Automatic Extraction of Idiom, Proverb and its Variations from Text using Statistical Approach

Chitra Garg[1], Lalit Goyal[2]

[1]M. Tech. Scholar, Department of Computer Science, Banasthali University, Rajasthan, India
chitragarg05@gmail.com
[2]Assistant Professor, Department of Computer Science, D.A.V College, Jalandhar, Punjab, India
goyal_aqua@yahoo.com

ABSTRACT

Natural languages are full of idiomatic uses, which while translating through present NLP system do not extract variations of idioms and proverbs. To overcome this problem, a new method to extract idioms / proverbs is proposed in this paper. The proposed methodology uses statistical method to automatically extract idioms and proverbs from the text along with their variations. The system is updated with a huge database of idioms and proverbs with all of their variations and then tested on a large text file of 'Panchatantra Tales'. The system gave an accuracy of more than 80%, which proves that our method is a successful approach in correctly interpreting and generating the translation of natural language.

Keywords----Natural Language, Proverb, Idiom, Statistical Approach, Idiomatic

I. INTRODUCTION

Idioms are phrases or expressions where the words together have a meaning that is totally different from the dictionary definitions of the individual words. Proverb is a simple and concrete saying that is popularly known and repeated, expresses a truth based on common sense or the practical experience of humanity. Proverbs have the figurative meaning instead of literal meaning. Idiom has figurative meaning as well as composite meaning. Idioms and Proverbs represent a key issue for various applications in NLP (Natural Language Processing) especially for machine translation. Translation quality may be affected by lack of adequate processing of Idioms and Proverbs [1]. From a linguistic perspective, Idioms and Proverbs are presumed to be part of speech that is contradictory to the principle of compositionality.

Idioms and Proverbs are numerous and they occur frequently in all languages. Identifying Idiom and proverb expression from text help us to translate it into another language or to get its meaning i.e. whether the words used in the text are to be taken by their literal meaning or by figurative meaning [2]. Identifying Idioms and proverbs is an important subtask so that computer will enable to recognize idioms and proverbs independently [3]. This differentiation has very much importance in many applications like machine translation, finding paraphrases, information retrieval etc [4]. Any NLP system will make mistakes in translation if it does not have knowledge of non-compositional idioms / proverbs. It is necessary to enable the system to recognize idioms and proverbs so that system can take figurative meaning instead of literal meaning. In this paper, Statistical Approach is used to identify Idiom and proverb from text with its all variations.

Take an example of variations of an idiom:
1. Dull as dish water
2. Dull as ditch water

Take an example of variations of a proverb:
1. Bad news travels fast
2. Bad news has wings

As in the above examples, phrases (1-2) have different term, but both have the same meaning [5].

Identifying idioms and proverbs from text increase efficiency of any system. It reduces the time as we take the complete meaning of an idiom and proverb instead of the composite meaning of individual words in translation. It also reduces searching time of a lexicographer. Identifying idioms makes the system able to respond intelligently to natural language input and improves the coverage of language resources [6]. This work is also useful for sign languages. Deaf students can get the idiomatic meaning by acting in place of literal meaning. Idioms and Proverbs

are also used to express the emotion and attitude of a person. . For example- a student learning English language finds an idiom in his reading, it is necessary for better understanding to know the attitude and emotions behind this medium [7]. Further it can be useful in web searching and parsing of text.

II. RELATED WORK

Monika Gaule et. al. [8] analyzed how to identify and translate idiomatic expression from English to Hindi. They describe the main problems and difficulties during idioms translation and identification of idioms is utmost important resource for machine translation system. They proposed a rule based approach for identification of idioms and used the Google translate system to translate idioms. They applied this resource on manually created testing data. Their system output is 70% accurate and shows the problem of bad translation due to errors of different categories like grammar agreement, part of speech, irrelevant idioms etc.

Monika et. al. [2], designed graphical user interfaces for extracting proverbs in machine translation from Hindi to Punjabi. They have used relational data approach. Hindi and Punjabi proverbs divided into two parts: static and dynamic. Static part is handled by regular expressions and dynamic part may have inflections. Static part will be matched in database and when match found, it gets the corresponding Punjabi meaning of the proverb. This approach gives result with 60-80% accuracy.

