limitations:

It is more efficient and accurate only when large data is given to the application for translation. It can be useful if you are trying to decipher whether a document contains particular information, but it may not be of much help if you are trying to decipher how the topic is related to the document. Hence, it cannot replace human professional translation.

III. THEORETICAL BACKGROUND

The language statistical can be expressed here, including the use of certain keywords, the order of letters and there frequencies of short words (in a combination of presence) is decisive and each language is descriptive in nature. For the extracted language features, the n-gram feature extraction method is used. The aim of this study, to date, compare of accuracy rate of the language identification approach based on classification methods, and so, reveal the most appropriate methods. For the identification of language Pre-processing and extracting features is the first to consider.

IV. METHODOLOGY, TECHNIQUES AND ALGORITHMS

An N-Grams Approach

An N-gram is an N-character slice of a longer string that is to be in consideration. Although in the literature the term can include the notion of any co-occurring set of characters in a string for example (e.g., an N gram made up of the first and third character of a word), in our study, we have use the term for contiguous slices only for a set of characters [1].

Technically speaking, one slices the string (set of character) into a set of overlapping N-grams, the LDTS system, use N-grams of several different lengths of characters simultaneously. We can also append blanks spaces to the beginning and ending of the string in order to help with matching beginning-of-word and ending-of-word situations effectively. (We will use the underscore character (—_‖) to represent blanks in our LDTS system). Hence, the word —WORD‖ would be composed of the following N-grams:

bi-grams: _W, WO, OR, RD, D_
tri-grams: _WO, WOR, ORD, RD_, D_ _
quad-grams: _WOR, WORD, ORD_, RD_ _, W_ _ _

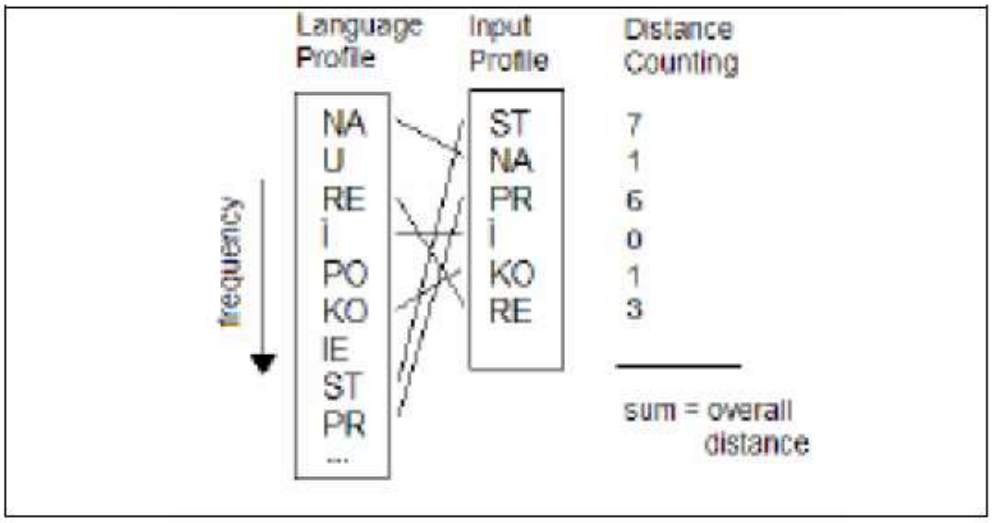

Fig 1. Shows the comparison between text and its equivalent language [1]

In general, a string of length k, padded with blanks, will have k+1 bi-grams, k+1tri-grams, k+1 quad-grams, and so on. The main advantage of LDTS system is that N-gram-based matching provides derives from its very nature specific, although, the whole string is decomposed into its small parts, any errors that are present or tend to affect only a limited number of those parts which will leaving the remainder intact. The of count N-grams are common for two strings and will get a measure of their similarity that is resistant to a wide variety of textual errors by LDTS.

