Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 45–50
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Determination of Idiomatic Sentences in Paragraphs Using Statement Classification and Generalization of Grammar Rules

Naziya Mahamdul Shaikh
Government College of Arts, Science and Commerce
Quepem-Goa, India
naziya1019@gmail.com

Abstract

The translation systems are often not able to determine the presence of an idiom in a given paragraph. Due to this many systems tend to return the word-for-word translation of such statements leading to loss in the flavor of the idioms in the paragraph. This paper suggests a novel approach to efficiently determine probability of any statement in a given English paragraph to be an idiom. This approach combines the rule-based generalization of idioms in English language and classification of statements based on the context to determine the idioms in the sentence. The context based classification method can be used further for determination of idioms in regional Indian languages such as Marathi, Konkani and Hindi as the difference in the semantic context of the proverb as compared to the context in a paragraph is also evident in these languages.

Keywords: idioms, rule-based generalization, classification

1. Introduction

Most translation systems are able to reasonably translate one language to another. But many fail when it comes to translation of the idioms in a language. Idioms are multi word expressions and the correct translation of such expression requires usage of completely different semantic context in a different language which is difficult to achieve using direct translation method. For translating idioms, system must first determine that the given sentence in a paragraph is an idiom and once identified, search for it from a large corpus storing the proverbs which requires a lot of processing. Therefore most systems prefer to treat idioms like any other general language sentence and provide literal meaning instead of a semantically equivalent statement in target language that would actually provide the flavor of the idiom to the given paragraph. Certain other systems would simply carry out search through a large corpus of stored idioms which is also very time-consuming. A new method is needed towards finding idioms in a paragraph efficiently. Further after determining idioms, processing is required to search for translation of such idioms with the correct words denoting semantic translation of the same in the target language. A new method of storage that can reduce the amount of time required for the search would greatly enhance the translation process of the idiomatic sentences.

2. Related Work

Various types of studies have been done on the structural compositions of the proverbs in different languages. One of the theories treats the proverbs as a species of metaphors (Coinnigh, 2013). Other anthropological, folkloric, and performance-based studies were carried out over the use of various proverbs in different cultural settings (Bradbury, 2002).

For finding the proverb in the given paragraph, certain methods have been used in the recent years. In one method, each statement in the paragraph is taken, all the inflections in the statement are removed and the statement is compared with the entire exhaustive list of stored proverbs for a match (Pisharoty et al., 2012). But this method does not provide an optimized approach. As the size of the input paragraph gets larger, the processing required for this method also increases greatly. In another study, idioms are treated as any other statement to be parsed into parts of speech tags until a pattern is recognized based on the threshold. Based on the activation achieved using grammar based parsing, a new additional process is added for the idioms (Stock, 1989). One of the methods uses a proverb translation system in which source language text are searched and their equivalents in target language are produced using single word information.(Brahmaleen et al., 2010) In this method, input is split into single word units and saved in one dimensional array. Proverb is also split into single word units and saved as a row in two dimensional array. The two arrays are compared to determine whether the statement is a proverb. A linguistic characteristic called verb-object collocations used to determine the asymmetry in the semantics between a verb in the statement and the corresponding object in the same statement is used in the determination of the idiomatic statements (Tapanainen et al., 1998). Yet another method used for determination of proverbs (Coinnigh, 2013) focuses on the metaphors present in the statements likely to be a proverb. This study analyzed the frequency, nature and form of metaphor in Irish-language proverbs. This study claimed metaphors could be secondary proverbial markers. For the translation of the proverbs, current method either uses direct storage of list of proverbs along with their translation to be searched with sequential accesses (Pisharoty et al., 2012) or uses complete ignorance of the proverb and does word-for-word literal translation for the proverb. Yet another method (Goyal and Sharma, 2011) uses a simple relational data approach after extracting the static part of the Hindi language proverb from the sentence, handled by using regular expression. The commonalities in the methods used in translation of proverb in Indian languages like Marathi were also studied and the methods used for translation were generalized into common strategies (Dash, 2016).

3. Proposed Work

This paper presents an idea to determine the idioms in a given paragraph using certain generalized rules related to the grammar which are common for most of the idiom statements in that language. The results of this method can

then be combined with classification based method to improve the probability of a given statement being an idiom.

