IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.6, Special Issue.10, Nov 2018 E-ISSN: 2347-2693

Grammar And Context Based Approach For Identification And Translation Of Proverbs Using Trie-Based Ontology

Naziya Shaikh[1*]

[1] Department of Computer Applications, Rosary College, Goa University, India

[*]Corresponding Author: naziya1019@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** Most current machine translation systems for translation from English to regional Indian languages ignore the presence of idioms in the text or return the exact literal meaning of the phrase in another language which loses the essence of the proverb. The main issues that arise include timely detection of the proverbs from a given paragraph and the separate processing required for translations of idioms into other languages. This paper presents a combination of natural language grammar-based approach and context-based approach towards detection of idioms in given English text and further presents a trie-based ontology that can be used to translate proverbs into regional languages. The grammar-based approach involves parsing English sentences and identifying the parts-of-speech tags and determining statistically the probability whether the given sentence is a proverb using certain grammar-based rules applicable for only proverbs. The context-based approach classifies and compares keywords in the proverbs and the keywords present in remaining part of the paragraph. Based on the combination of these two approaches, the proverb can be determined with better accuracy. For quick translation of detected proverbs into regional languages, keyword based priority search can be implemented on previously developed trie-based ontology using parts-of-speech tags.

***Keywords*—**Machine translation, Proverbs, Idioms, English, Regional Languages, grammar-based approach, classification, context-based approach, trie-based ontology

I. INTRODUCTION

Machine translation of English into regional languages has progressed well with most common languages spoken in India already being translated into English and vice versa. Yet some aspects of the natural languages like proverbs which are not used very often are usually ignored during the translation. The reason for this as mentioned in [1] is that these idioms do not make language based sense. Therefore, we need a language specific rule-based criteria and algorithm with a separate ontology for these translations. Even before translating the idioms, there is a need to detect these sayings or proverbs from a large paragraph given as an input for translation. Apart from that, the idioms are usually influenced by the environment and region where that language has evolved. As a result it is difficult to find corresponding proverbs with exact same implication in the target language during translation. Therefore sometimes we need to translate them into a small paragraph explaining the idiom in another language. Yet the current solution only provides a literal translation which loses the essence of the proverb as these idioms do not create any language –based semantics.

This is the case for many regional Indian languages including hindi, Marathi, Konkani, etc. For illustration, translation from English to Marathi language has been considered. Consider the following examples of idioms in English language for which there are available corresponding proverbs in Marathi language.

1. English Saying: ‘Tit for Tat’

Corresponding saying in Marathi language: ‘जस्यास तसे’

2. English Saying: ‘The grass is always green on the other side.’

Corresponding saying in Marathi language: ‘दुरून डोंगर साजरे.’

3. English Saying: ‘Too many cooks spoil the broth’

Corresponding saying in Marathi language: ‘अति तेथे माती.’

4. English Saying: ‘Empty vessels make more noise’

Corresponding saying in Marathi language: ‘उथळ पाण्याला खळखळाट फार.’

The above sayings in English translate into Marathi as a completely different proverb, but with same implication. There are other sayings in English, which do not have corresponding proverbs in Marathi language. Consider for example the proverbs: “Two wrongs don't make a right, “Justice delayed is justice denied”. In such cases, appropriate translation would be to explain the implied meaning of the

© 2018, IJCSE All Rights Reserved

proverb in target language. But most of the machine translation systems use literal translation instead of the implied meaning.

Consider the following examples entered into google translate system to translate from English to Marathi, the results are literal translations on these proverbs as shown below:

1. English proverb: Two wrongs don't make a right

Google translate converted this into Marathi as: ‘दोन चुकीच्या अधिकार करू नका’

2. English proverb: Justice delayed is justice denied

Google translate converted this into Marathi as: ‘विलंब न्याय न्याय नाकारला जातो’

3. English proverb: The grass is always green on the other side

Google translate converted this into Marathi as: ‘गवत नेहमी दुसऱ्या बाजूला हिरवा आहे’

4. English proverb: Too many cooks spoil the broth

Google translate converted this into Marathi as: ‘बरेच स्वयंपाकी रस्सा पाणी घालणे’

5. English proverb: Empty vessels make more sound

Google translate converted this into Marathi as: ‘रिकामी भांडी अधिक आवाज करू’

As we can see, most of the translations done are completely dictionary based literal translations.