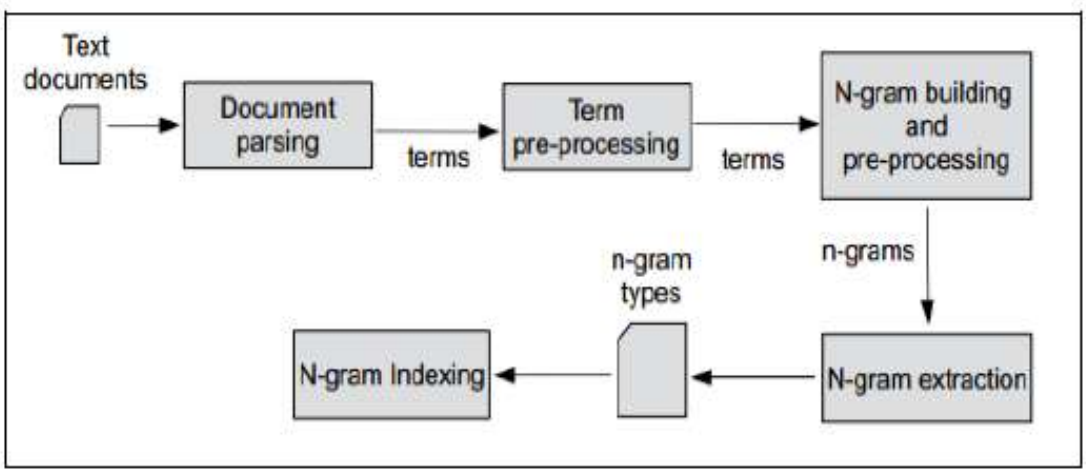

Fig 2. An architecture of an n-gram extraction framework [9]

In Figure 2, shows that the architecture of an n-gram extraction framework. This framework usually includes:
1) Document parsing – Input document terms are parsed here.

2) Term pre-processing – Here, various techniques like stemming and stop-list are applied for the reduction of terms for n-grams approach in LDTS.

3) N-gram building and pre-processing – Here an n-gram sequence of n terms is created. Not shared every time (sentences or paragraphs). It means, the last term of a sentence is the last term of an n-gram and the next n-gram begins by the first term of the next sentence in LDTS.

4) N-gram extraction – Here the removal of duplicate terms in n-grams are specified. It will result into the collection of n-gram type with the frequency enclosed to each type. For example, n-gram types with a low frequency are removed. Evidently, it is not appropriate to apply this post-processing in any application. It can be used only when we do not need these n-gram types which is not our case.

5) N-gram indexing –Here the indexing of n-gram approach shown by LDTS.

A data structure is applied to speed up access to the tuple n-gram, id, frequency, where n-gram is a key; it means the n-gram is an input of the query and id and frequency form the output. And that solution is usable for Boolean or Vector models as an input. Might be, it is necessary to create new data structures for specific text document and query models, for this we have to consider the global storage of the tuples for the whole document collection. Existing techniques of the n-gram extraction suppose only external sorting algorithms. These LDTS methods must handle a large number of duplicate n-grams which are eliminated after the frequency get computed. It results in high time and space overhead.

In this paper, we show a time and space efficient LDTS method for the n-gram extraction; here we do not consider various document parsing methods, pre-processing of terms, and n-gram building and pre-processing. We show that we can use well data structures that is known in the area of database management systems and physical database design for this task to be carried out. In this way, we utilize the same data structures for the n-gram indexing and the n-gram extraction. Additionally, we show a high scalability of our method; it is usable for large document collections including up-to 109 n-grams regardless the amount of the main memory.

V. STATISTICAL MACHINE TRANSLATION

The statistical MT models take in consideration to view that every sentence in the target language is a translation of the source language sentence with some probability in LDTS. For best translation, of course, the sentence that has the highest probability get considered. The key problems in statistical MT are: estimating the amount of probability of a translation, and efficiently finding the sentence with the highest probability [4].

