International Journal of Engineering Science Invention Research & Development; Vol. IV, Issue IX, MARCH 2018
www.ijesird.com, E-ISSN: 2349-6185

# A PAPER ON INTEGRATED SYSTEM BASED ON HINDI AND PUNJABI LANGUAGE

Meenakshi Sharma[1], Dr Satheesh Kumar Nagineni[2]
P.hd Scholar student[1] , Associate professor[2], Computer Science and Engineering Deptt., OPJS University, churu (Rajasthan)
ermeenakshi89@gmail.com[1]

***Abstract:*** **- QG (Question generation), QA (Question Answering), SD (Spell Detection), SC (Spell Correction) are key challenges facing systems that interplay with natural languages. There are many individual existing systems presented on QG, QA, SD and SC. This paper proposes an integrated system that will automatically generate questions from a given paragraph (Consists of text based on useful information) with this, paragraph contains these question answers also. The system will work on two languages that are Punjabi and Hindi. Before QG system, NER system will be developed, because it is a base of QG system. Along with this work, this system will also detect the errors in spellings; if there exist after that system will correct that errors either manually or automatically. For this, an Algorithm named as NQS will be generated. This is a combination of various approaches like Dictionary look up, Rule Based, SMT, Edit Distance, Example Based approach and will use linguistic features of the Punjabi language and Hindi language. SMT will be extended up to 8 grams.**

***Keywords:-*** **Question Generation, Question answering, NQS(Named Entity recognition+Question Generation+Spell Correction), SMT (Statistical Macnine Translation)**

## I. INTRODUCTION

Natural Language processing is an area of Computer Science, Linguistics and Artificial Intelligence, which is used to Interact Human and Computer. The Imperative objective of NLP is Natural Language Generation and Natural Language Understanding. Named Entity Recognition (NER) is an application of NLP. It is also a subtask of the many applications of NLP like Information Retrieval, Information Extraction, Question Answering, Machine Translation, Text Summarization.

NER can be identified as a fundamental subtask of information extraction, which is commonly used across different languages for data and text analytics. The task of NER can be simply defined as locating and categorizing words in an input document into different predefined named entity classes such as person names, location names and organization names. [1]

Question generation is an important aspect of NLP. It is the task of generating reasonable questions from an input, which can be structured (e.g. a database) or unstructured (e.g. a text). In this paper, we narrow the task of QG down to taking a natural language text as input (thus textual QG), as it is a more interesting challenge that involves a joint effort between Natural Language Understanding (NLU) and Natural Language Generation (NLG). Simply put, if natural language understanding maps text to symbols and natural language generation maps symbols to text, then question generation maps text to text, through an inner mapping from symbols for declarative sentences to symbols for interrogative sentences, as shown in Figure 1. Here we use symbols as an organized data form that can represent the semantics of natural languages and that can be processed by a machinery, artificial or otherwise. [10] Automatic spellchecking constitutes one of the major areas in the field of Natural Languages Processing (NLP) and has been the subject of different research studies since the 1960 [4]. Spell-checking give suggestions regarding the solutions closest to an erroneous word. In this paper, two types of spell Correction methods are used; first is manually and second one is automatically. This implies the drawing up of sets of error sequencing as well as the development of the methods that will help lay out the most plausible alternative solutions.

## II. LITERATURE REVIEW

*A. Ananya - A Named-Entity-Recognition (NER) System for Sinhala Language, S.A.P.M. Manamini, A.F. Ahamed, R.A.E.C. Rajapakshe, G.H.A. Reemal, S. Jayasena,*

*G.V. Dias, S. Ranathunga, 978-1-5090-0645-8/16/$31.00 ©2016 IEEE*

This paper represent data-driven techniques to detect Named Entities in Sinhala text, with the use of Conditional Random Fields (CRF) and Maximum Entropy (ME) statistical modeling methods. Results obtained from experiments indicate that CRF, which provided the highest accuracy for the same task for other languages outperforms ME in Sinhala NER as well. Furthermore, they identify different linguistic features such as orthographic word level and contextual information that are effective with both CRF and ME Algorithms [1].

*B. Frequency based Spell Checking and Rule based Grammar Checking, Shashi Pal Singh, Ajai Kumar, Lenali Singh, Mahesh Bhargava, Kritika Goyal, Bhanu Sharma, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016*

In this paper, they presented about English. English is a language that is spoken by around 380-420 million people on this planet and understanding it is not at all easy. The meaning of a sentence varies according to the context and the tone of the speaker. To convey the thoughts efficiently, the knowledge of the language and its various rules is very important as thoughts take the form of words and the words take the form of action. One should aim to minimize the errors while using the language. Lesser is the number of mistakes, better will be the communication. To aid in achieving this goal, they created a frequency based spell checker and a rule based grammar checker for English language. The grammar checker focuses on detecting and correcting tense related mistakes [2].