3.1 Determination of Statement as an Idiom

3.1.1 Classification Method

The classification method includes classifying a given statement into a general category. Idioms in a paragraph generally tend to belong to a different context as compared to the actual paragraph content. Using this fact, the category of the entire paragraph is compared with the category of the statement. The difference between the two categories is measured. The higher the difference more is the probability of any given sentence being an idiom. The same concept can be enhanced and further developed cross linguistically for regional languages like Hindi, Marathi and Konkani because the proverbs in most languages generally consist of a completely different semantic context as compared to the context of the remaining paragraph.

3.1.2 Generalized Rules Method

Certain rules of the English language usually apply for most of the idioms in a paragraph. After observation of several proverbs in the language, we can assume in general, that the rules include facts like - most of the idioms tend to be in present tense, idioms are usually said in active voices, an idiom statement usually does not contain a personal pronoun unless the pronoun is placed at start of the statement along with a subject pronoun in the same sentence. For example, consider a statement like: "He will fail as he has not planned properly" versus an idiomatic statement like "he who is failing to plan, is planning to fail". In the first statement, personal pronoun 'He' is not followed by a subject pronoun. Whereas in second statement, a subject pronoun 'who' follows.

Also presence of a clause is very common whenever a sentence includes an idiom. Consider for example the phrases in the paragraph like: "But you know what they say, don't judge a book by its cover" or "It has been rightly said that whenever there is a will, there will always be a way". In such cases, there is high amount of usage of certain limited set of words such as 'realized', 'told', 'said', 'knew', 'say' and few more. All these rules can be generalized and used as criterias to indicate the presence of an idiom in a sentence. An algorithm can be written to check for these rules in each statement and accordingly assign weightage to each criteria for that statement. The higher the scoring for each criteria, more is the probability of the statement being an idiom.

3.2 Development of POS Tags Based Data Structure for Better Access of the Proverb Translations

To provide the translation of proverbs, we need to store the proverbs and their corresponding semantically translated meaning in target languages. But due to the large number of idioms, which are required to be stored, a direct storage can cause processing issues during search operations. Therefore instead we can use a different approach for storage of these proverbs in the data structure. Ontology can be prepared which contains major adjectives and nouns in the idioms as the beginning of the access search. As the nouns and adjectives are not very common throughout the idioms, the search list can be easily filtered and the amount of search comparisons required can be lowered. When a proverb is identified in a paragraph, the proverb statement will be POS tagged and various nouns and adjectives in the proverb would be identified. Based on the nouns and adjectives in the proverb, ontology prepared as mentioned before can be searched to find the proverb and then display the corresponding semantic translation meaning.

4. Implementation

4.1.1 Classification Method

For implementation of the classification method, we find the categories of the sentence and the rest of the paragraph without that sentence. In this implementation, Application Programming Interface (API) provided by IBM Watson has been used for testing the categories. The model was not trained for any specific requirement, only the general version was used for this implementation. The general version of this API provided classification into already specified sets of categories by IBM Watson including 'style and fashion', 'law and politics', 'science', 'spirituality', 'parenting', etc. Member sets were created with various categories available in the classifier in such a way that categories that are semantically closer in context are placed into same member set. Each member set was arranged according to the similarities between the categories within it in order to determine the proximity of the categories for member sets. Based on these member sets, the difference between the categories of the given statement and the categories of the paragraphs excluding that given statement was determined. In this implementation, for simplicity, only the following grading of the category classes has been considered based on the difference determination using member sets:

Completely not matching - 4.5
Matching to some extent - 3.5
Matching well - 2
Complete match - 1.5

The following paragraph is used as sample to demonstrate the method for finding the idioms. Similarly other paragraphs were checked using the same method and the similar grading was obtained for those other paragraphs.

Paragraph:
"There is this boy in my neighborhood. He is Very strangely dressed and remains quite aloof. He has tattoos all over his body. But you know what they say, don't judge a book by its cover. So I went ahead and tried to converse with this boy. And I was right, this boy was indeed had a very interesting nature. He was just a teenager in his growing phase. That is when I realized to understand others we need to put ourselves in other people's shoes."