We suppose that the sentence f to be translated was initially conceived in language E as some sentence e. [5] While doing communication e was corrupted by the channel to f. Now, let's assume that each sentence in E is a translation of with probability, and the sentence that we choose as the translation (ˆe) is the one that has the highest probability. In mathematical terms [Brown et al.,1990].

$$\hat{e} = \arg\max_{e} P(e|f)$$

Initially, $P(e|f)$ should depend on two factors:
1. The type of sentences that are more likely in the language E. This is called as the language model — P(e).
2. The sentences in E converted into sentences in F. This is called the translation model — P(f|e).

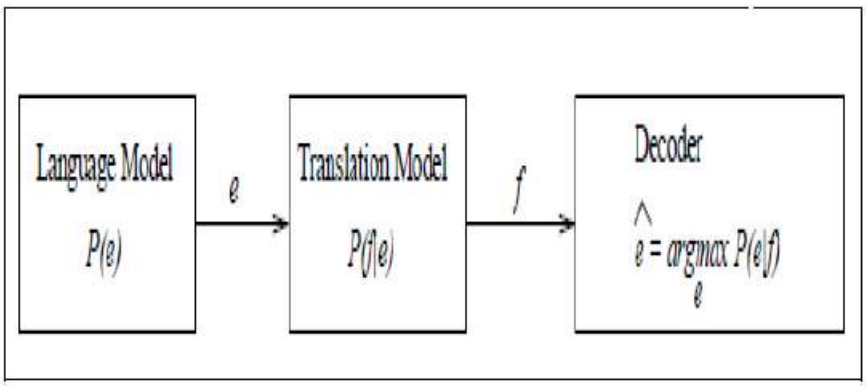

Fig 3. The Noisy Channel Model for Machine Translation [1]

$$\hat{e} = \arg\max_{e} \frac{P(e)P(f|e)}{P(f)}$$

Hence *f* is fixed, we removed it from the maximization & we will get following equation,