*C. Automatic Question Generation for Intelligent Tutoring Systems, Riken Shah, Deesha Shah Prof. Lakshmi Kurup, 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*

In this paper, they presented the system of automatic MCQs (Multiple Choice Questions) generation for any given input text along with a set of distractors. The system was trained on a Wikipedia-based dataset consisting of URLs of Wikipedia articles. The important words (keywords) which consist of both bigrams and unigrams are extracted and stored in a dictionary along with many other components of the knowledge base. They had used Inverse Document Frequency (IDF) measure for ranking the extracted keywords and Context-Based Similarity approach using Paradigmatic Relation discovery techniques for generation of distractors. In addition, the question generation phase includes eliminating sentences starting with Discourse Connectives to avoid a question with incomplete information. They had obtained significant accuracy compared to many similar approaches. The results were quite promising considering that there is no human intervention. Though they had developed their system for the field of physics, it can be extended to any field [3].

## III. EXISTING APPROACHES

### A. DICTIONARY LOOK UP APPROACH

In this technique, a parallel corpus will be created for both Punjabi and Hindi named entities which include the names of males, females, countries, locations, states, rivers, places, Bird, Animal etc. for and results will be calculated by comparing the input text with the stored words one by one. This is the easiest and fastest method to obtain the results but works only if the word which is to be translated is present in the database.

### B. RULE BASED APPROACH

A rule based system consists of collection of rules called grammar rules, lexicon and software programs to process the rules. A rule-based system is a set of "if-then" statements that uses a set of assertions, to which rules on how to act upon those assertions are created. In software development, rule-based systems can be used to create software that will provide an answer to a problem in place of a human expert. This type of system may also be called an expert system. Rule-based systems are also used in AI (artificial intelligence) programming and systems [10]. Rules

are written with linguistic knowledge gathered from linguists for both Punjabi and English languages. Rule based system plays a vital role in the process of creating question generation system because handcrafted rules are created according to the grammatical rules of the language used in the system. Rules can be created to extract the names, location names, date formats, name of various places to generate the questions by the system.

### C. EXAMPLE BASED APPROACH

In this approach a set of patterns along with all their possible questions are stored into the database. When an input sentence will be given to the system, system compares the pattern of the input to the patterns stored into the database. If the pattern matches with the pattern stored into the database then set of questions with the help of available stored questions are get generated. If pattern does not found then system will generate the questions on the basis of Rule Based approach.

### D. EDIT DISTANCE APPROACH

Edit distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other. Edit distances find applications in natural language processing, where automatic spelling correction can determine candidate corrections for a misspelled word by selecting words from a dictionary that have a low distance to the word in question [9].

### E. STATISTICAL MACHINE TRANSLATION APPROACH

Statistical machine translation is a data-oriented statistical framework for translating text from one natural language to another based on the knowledge. During translation, the collected statistical information is used to find the best translation for the input sentences, and this translation step is called the decoding process. There are three different statistical approaches in MT, Word-based Translation, Phrase-based Translation, and Hierarchical phrase based model [11].

## IV. CONCLUSION AND FUTURE WORK

The proposed System is implemented in ASP.net and the performance of the proposed technique is compared with some commonly used techniques. A web based interface is created to provide the input to the system and to obtain the results from the system. In existing systems, the results have taken by using only one type of technique. But In this Proposed System, hybrid approach will used to generate the questions. Hybrid approach consist of Dictionary look up, Rule Based, SMT, Edit Distance, Example Based approach and will use linguistic features of the Punjabi language and Hindi language. SMT will be extended up to 8 grams. These approaches works in the sequential order i.e. if one approach fails to produce the desired output then another approach will try to produce the result. In Future, one more Indian language that is English language can also be added make the system language independent.

## REFERENCE-

[1] S.A.P.M. Manamini, A.F. Ahamed, R.A.E.C. Rajapakshe, G.H.A. Reemal, S. Jayasena, G.V. Dias and S. Ranathunga," Ananya - A Named-Entity-Recognition (NER) System for Sinhala Language", IEEE-2016

[2] Shashi Pal Singh, Ajai Kumar, Lenali Singh, Mahesh Bhargava, Kritika Goya1, and Bhanu Sharma, "Frequency based Spell Checking and Rule based Grammar Checking ", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016

[3] Riken Shah ,Deesha Shah and Prof. Lakshmi Kurup, "Automatic Question Generation for Intelligent Tutoring Systems", 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)- 2017

[4] Karen Kukich, " Techniques for Automatically Correcting Words in Text ",ACM Computing Surveys,Vol.24,No.4,pp, 377-439-December 1992.

[5] Frederick J.Damerau, " A technique for computer detection and correction of spelling errors ",Communications of the Association for Computing Machinery-1964

[6] Vladimir Levenshtein, " Binary codes capable of correcting deletions, insertions and reversals ", SOL Phys Dokl,pp, 707-710-1966

[7] J. J. Pollock and A. Zamora, " Automatic Spelling Correction in Scientific and Scholarly Text ",Communications of the ACM,27(4),pp, 358-368-1984.

[8] E. Ukkonen, " Approximate string matching with q-grams and maximal matches ", Theoretical Computer Science, 92,pp, 191211-1992.

[9] https://en.wikipedia.org/wiki/Edit_distance
[10] https://www.webopedia.com/TERM/R/rule_based_system.html
[11] Sakshi Goyal , Er.charandeep Singh Bedi,"
A review on SMS Text Normalization using Statistical Machine translation Approach", 2nd International Conference on CommunicationSystem-2017