Ashwini Aggarwal et. al. [9] describe an approach for automatic extraction of multiword expression of specific kinds from a moderate size untagged corpus of Bengali language using morphological analysis and statistical method .It is a method to handle sparse linguistic data. In this paper first of all noun verb, adjective verb and adverb verb collocation is extracted. Possible MWs candidates are extracted from the sentences and assigned a significant value based on statistical parameter like co-occurrences and individual frequencies. Then the list of different classes of MWs is finally sorted in the descending order of significance value.

Tim Van de Cruys et. al. [5] describes a fully unsupervised and automated method for large-scale extraction of multiword expressions from large corpora. The intuition for extracting multiword expression is that a noun within a MWE cannot be substituted by a semantically similar noun. Noun clustering is automatically extracted to implement this intuition. Noun clustering means cluster of semantically related nouns. To formalize the intuition of non-compositionality, a number of statistical measures are developed. They try to capture the MWE's non-compositionality bond between a verb-preposition combination and its noun. Approach given by them has been tested on Dutch and assessed automatically by Dutch lexical resources.

III. PROPOSED METHODOLOGY

As discussed above, to overcome the difficulty in extracting the variations in idioms/proverbs, a different system is proposed which uses statistical approach to translate the idioms/proverbs. The flow chart in fig.1 describes the complete methodology of proposed idiom/proverb extraction system.

In the present system, we have created a huge database of proverbs and idioms along with their variations. The proposed proverb/idiom extraction system has two options, one for idioms and another is for proverbs. The user has to choose one of the two options to search idioms / proverb in the text file. Later on, user has to input idiom / proverb to be searched and the text file in which searching is to be performed. After the file selection, the system will fetch the user entered idiom / proverb and its variations from the stored database. Using the KMP pattern matching algorithm, the system will match the fetched idiom / proverb with the text file and perform the tagging of the searched idiom / proverb. The output will be the tagged text file, which indicates the searched idioms or proverb. KMP pattern matching algorithm searches the occurrences of a word within a main text.

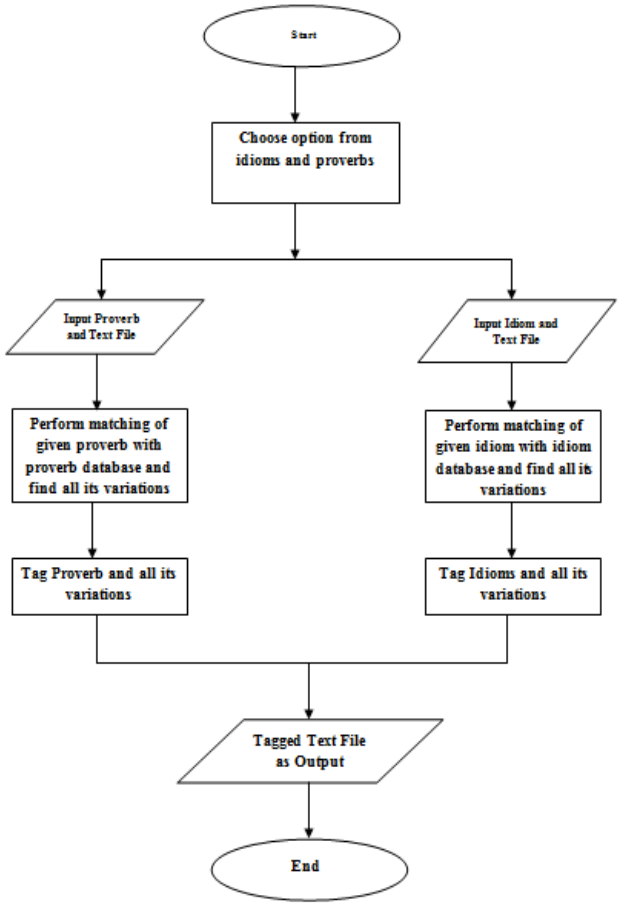

Fig. 1 Flow Chart of Proverb/Idiom Extraction System

Proposed Proverb / Idiom system contains three units as follows.

