Identification of Idiomatic Expressions in Konkani Language Using Neural Networks

Naziya Shaikh, Jyoti Pawar
Government College Borda, Goa University
Margao Goa India, Panjim Goa India
naziya.gcq@gmail.com, jdp@unigoa.ac.in

Abstract

The task of multi-word expressions identification and processing has posed a remarkable challenge to the natural language processing applications. One related subtask in this arena is correct labelling of the sentences with the presence of idiomatic expressions as either literal or idiomatic sense. The regional Indian language Konkani spoken in the states located in the west coast of India lacks in the research in idiom processing tasks. We aim at bridging this gap through a contribution to idiom identification method in Konkani language. This paper classifies the idiomatic expression usage in Konkani language as idiomatic or literal usage using a neural network-based setup. The developed system was able to successfully perform the identification task with an accuracy of 79.5 % and F1-score of 0.77.

1 Introduction

India is a vast country with huge number of different languages spoken from one corner of the country to another. As compared to the deep learning research available in English language with respect to various arenas like question-answering, parts-of-speech tagging, idioms and metaphor detection and paraphrasing, etc., the Indian languages have not been studied in terms of natural language processing in that depth. This might be due to lack of digital resources in the past. But that gap is being filled now with large number of datasets and resources in Indian Languages surfacing in recent years through various projects, making the deep learning research feasible in many of the Indian languages. Research on classification of language constructs like idioms, similes, metaphors, etc. is still in its base form with current research in this context still referring to use of rule-based methods. This motivated us to test the application task of identification of idiomatic sense in the statements with the presence of potentially idiomatic expressions in Konkani language using the neural network. Konkani is a language spoken by over 2.5M speakers on the west coastal states of India (census, 2011). The potentially idiomatic expressions refer to class of idiomatic expressions that can be used in a sentence in both literal way as well as the idiomatically sensed way. This makes it difficult to get the clear semantic understanding of the statement. The translation of such sentences becomes a challenging machine translation task.
Consider for example the idiom:

खोबरें करप

kʰɔbrɛ̃ kərəp

This idiom literally means – “to make a product out of crushed dry coconut”. This multi-word expression is used in the idiomatic way to indicate “ruining or messing up of something”. When used in this sense, this idiom has no semantic relation to the constituent words that refer to the product made out of dry coconut. The example sentence using this expression in the idiomatic sense is shown in table 1.

Sentence: ताणें सगल्या कामाचें खोबरे करून दवरलां, आतां आमकां काम जाग्यार उडोवपाक खूब त्रास जाता.
taɳɛ̃ səglja kamat͡ʃɛ̃ kʰɔbrɛ̃ kərun dəvərlã, atã amkã d͡zagjar uɖoʊpak kʰub tras d͡zata
Translation: He has messed up all this work and now we are facing trouble in organizing it.

Table 1: Example sentence with the multi-word expression used in the idiomatic sense.

With this system that can identify the sense in which these idiomatic expressions have been used can be very useful in such a scenario.
The paper is organized into 6 sections beginning with the abstract. This is followed by introduction of the basic idea of this paper. Further, the related

work available in English language and other related Indian languages is explained. This is followed by the section on proposed method using the RNN BiLSTM network. The next section specifies the implementation details including the dataset used and the experimental setup. Then we discuss the results of the experiments, followed by conclusion for this experiment.

2 Related Work

The methods for detection of idiomatic senses in English language sentences range from the use of rule-based techniques to the use of deep neural networks methods. The rule-based techniques focused on the type and token detection depending on the lexical properties of the idioms like lexical fixedness (Fazly et al., 2009). These rule-based techniques for idiom identification were also tested for the Indian language Hindi by (Priyanka and Sinha, 2014). Further methods used the statistical techniques of distributional semantics of a given sentence for idiom identification (Salton et al., 2016, Peng et al., 2019). The deep learning techniques used for idiom identification in English language include use of a neural networks with added input features like parts of speech tagging and the ensemble method that incorporates multiple modules of networks for the classification task.