$$\hat{e} = \arg\max_{e} P(e)P(f|e)$$

VI. APPLICATION FLOW

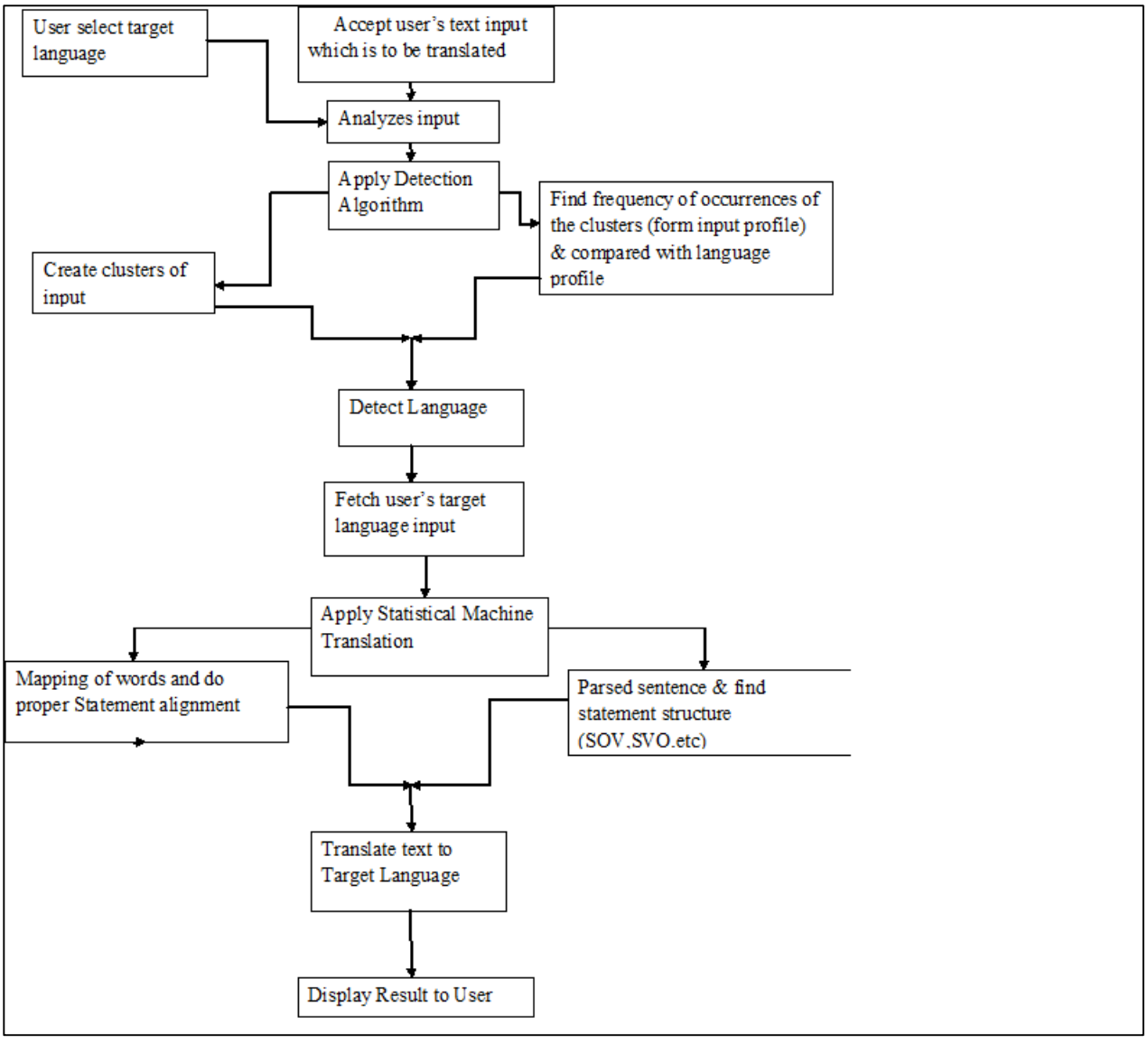

Fig 4. LDTS system FLOW

Module 1: Inputting text document to the Application

Here user can gives the text document as a input to the Application & choose the target language in which he wants to convert.

Module 2: Analysis of Text Document
Application will process the document & it will calculate the language profile needed to identify the language.

Module 3: Identification of Languages
The specific language profile calculated by the Module 2, which is used to identify the language of the text by LDTS system N-Gramm Approach.

Module 4: Translation of the Text document.
It will translate the text document to the desired language as per the user's input given to the system.

VII. CONCLUSIONS

Our method in this article introduced a translation approach to language modeling. Specifically, we used n-gram approach model to detect an input text and using statistical machine translation input text is translated into desired language as an output. The tools and methodologies which can be used to develop a translation system that translates English to other morphologically rich languages. In future, advancement of this system lies in integrating with speech recognition systems that will translates the speech into the desired lamguages.

REFERENCES

[1] Cavnar, W. B. and J. M. Trenkle : N-Gram Based Text Categorization. In: Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics (1994), 161-175.

[2] Dunning, T.: Statistical Identification of Language. In: Technical report CRL MCCS-94--273, Computing Research Lab, New Mexico State University (1994).

[3] N-Gram Based Statistics Aimed at Language Identification, Tomáš ÖLVECKÝ Slovak University of Technology Faculty of Informatics and Information Technologies Ilkovičova 3, 842 16 Bratislava, Slovak Republic olvecky@stonline.sk.

[4] Statistical Machine Translation ADAM LOPEZ University of Edinburgh

[5] Statistical Machine Translation Ph.D. Seminar Report by Ananthakrishnan Ramanathan

[6] Use of statistical N-gram models in natural language generation for machine translation,Liu, F.-H. ; Liang Gu ;

[7] Yuqing Gao ; Picheny, Michael Publication Year: 2003 IEEE CONFERENCE PUBLICATIONS Communication and Processing, 2007 IEEE International Conference on, 2007, pp. 209– 216.

[8] J.Pomikalek and P. Rychl ´ y, ―Detecting Co-Derivative Documents in ´ Large Text Collections,‖ in Proceedings of the
Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco. European Language Resources Association (ELRA), 2008, pp. 132–135. P. Xu, D. Karakos, and S. Khudanpur, ―Self-supervised dis-criminative training of statistical language models,‖ in Proc.IEEE ASRU, 2009.