The results found after analyzing the given paragraph using the classifier are as shown in the table below. The table shows the category of the given statement, the category of the entire paragraph excluding the given statement and the score which is calculated using the grading based on the difference of the categories of the sentence and the remaining paragraph.

Sentence	Sentence Category	Paragraph Category	Score
There is this boy in my neighborhood.	(real estate / low income housing)-0.77 (/ home and garden / gardening and landscaping / yard and patio)- 0.25 (/ food and drink / food and grocery retailers / bakeries)-0.16	(/ style and fashion / footwear / shoes)-0.95 (/ style and fashion / body art)-0.44	3.5
He is very strangely dressed and remains quite aloof.	(/ law, govt and politics / politics)-0.43 (/ pets / cats)-0.36 (/ business and industrial / company / merger and acquisition)-0.22	(/ style and fashion / footwear / shoes)-0.95 (/ style and fashion / body art)-0.42	3.5
He has tattoos all over his body.	(/ style and fashion / beauty / tattoos)-1.00 (/ style and fashion / body art / hobbies and interests)	(/ style and fashion / footwear / shoes)	1.5
But you know what they say, don't judge a book by its cover.	(/ art and entertainment / books and literature)-0.98 (/ law, govt and politics / government / courts and judiciary)-0.05 (/ business and industrial / company / bankruptcy)-0.03	(/ style and fashion / footwear / shoes)-0.96 / style and fashion / body art 0.42	3.5
So I went ahead and tried to converse with this boy.	(/ style and fashion / footwear / sneakers)-1.00 (/ style and fashion / footwear / shoes)-0.02 (/ shopping / retail / outlet stores)	(/ style and fashion / footwear / shoes)- 0.96 (/ style and fashion / body art)-0.42	1.5
And I was right, this boy indeed had a very interesting nature.	(/ science)-0.54 (/ law, govt and politics / politics)-0.48 (/ religion and spirituality)-0.48	(/ style and fashion / footwear / shoes)- 0.96 (/ style and fashion / body art)-0.42	3.5
He was just a teenager in his growing phase.	(/ family and parenting / parenting teens)-0.96 (/ family and parenting / children)-0.10 (/ family and parenting)-0.06	(/ style and fashion / footwear / shoes)-0.96 (/ style and fashion / body art)-0.42	4.5
That is when I realized to understand others we need to put ourselves in other people's shoes.	(/ style and fashion / footwear / sneakers / style and fashion / footwear)	(/ style and fashion / footwear / shoes)-0.96 (/ style and fashion / body art)-0.42	1.5

Table 1: Analysis of the paragraph using the classifier

4.1.2 Generalized Rules Method

In the implementation of this method, we first generate tense of the sentence using Stanford POS Tagger (Toutanova and Manning, 2000) which provides the parts of speech tags for each word in the sentence. If the tense is any form of present tense, we assign a score based on this criterion. In this implementation, we have assigned a comparative score of 2.5 for the statement to be in present tense as this rule applies for most of the English proverbs. We further evaluate the statement to check whether the statement contains a clause and a score is assigned based on this criterion. If there are words like realized, told, etc. present along with the clause, then some more points are given to the criterion as the clauses containing these words are very common in the idiomatic statements in the paragraphs. This is followed by determination of Active/Passive voice of the sentence using Dependencies of the sentence provided by Stanford Core NLP package (Toutanova et al., 2003). After this, using the POS tags of the sentence generated before, a personal pronoun (such as he/she) is searched in the sentence. If a personal pronoun is found, we further search for a subject pronoun (such as who) in the same sentence. We also check the position of the personal pronoun in the sentence. After observation of different pronouns in the language, the rule has been generalized stating that if only a personal pronoun is present, sentence is usually not found to be an idiom. Therefore we decrement the score in this case. Whereas if personal pronoun is present and also subject pronoun is present and the personal pronoun is placed before the subject pronoun, then the statement may be an idiom. So we add to the score based on this criterion.

A general algorithm has been written using Java packages to check for the tense, clauses (if found each clause is processed separately), personal pronouns such as He/She, subject pronouns(such as who) and certain words which are used often while using proverbs in the paragraph for example: (realized, said, told, etc). According to the various possibilities of the languages and probabilities of the rules, a certain grade is assigned to each criteria and sentences are evaluated accordingly.