II. Related Work

Existing approach uses a simple relational database with entire text of proverbs stored and matched as mentioned in [1]. In this paper, the translation is done using 2 tables, one holds the source language phrases and the other holds the target language meanings or corresponding phrases. The source language table, also holds the number (count) of the corresponding phrases in the target language and a pointer to the entry of the first phrase in the target language. The count column is included in source table since a single phrase in the source language may map to more than one phrase in the target language. For translating, each word in the sentence is mapped on to the phrases of the source language linearly, once the entire match is encountered, the corresponding target language phrases are accessed based on the pointers. Count was already stored, so all the corresponding phrases are retrieved and using word sense disambiguation, most accurate phrase is chosen.

This method requires a lot of processing as entire strings are stored and searched through in order to detect any proverb, as well as to acquire the corresponding meaning. This processing adds to the overall translation time making the machine translation slow. Also, as proverbs are very few in a paragraph, or might not even exist in the text to be translated, using an entire database search leads to performance and storage issues in machine translation systems. Therefore, some other existing systems prefer speed and performance over translation of few proverbs. These systems completely ignore the existence of the proverbs in the paragraph, and translate the proverbs in a literal fashion. Consider google translate example given previously, wherein the translation speed was very fast, but the translation was literal. As the translation systems already have a large corpus to scan through for normal translation using various methods like direct translation, statistical translation, example-based translation, etc., most existing systems have not explored the proverbs translations in detail.

III. Methodology

The proposed methodology includes the development of an ontology which uses the most important nouns and adjectives in proverbs to filter out the search for required proverbs for translation in lesser amount of time, as the first activity in the process. Further, an algorithm with various grammar-based rules can be used to identify the proverbs in a given input source paragraph. To deal with translation of idioms present in English language into regional languages, an algorithm to optimally access the trie-based ontology can be utilized. Accordingly we can translate the source language proverb into corresponding target language proverb or into some target language statements that would have the same meaning as the proverb.

A. Identification of the proverb from source language paragraphs

For the identification of proverbs, this paper proposes combination of context-based approach and natural language grammar based approach towards detection of idioms in given English text. The context-based approach is used based on the fact that most proverbs do not relate semantically to the rest of the context that is present in the paragraph. Thus, if we classify using the keywords in a given sentence and include the general categories that can be related to the keywords in that sentence, there is a high possibility that the keywords from the rest of the paragraph will not belong to the same categories.

The grammar-based approach works on the fact that was observed in data sets consisting of around 1500 English proverbs. It was observed that proverbs are either clauses or sentences that are usually in present tense and active voice and do not include personal pronouns like ‘he/she’ in the sentences except for the beginning of the sentence in certain cases. It has been also observed that whenever a proverb consists of personal pronouns at the beginning of the sentence, it also consists of a subject pronoun like ‘who’ in the same sentence.

Based on the above observation of the general custom grammar rules in the proverbs, we could devise an algorithm to determine the possibility of given sentence in the source paragraph to be a proverb.

Consider for example the proverbs given below:

1. He who knows nothing doubts nothing.
2. He has enough who is content.
3. He who digs a pit for others falls into it himself.
4. All good things come to he who waits.

There are exceptions to this grammar rules, but considering the number of these exceptions among the large number of proverbs, we can say that these grammar rules can be applied in general to detect most probable proverbs in a paragraph. To increase the precision of the detection, we can use a counter ‘proverb-weight’ as an indicator of whether a given sentence is a proverb or not. Based on the various grammar-based tests, weights can be added. The higher the weight, more is the possibility of the given sentence being a proverb.