The most recent shared task on the idiom identification in English and Portuguese language was conducted by (Madabushi et al., 2022) to determine idioms in few-shot, zero-shot, one-shot and fine-tuning settings over the given datasets. The task was to use various methodologies for identification over the provided datasets. The dataset consisted of statements with potential idiomatic usages along with the corresponding previous and the next sentences for the given statement. Each statement was given the classification as idiomatic or literal. Various methods used for this task include use of mBERT along with LSTM and TextCNN (Daminglu,2022), BERT finetuning over feature-based sentence transformer (Itkonen et al., 2022), span-wise identification over BERT model (Yamaguchi et al., 2022), large language models (Jakhotiya et al., 2022) for classification of idiomatic sense in a sentence. Another mBERT and BiLSTM combination network (Tedeschi et al., 2022) was used to identify idioms in 10 different languages. The task of identification as well as localization of potentially idiomatic expressions in a sentence by (Zeng and Bhat, 2021) used multi-level neural model along with attention mechanism combined with the contextual embedding references. Other methods for improved idiom identification in English include the ensemble method to incorporate common knowledge into the BERT structure (Briskilal and Subalalitha,2022).

Although wide range of efforts have been made in this area, they are all mainly limited to English and few other languages like Portuguese. Corresponding research in low-resourced Indian languages is lacking in this arena. The computational research currently done on idiomatic constructions in Hindi includes the use of a rule-base constructed manually based on the idiom properties in Hindi that is used to determine possibility of idiomaticity. (Priyanka and Sinha, 2014).

3 Proposed Methodology

This paper builds a sequential model through a Bi-directional long short term memory neural network for the purpose of idiomatic statement identification in Konkani language. The basic architectural dimensions of the system are depicted in figure 1.

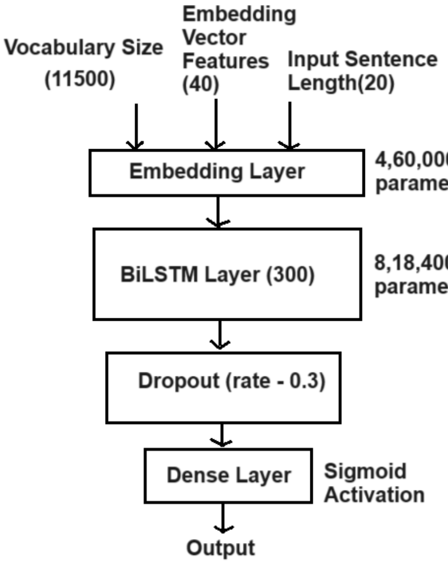

Figure 1: System Architecture.

The method depends on the feature extraction of the sentences in the training data through static embeddings using the sequence information in both directions. The static embeddings generated from the data using one-hot representation as the base embedding were further passed on to a

BiLSTM network for extraction of the features as per the binary classification associated with the labels ‘idiom’ and ‘literal’ provided in the training dataset. This model was trained using the gradient approach with Adam optimizer based on the logarithmic loss function. The output from this layer was connected to the dense layer with the sigmoid activation to derive the final output classification in form of binary labels ‘idiom’ or ‘literal’.

4 Implementation Details

4.1 Dataset Used

The in-house dataset used for this implementation consists of total 4332 Konkani language sentences out of which 2216 instances are labelled to be of idiomatic sense and 2116 instances are labelled to be of literal sense. The dataset instances consist of Konkani language examples spoken on daily basis in the local dialogue. The dimensions of the dataset used is provided in table 2.

No. of potentially idiomatic expressions:	**817**
Total No. of Sentences:	**4332**
No. of sentences labelled with ‘idiom’ sense:	**2216**
No. of sentences labelled with ‘literal’ sense:	**2116**

Table 2: Dimensions of the in-house dataset of Konkani idioms.

Table 3 shows an example listing from the dataset with the label annotated as an ‘idiom’.