The following three tables describe the evaluation of the sample paragraph using the rule generalization method for idiom detection.

Sentence	Clause 0.5	Tense Score Present = 2.5
There is this boy in my neighborhood.	0	Simple Present 2.5
He is Very strangely	0	Present

dressed and remains quite aloof		2.5
He has tattoos all over his body.	0	Simple Present 2.5
But you know what they say, don't judge a book by its cover.	0.5	Simple Present 2.5
I went ahead and tried to converse with this boy.	0	Past 0
And I was right, this boy indeed had a very interesting nature.	0	Simple Past 0
He was just a teenager in his growing phase.	0	Past Continuos 0
That is when I realized, to understand others we need to put ourselves in other people's shoes.	0.5	Simple Present 2.5

Table 2: Evaluation of sample paragraph using the tense rule and clause rule generalization method

Table 2 checks each statement in the paragraph for the presence of a clause and the tense of the statement. The scoring of the criteria is currently done manually on the 5 point scale depending on how much a rule can actually determine the presence of a proverb. The presence of clause is given the score of 0.5 because, simply presence of clause is very common in a paragraph and in itself is not a very good indication of the statement being a proverb. Whereas, if a statement is in present tense, possibility of statement being a proverb is increased, therefore a score of 2.5 is given if the statement is in present tense. Similarly the scores have been assigned for each criterion according to the capacity of the given criteria to determine the statement as a proverb. Table 3 checks each statement for the presence of a clause along with presence of certain words common in the idiomatic sentences. It also checks presence of the personal pronoun (PP) and the subject pronoun (SP) and adds or decreases the score according to the rule.

Sentence	Clause + that 1	Words: Realized, told, said, knew, say - 0.5	Voice Active – 1.5 Passive - 0	PP + SP (+1) Only PP (-1)
There is this boy in my neighborhood.	0	0	1.5	0
He is Very strangely dressed and remains quite aloof	0	0	0	-2
He has tattoos all over his body.	0	0	1.5	-2
But you know what they say, don't judge a book by its cover.	0	0.5	1.5	0
I went ahead and tried to converse with this boy.	0	0	1.5	0
And I was right, this boy indeed had a very interesting nature.	0	0	1.5	0
He was just a teenager in his growing phase.	0	0	1.5	-2
That is when I realized, to understand others we need to put ourselves in other people's shoes.	1	0.5	1.5	0

Table 3: Evaluation of sample paragraph using the active passive voice and pronouns rules

Sentence	Total Score
There is this boy in my neighborhood.	3.5
He is Very strangely dressed and remains quite aloof	0.5
He has tattoos all over his body.	2
But you know what they say, don't judge a book by its cover.	4.5
I went ahead and tried to converse with this boy.	1.5
And I was right, this boy indeed had a very interesting nature.	1.5
He was just a teenager in his growing phase.	-0.5
That is when I realized, to understand others we need to put ourselves in other people's shoes.	6

Table 4: Evaluation of sample paragraph using the rule generalization method- final total of all rule criterions

After the analysis using the generalized rules method, we have rated the sentences based on whether they follow various grammar rules usually followed by most of the idiomatic sentences. Using the classification method, we have further analyzed and rated the same statement based on the difference in its semantic context with the context of the entire paragraph. Therefore a statement getting a highest score using these two methods can be assumed to have higher probability of being an idiomatic statement.

Based on the analysis of the rule based and classification method, the total score was calculated as the combination score of the grading assigned by the two methods as shown in the following table.

Sentence	Classification method	General language rule based method	Total score
There is this boy in my neighborhood.	3.5	3.5	7
He is Very strangely dressed and remains quite aloof	3.5	0.5	4
He has tattoos all over his body.	1.5	2	3.5
But you know what they say, don't judge a book by its cover.	3.5	4.5	8
I went ahead and tried to converse with this boy.	1.5	1.5	3
And I was right, this boy indeed had a very interesting nature.	3.5	1.5	5
He was just a teenager in his growing phase.	4.5	-0.5	4
That is when I realized, to understand others we need to put ourselves in other people's shoes.	1.5	6	7.5

Table 5: Total score calculated by the combination score of the two methods

As we can see, the two idioms in the paragraph got detected with the highest score. A data set consisting paragraphs from different domains was tested using the similar method of implementation as mentioned above with the sample paragraph.