[9] Index-Based N-gram Extraction from Large Document Collections Michal Kratk ´ y, Radim Ba ´ ca, David Bedn ˇ a´ˇr, Jiˇr´ı Walder, Jiˇr´ı Dvorsky, Peter Chovanec

[10] Z. Ceˇ ska, I. Han ˇ ak, and R. Tesa ´ ˇr, ―Teraman: A tool for n-gram extraction from large datasets,‖ in Intelligent Computer

[11] Q.F. Tan, K. Audhkhasi, P. Georgiou, E. Ettelaie, and S. Narayanan, ―Automatic speech recognition systemchannel modeling,‖ in Proc. Interspeech, 2010.

[12] Li, Wang, S. Khudanpur, and J. Eisner, ―Unsupervised discriminative language model training for machine translation using simulated confusion sets,‖ in Proc. Coling, 2010.

[13] A. Stolcke, "Entropy-based pruning of backoff language mod-els," in Proceedings ofthe DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, 1998.

[14] ing Zhang, Almut Silja Hildebrand, and Stephan Vogel, "Distributed language modeling for n-best list re-ranking," in Proceedings ofthe 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, July 2006, Association for Computational Linguistics.

[15] M. Elhadad and J. Robin, ―An overview of SURGE: a reusable comprehensive syntactic realization component,‖ Department of Mathematics and Computer Science, Tech. Rep., 96-03,1996.

[16] H. Inozuka, M. Hosoi, Y. Aramomi, and K. Murata, ―Sentence generation system by IPAL (SURFACE/DEEP),‖ IPSJ Technical Report.NL., vol. 99, no. 22, pp. 113–119, 1999.

[17] H. Hong, S. Lim, and S.-B. Cho, ―Autonomous language development using dialogue-act templates and genetic.

[18] Sychev O. A., Mamontov D. P., ―Automatic Error Detection and Hint Generation in the Teaching of Formal Languages Syntax Using Correctwriting Question Type for Moodle LMS‖, 978-1-5386-7531-1/18/$31.00 ©2018 IEEE

[19] Meredita Susanty, Sahrul, Ahmad Fauzan Rahman, Muhammad Dzaky Normansyah and Ade Irawan,‖ Offensive Language Detection using Artificial Neural Network‖, 978-1-5386-8448-1/19/$31.00 ©2019 IEEE

[20] Mrs. Thejaswini S. and Indupriya C,‖ BIG DATA SECURITY ISSUES AND NATURAL LANGUAGE PROCESSING‖, Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.

[21] Muhammed Enes Almahdi ; Yusuf Sinan Akgül ,‖ Automatic Detection of Word Substitutions within a Language over Periods of Time‖, (UBMK'19) 4rd International Conference on Computer Science and Engineering, DOI: 10.1109/UBMK.2019.8906998.

[22] Suraj Sharma ; Sabitra Sankalp Panigrahi ; Biswajit Paul ; Narayan Panigrahi ,‖ Detection of Topic from Unstructured Text With Mixed Languages‖, 2018 International Conference on

Information Technology (ICIT), 978-1-7281-0259-7/18/$31.00 ©2018 IEEEDOI 10.1109/ICIT.2018.00040.

[23] Palani Thanaraj Krishnan ; Parvathavarthini Balasubramanian ,‖ Detection of Alphabets for Machine Translation of Sign Language Using Deep Neural Net‖, 2019 International Conference on Data Science and Communication (IconDSC), DOI: 10.1109/IconDSC.2019.8816988 ,978-1-5386-9319-3/19/$31.00 © 2019 IEEE.

[24] Kyeong-Hwan Kim ; Chang-Sung Jeong ,‖ Fake News Detection System using Article Abstraction‖, 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), DOI: 10.1109/JCSSE.2019.8864154 978-1-7281-0719-6/19/$31.00 ©2019 IEEE