A. Input Unit

The Proverb/Idiom and text file is entered as input. Here in Fig. 2, idiom “keep an eye on” is given as input.

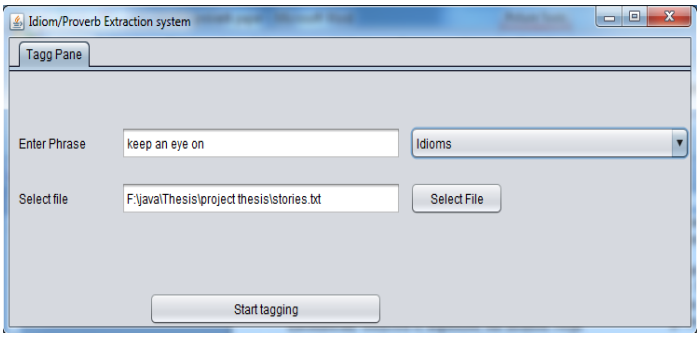

Fig. 2. Input Unit for Idiom

In Fig. 3 Proverb “wise is stronger than the strong” is given as input.

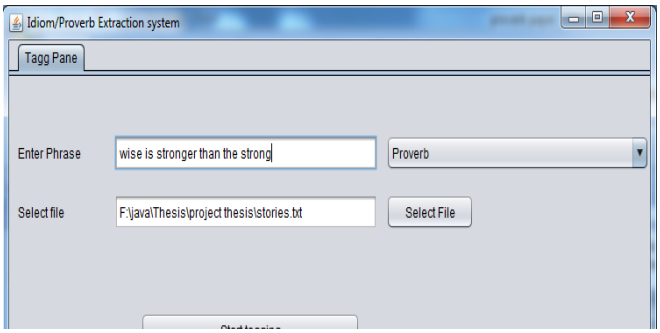

Fig. 3 Input Unit for Proverb

B. Processing Unit

The given input idiom or proverb is matched with the database and fetched through its variations. Using KMP pattern matching algorithm, system will tag all the input proverb or idiom with all its variations in the entered text file.

C. Output Unit

In this unit, the result comes after processing of idiom or proverb. Here Fig. 4 shows the resultant idiom tagged text file.

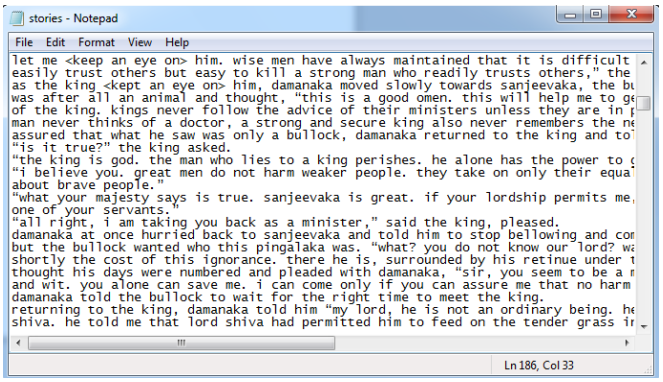

Fig. 4 Tagged Idioms in Text File

Here Fig. 5 shows the resultant Proverb tagged text file.

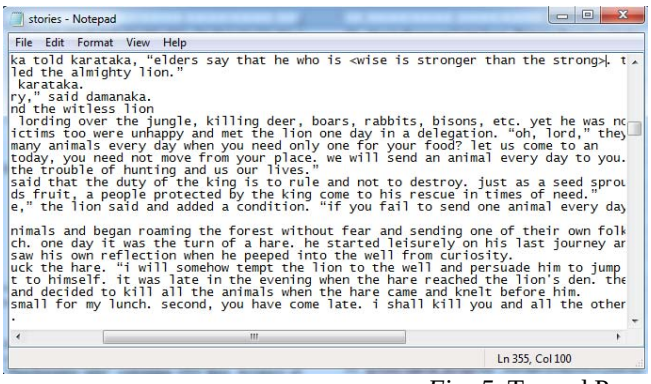

Fig -5: Tagged Proverbs in Text File

IV. RESULTS

System's quality is measured by the usual notion of the information-retrieval criteria. A parameter that is used to evaluate the proposed system is accuracy.