Idiom:	भिजत दवरप bʰid͡zət dəʊrəp Wet/Soaked + (to-keep)
Idiom meaning:	Keep waiting/ on hold
Literal meaning:	To keep for soaking in water
Sentence:	फिरयादिचो वकील हजर नाशिल्लो म्हणून न्यायाधिशान केशीचो निकाल भिजत दवरलो. firjadit͡sɔ vəkil həd͡zər naʃillɔ mʰəɳun vjajadʰiʃan keʃit͡sɔ nikal bʰid͡zət dəvərlɔ. Word-to-word: Complainants advocate present not (that is why) judge (of-case) result soaked kept. Translation: Due to absence of complainant’s advocate, the judge kept the case result on hold.
Label:	Idiom

Table 3: Example sentence in the dataset with annotation as ‘Idiom’ label.

Table 4 shows an example listing from the dataset with the label annotated as ‘literal’.

Idiom:	भिजत दवरप bʰid͡zət dəʊrəp Wet/Soaked + (to-keep)
Idiom meaning:	Keep waiting/ on hold
Literal meaning:	To keep for soaking in water
Sentence:	शेजांन्नीन चण्यांचो रोस करपाक चणे भिजत दवरल्यात. ʃɛd͡zannin t͡səɲjãt͡sɔ ros kərpak t͡səɳɛ bʰid͡zət dəvərlɔ. Word-to-word: (Neighbor-female) (of -the peas) curry (to-make) peas soaking kept. Translation: (Neighbor-female) has soaked peas in water to make curry.
Label:	Literal

Table 4: Example sentence in the dataset with annotation as ‘Literal’ label.

4.2 Experimental Setup

The training data is encoded in one-hot embedding and provided to the embedding layer of the network to generate static embeddings. The embeddings are further connected to the bi-directional LSTM network with 300 layers that trains the parameters for feature extraction. The vocabulary size for the establishment of one-hot vector is considered to be 11500 and the length of the one-hot embedded sequence is fixed to be 20 through the addition of padded sequences. The model is set to train itself for a maximum number of 40 features. The dropout layer has been added with a dropout rate of 0.3 in order to prevent overfitting due to the small size of the dataset. The output layer is the dense layer with the sigmoid activation that provides clear binary classification. The loss measured by this model is the cross-entropy loss adjusted for binary values using the logarithmic loss function and gradient loss calculation optimizer used is Adam optimizer. Total number of 4 epochs were used for completing the training phase. The testing is done through a train-test-split with 33% of the instances reserved for testing.

5 Results and Analysis

The result of this experiment proved the efficiency of using simple static embeddings over the BiLSTM layer for the purpose of classification and usage label identification of the statement containing the idiomatic expression. The system created was able to classify any given sentence as idiom or literal with an accuracy of 79.5%. The dataset being 2% more skewed and imbalanced towards idiom label classification, the F1-score was calculated to be 0.77.

6 Conclusion

This paper uses a BiLSTM network over the static embeddings to classify the sense of any idiomatic expression in a Konkani sentence. The created system was able to successfully classify the usage of the idiomatic expressions in the Konkani sentence into the binary labels – ‘idiom’ and ‘literal’ with reasonable accuracy. This model serves as the first step using the neural networks towards the research of multi-word expressions in Konkani language. The future work can include the inculcation of contextual embeddings and attention mechanisms for the classification and processing of idiomatic statements.

References

Afsaneh Fazly, Paul Cook, Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics vol.35*. ©2009 Association for Computational Linguistics.

Agrawal, R., Kumar, V., Muralidharan, V. and Sharma, D. (2018). No more beating about the bush: A Step towards Idiom Handling for Indian Language NLP. *In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Mayazaki, Japan, May. European Language Resource Association (ELRA).

Atsuki Yamaguchi, Gaku Morio, Hiroaki Ozaki, Yasuhiro Sogawa. Hitachi at SemEval-2022 Task 2: On the Effectiveness of Span-based Classification Approaches for Multilingual Idiomaticity Detection. *In the Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 135 – 144, July 14-15, 2022. ©2022 Association for Computational Linguistics.

Daminglu. 2022. daminglu123 at SemEval-2022 Task 2: Using BERT and LSTM to do Text Classification. *In the Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, July 14-15, 2022. © 2022 Association for Computational Linguistics.