5. Results

Various paragraphs containing an idiomatic statement were tested based on the method specified in the implementation section. The sample data contained paragraphs of various lengths. First the classification method was applied on every statement of the paragraph and scores were assigned based on difference in categories. This was followed by the grammar rule generalization method used to test all the sentences in the same paragraph. A score was assigned based on the number and weightage of criterions that the statement follows.

It was observed that most of the proverbs were detected with the highest scoring in the analysis. While for few of the paragraphs, the above method failed to detect idioms correctly. It was also observed that as the number of idioms in a paragraph grew higher, the algorithm failed to work. But considering the fact that usually number of idioms in a paragraph is usually limited, this algorithm can be assumed to work in general for most of the cases.

6. Conclusion

This paper proposed the idea of using categorization and generalization based on rules in order to detect idioms in a given paragraph with more efficiency. The classification method included comparing the category of the entire paragraph with the category of the statement and determining the highest difference between the categories. The generalization method included algorithm to determine whether a sentence follows certain common POS based rules followed by most sentences which are idioms. This paper also proposed the use of POS tags based ontology for the storage of idioms and their corresponding translations with similar meaning in other languages.

7. Bibliographical References

Bradbury N., (2002). Transforming Experience into Tradition: Two Theories of Proverb Use and Chaucer's Practice. In Oral Tradition (Volume 17 Issue 2, 2002), pages 261-289.

Brahmaleen S., Singh A. and Goyal V., (2010). Identification of Proverbs in Hindi Text Corpus and their Translation into Punjabi. In Journal of Computer Science and Engineering, (Volume 2 Issue 1, July 2010), pages 32-37.

Coinnigh M., (2013). The Heart of Irish-Language Proverbs? An Linguo-Stylistic Analysis of Explicit Metaphor. In Proverbium: Yearbook of International Proverb Scholarship (volume 30, 2013), pages 113-150.

Dash B., (2016). Filching commonality by translation of proverb in Indian Linguistic Scene. Translation Today (Volume10, Issue-I, June 2016), pages 15-32.

Goyal V. and Priyanka, (2009). Implementation of Rule Based Algorithm for Sandhi-Vicheda of Compound Hindi Words. International Journal of Computer Science Issues, Volume 3(2009), pages 45-49.

Goyal V. and Sharma M., (2011). Extracting proverbs in machine translation from Hindi to Punjabi using regional data approach. International Journal of Computer Science and Communication (Volume 2, No. 2, July-December 2011), pages 611-613.

Mukerjee A., Soni A., and Raina A., (2006). Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. In proceedings of Coling/ACL Workshop on Multi-Word Expressions, Sydney, July 23, 2006.

Pisharoty D., Sidhaye P., Utpat H., Wandkar S. and Sugandhi R.,(2012). Extending Capabilities of English to Marathi Machine Translator. IJCSI International Journal of Computer Science Issues, (Volume 9, Issue 3, No 3, May 2012), ISSN (Online): 16940814

Sriram V, (2005). Relative compositionality of multi-word expressions: a study of verb-noun (V-N) collocations. In Proceedings of International Joint Conference on Natural Language Processing - 2005, Jeju Island, Korea.

Sriram V and Joshi A., (2005). Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In Proceedings of Human Language Technology Conference/Conference on

Empirical Methods in Natural Language Processing (HLT/EMNLP) - 2005, Vancouver.

Sriram V and Joshi A., (2004). Recognition of Multi-word Expressions: A Study of Verb-Noun (V-N) Collocations. In the proceedings of ICON 2004, Dec 2004, Hyderabad, pages 19-22.

Stock O., (1989). Parsing with flexibility, dynamic strategies, and idioms in mind. In Computational Linguistics (Volume 15 Issue 1, March 1989).

Tapanainen P., Piitulainen J. and Jarvinen T., (1998). Idiomatic object usage and support verbs. In the proceedings of 36th Annual Meeting of the Association for Computational Linguistics and Coling/ACL 17th International Conference on Computational Linguistics (volume 2, August 1998), pages 1289-1293, Montreal, Quebec, Canada.

Toutanova, A., Klein D., Singer Y. and Manning C., (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of 2003 HLT-NAACL, pages 252-259.

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