$$\text{Accuracy (\%)} = \frac{\text{Total tagged idioms ,proverbs and their variation}}{\text{Total number of idioms, proverbs and their variations}} \text{ X 100}$$

Accuracy is directly proportional to the size of the database. Bigger database leads to higher accuracy. For calculating the accuracy we have taken a text file of "Panchatantra tales" containing 1600 lines. Accuracy of our proposed system is 80.62%.

CONCLUSIONS

Identification of various idioms / proverbs and their variations can be done using the proposed statistical approach of extracting idioms and proverbs. The proposed approach is checked against a text file of Panchatantra Tales and the result shows an accuracy of more than 80 %, which proves that our method is a successful approach to get an idea of the non-compositionality of idioms / proverbs in a fully automated way. All the generated information is useful in correctly interpreting and generating the translation of natural language.

REFERENCES

[1] Dhouha Bouamor, Nasredine Semmar, Pierre Zweigenbaum, "Identifying bilingual Multi-word Expressions for Statistical Machine Translation", International Conferences on Language Resources and Evaluation, May 2012

[2] Monika Sharma, Vishal Goyal, "Extracting Proverbs in Machine translation from Hindi to Punjabi using Relational Data Approach ",International Journal of Computer Science and Communication, July-December 2011, Vol. 2, PP. 611-613,

[3] Eugenie Giesbrecht, Graham Katz, "Automatic Identification of Non-Compositional Multiword Expressions using Latent Semantic Analysis", Proceeding of the workshop on Multiword Expressions, Associations for Computational Linguistics, July 2006,PP 12-19

[4] Begona Villada Moiron, Jorg Tiedemann, "Identifying idiomatic expression using automatic word alignment", Proceedings of the workshop on Multiword Expressions in a multilingual context, 11th Conference of the European Chapter of the Association for Computational Linguistics, April 2006, PP 33-40

[5] Begona Villada Moiron , Tim Van de Cruys , "Semantics-based Multiword Expression Extraction", Proceedings of the Workshop on a Broader Perspectives on Multiword Expressions, Associations for Computational Linguistics ,PP 25-32

[6] Beate Dorow, Dominic Widdows, "Automatic Extraction of Idioms using Graph Analysis and Asymmetric Lexicosyntactic Pattern", ACL 2005 Workshop on Deep Lexical Acquisition, June 30,2005

[7] Lei Wang, Shiwe n Yu, "Construction of a Chinese Idiom Knowledge Base and Its Applications" Proceeding of the Workshop on Multiword Expressions: from Theory to Application (MWE 2010), 23 rd International Conference on Computational Linguistics, Aug 2010, PP. 11-18

[8] Monika Gaule, Dr. Gurpreet Singh Josan, "Machine Translation of Idioms from English to Hindi", International Journal of Computational Engineering Research, October 2012, Vol. 2 Issue 6

[9] Aswhini Agarwal, Biswajit Ray "Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenarios", Proceedings of the International Conference on Natural Language Processing (ICON 2004). Allied Publishers, Dec 2004, PP. 165 - 172

BIOGRAPHIES

CHITRA GARG received B.E degree in Information Technology from University of Rajasthan, India in 2009 and she is pursuing her M. Tech. in Computer Science and Engineering from Banasthali University, Rajasthan, India. Her area of interest is Natural Language Processing.

Mr. Lalit Goyal received M. Tech. degree in Computer Science & Engineering from Punjabi University, Patiala, India under the guidance of Dr. Gurpreet Singh Lehal and pursuing Ph.D under the guidance of Dr. Vishal Goyal. His area of interest is Natural Language Processing and Image Processing. Currently he is working as Asst. Prof. in D.A.V College, Jalandhar, Punjab , India He has published around 5 National and 3 International publications