For obtaining grammar based rules, on any sentence, we first need to parse the sentence and provide parts of speech tags to words in the sentence. Once we have done that, we can use parts-of-speech (POS) taggers[5][6] and apply algorithms to verify the tense of the sentence to determine if it is in ‘present tense’ based on the verbs tagged using POS tagger. We can also use the POS tagger to find out the presence as well as placement of personal pronouns as well as subject pronouns in the given sentence.

The following is the overall algorithm to identify the proverbs in a given English sentence.

1. Define two indicators proverb-weight (used to measure the grammar-rules based possibility of the sentence being a proverb) and difference-counter (used to measure the context-based possibility of the sentence being a proverb).

2. For each sentence in the paragraph,

- **I.** Parse the given sentence using POS tagger.
    - **a.** Identify the tense of the sentence. If the sentence is in present tense, add certain appropriate weightage to the indicator ‘proverb-weight’.
    - **b.** Identify the voice of the sentence. If the sentence is in active voice, add certain appropriate weightage to the indicator ‘proverb-weight’.
    - **c.** Check if the personal pronoun is present in the beginning of the sentence. If present, check if the subject pronoun is present in the same sentence. If both pronouns are present in the same sentence, add certain appropriate weightage to the indicator ‘proverb-weight’.
      Else, reduce certain appropriate weightage from the indicator ‘proverb-weight’.
- **II.** Categorize the keyword from the given sentence into related classes/categories. Compare the categories of the keywords from the rest of the paragraph with the classes/categories obtained for current sentence. If the categories are varied and do not belong to any related areas, then there is a large possibility that the given sentence is a Proverb. Based on the difference in the areas of categorization, assign the ‘difference-counter’ for the given sentence.
- **III.** Based on the combined values of the two indicators we need to define a threshold to decide whether a given sentence is a proverb. Higher the values of these indicators, more is the possibility of the sentence being a proverb.

In the first step, the English sentence is parsed using the POS tagging and the pronouns, nouns, and verbs in the sentence are determined. Further, we need to identify the tense of the sentence using the verbs tagged by POS tagger. If tense is present tense, we add some weightage to the counter ‘proverb-weight’. Then we identify the voice of the sentence using POS tags. If the sentence is in active voice, add certain appropriate weightage to the indicator ‘proverb-weight’.

Next, we need to find if there is any personal pronoun in the sentence and also check its position. If there is a personal pronoun present at the beginning of the sentence, it can imply that there is either some possibility of the clause still being an idiom or the given clause could be a regular English sentence. To verify the former possibility, we check for the subject pronoun like the word ‘who’. If there is no subject pronoun in the sentence, it implies that it is a regular English sentence. Hence we denote this, by reducing some points from the variable ‘proverb_weight’. If there is a subject pronoun to the sentence, it is probably an idiom, to denote this, we add some points to the variable ‘proverb_weight’. Otherwise if there is a personal pronoun present at any other place in the sentence, probability of that sentence being an idiom is very low, and hence we assume there is no proverb.

After the evaluations, if the combined value of proverb-weight and difference-counter is low, it implies that it is a normal English sentence and that it can be directly processed using the machine translator. If the value is high (above the threshold), sentence can be a proverb and should be further verified in the next phase, wherein nouns/adjectives from the proverbs would be matched with the trie based dictionary structure.

B. Development of a trie based ontology

For the overall translation of proverbs into target language proverbs, we need to have an entire database of the proverbs. This paper proposes that, instead of having the data stored in form of phrases, the entire set of proverbs in the English language could be analyzed, and a new structure can be built, which uses a tree based approach, starting from most important adjectives / nouns/ verbs in a given proverb. Therefore we could use the trie data structure and place a search using most important words as the higher nodes in the trie, followed by the less important nodes at each level, thus making the search of a proverb simpler and faster. The database will hence be divided into smaller parts and will be easier to access. Also as we are carrying out word sense disambiguation during the search itself locating correct corresponding proverbs would be easier.