Grace Muzny, Luke Zettlemoyer. 2013. Automatic Idiom Identification in Wiktionary. *In the proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, 1417-1421. 18-21 October 2013. ©2013 Association for Computational Linguistics.

Giancarlo D. Salton, Rober J. Ross, John D. Kellcher. 2016. Idiom Token Classification using Sentential Distributed Semantics. *In the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 194-204. August 7-12, 2016. ©2016 Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and Methods for Exploration of Idiomaticity in Pre-trained Language Models. *EMNLP 2021*, Punta Cana, Dominican Republic, 3464-3477. © Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. *In the Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval2022)*. © Association for Computational Linguistics.

J Briskilal, C. N. Subalalitha. 2022. An ensemble model for classifying idioms and literal texts using BERT and RoBERTA. *Information Processing and Management*, volume 59, issue 1, Jan. 2022, 102756.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova . 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In the proceedings of the 2019 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and short papers), Minneapolis, Minnesota, 4171-4186. ©Association for Computational Linguistics.

Jing Peng, Katsiaryna Aharodnik, Anna Feldman. 2018. A Distributional Semantics model for Idiom Detection: The case of English and Russian. *In the Proceedings of the Conference on special session on Natural Language Processing in Artificial Intelligence*. https://doi.org/10.5220/0006733806750682.

Manali Pradhan, Jing Peng, Anna Feldman, Bianca Wright. 2017. Idioms: Humans or machines, it's all about context. *In the Proceedings of CICLing 2017*, Budapest, Hungary. April 2017.

Miriam Amin, Peter Fankhauser, Marc Kupietz, Roman Schneider. 2021. Data-driven Identification of Idioms in Song Lyrics. *In the Proceedings of the 17th Workshop on Multiword Expressions*, Bangkok, Thailand, 13-22. August 6, 2021. © Association for Computational Linguistics.

Priyanka and R. M. K. Sinha. 2014. A system for identification of idioms in Hindi. *In the Proceedings of the Seventh International Conference on Contemporary Computing (IC3)*, Noida, India, 2014, pp. 467-472. https://doi.org/10.1109/IC3.2014.6897218

S. Abarna, J. I. Sheeba, S. Pradeep Devaneyan. 2022. An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning. *Measurement: Sensors 24 (2022) 100434*. © 2022. The Authors. Published by Elsevier Ltd.

Sami Itkonen, Jörg Tiedemann, Mathias Creutz. 2022. Helsinki-NLP at SemEval-2022 Task 2: A Feature-Based Approach to Multilingual Idiomaticity Detection. *In the Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 122 – 134, July 14-15, 2022. ©2022 Association for Computational Linguistics.

Statement 1: Abstract of speakers' strength of languages and mother tongues-2011. Retrieved 7 June 2023 from www.censusindia.gov.in, Office of the Registrar General and Census Commissioner, India.

Simone Tedeschi, Federico Martelli, Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. *Findings of the Association for Computational linguistics: NAACL 2022*, 2715-2726. July 10-15, 2022. ©2022 Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, Vered Shwartz. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, vol. 10, 589-606, 2022. © 2022. Association for Computational Linguistics.

Yash Jakhotiya, Vaibhav Kumar, Ashwin Pathak, Raj Shah. SemEval-2022 Task 2: It Takes One to Know One? Idiomaticity Detection using Zero and One-Shot Learning. *In the Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States,165-168. July 2022. © 2022 Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.semeval-1.19

Yuanchao Liu, Bingquan Liu, Lili Shan, Xin wang. 2018. Modelling context with neural networks for recommending idioms in essay writing. *Neurocomputing* 275 (2018) 2287-2293.

Yuri Bizzoni, Marco S. G. Senaldi, Alessandro Lenci. 2018. Finding the Neural Net: Deep-learning Idiom Type Identification from Distributional Vectors. *Emerging Topics at the Fourth Italian Conference on Computational Linguistics* (Part 1), 28-41.

Ziheng Zeng, Suma Bhat. 2021. Idiomatic Expression Identification using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546-1562.