The creation of such ontology requires analysis of a large language corpus in English consisting of sayings as well as corresponding meaning and an expert with grasp on both source and target languages. But once created, the processing speed of the machine translation systems, which translate the proverbs would increase highly.

Once the ontology is created, during the translation, we need algorithm to POS tag the sentence, choose most important nouns, verbs and adjectives, then access the trie based ontology based on those nouns and adjectives and search through that point to reach the complete proverb. Once complete proverb is reached, pick up the correct corresponding meaning or corresponding saying that was already saved during the creation of this trie-based ontology. Then add this to overall translated paragraph.

IV. DISCUSSION

Existing approach that uses a simple relational database with entire text of proverbs stored and matched requires a lot of processing. The proposed method will highly improve the speed of search as it just the stemmed POS tags that are being searched with the help of the trie data structure. Also, the detection of the proverbs from the paragraph would be improvised as it would no more require any linear analysis of the database to detect the proverbs.

The creation of an entire trie based ontology, based on each stemmed root of the nouns, adverbs which consists of a very wide set of keywords requires explicit expertise in the language as well as large amount of human effort. An automated POS based algorithm for automating the creation of this ontology could be a further enhancement to this translation process.

V. CONCLUSION AND FUTURE SCOPE

This paper proposed the translation of proverbs in a source language using a natural language grammar based approach combined with the context-based approach towards detection of idioms in given source text and further utilizing a trie-based ontology that can be used to convert these idioms into other language translations. The grammar-based detection involved parsing the sentences using parts of speech tagger and analyzing the tokens for various grammar rules that are mostly specific to the proverbs. These included searching for sentences with present tense, active voice, for those sentences where personal pronouns are present along with subjective pronouns.

Context-based detection categorized the keywords in the sentence and compared the difference between the categories of the given sentence with the categories of the rest of the sentences in the paragraph. Once the proverbs were found, they were optimally searched through the trie-based ontology to reach to the translated texts, which were then added to the overall translated paragraphs in the target language. This approach has many advantages over the existing approach including improved overall translations, enhanced speed and performance. In this approach the creation of the trie based ontology is critical and an automated algorithm to do the same could be a future enhancement to this research.

REFERENCES

[1] D. Pisharoty, P. Sidhaye, H. Utpat, S. Wandkar, R. Sugandhi, “Extending capabilities of english to marathi machine translator”, IJCSI International Journal of Computer Science Issues, Vol.9, Issue No.3, May 2012. ISSN (Online): 1694-0814.

[2] M. Sharma, V. Goyal, “Extracting proverbs in machine translation from hindi to punjabi using regional data approach”, International Journal of Computer Science and Communication, Vol. 2, No. 2, pp. 611-613, July-December 2011.

[3] V. K. Birla, M. N. Ahmed, V. N. Shukla, “Multiword expression extraction – text processing”, In the Proceedings of ASCNT – (2009), CDAC, Noida, India pp. 72-77, 2009.

[4] V. Goyal and Priyanka, “Implementation of rule based algorithm for sandhi-vicheda of compound hindi words”, International Journal of Computer Science Issues, No. 3, pp. 45-49, 2009.

[5] K. Toutanova, C. D. Manning, “Enriching the knowledge sources used in a maximum entropy part-of-speech tagger”, In the Proceedings of the 2000 joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70, 2000.

[6] K. Toutanova, D. Klein, C. Manning, Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network”, In the Proceedings of 2003 HLT-NAACL, pp.252-259, 2003.

[7] R. Balyan, S. K. Naskar, A. Toral, N. Chatterjee, “A diagnostic evaluation approach targeting MT systems for indian languages”, In Proceedings of the 2012 Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pp. 61-72, COLING 2012, Mumbai, December 2012.

[8] L. R. Nair, P. S. David, “Machine translation systems for indian languages”, International Journal of Computer Applications, Vol. 39, No. 1, February 2012. ISSN: 0975-8887.