IDIOMATIC EXPRESSION IDENTIFICATION

B. Tech. Project Stage-II Report

Submitted by

SHRIJEET VILAS RAMTEKE	**111803140**
RUTVIK GANESH MOHARIL	**111803166**
AKASH ANANDA MALAVEKAR	**111803171**

Under the guidance of

Prof. Vibhavari Kamble

College of Engineering, Pune

DEPARTMENT OF COMPUTER ENGINEERING

AND

INFORMATION TECHNOLOGY,

COLLEGE OF ENGINEERING, PUNE-5

April-May 2022

DEPARTMENT OF COMPUTER ENGINEERING

AND

INFORMATION TECHNOLOGY,

COLLEGE OF ENGINEERING, PUNE

CERTIFICATE

Certified that this project, titled "IDIOMATIC EXPRESSION IDENTIFICATION" has been successfully completed by

SHRIJEET VILAS RAMTEKE	**111803140**
RUTVIK GANESH MOHARIL	**111803166**
AKASH ANANDA MALAVEKAR	**111803171**

and is approved for the partial fulfillment of the requirements for the degree of "B.Tech. Computer Engineering/Information Technology".

SIGNATURE
Prof. Vibhavari Kamble
Project Guide
Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.

SIGNATURE
Dr. Vahida Attar
Head
Department of Computer Engineering
and Information Technology,
College of Engineering Pune,
Shivajinagar, Pune - 5.

Idiomatic Expression Identification

ORIGINALITY REPORT

10%	9%	5%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	aniketbhadane.github.io Internet Source	4%
2	towardsdatascience.com Internet Source	2%
3	fugumt.com Internet Source	2%
4	www.esro.org Internet Source	1%
5	www.aclweb.org Internet Source	<1%
6	"Chinese Computational Linguistics", Springer Science and Business Media LLC, 2021 Publication	<1%
7	A Fathima Shirin, C Raseek. "Replacing Idioms Based on Their Figurative Usage", 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), 2018 Publication	<1%

Project Guide

Kamble
13/04/22

Abstract

A sentence is said to be an idiom when its literal and contextual meanings are different. These forms of statements play an important function in the tasks of information abstraction and tasks associated with system understanding. There is a lack of available datasets for training and testing this problem of identifying the nature of sentence, whether it is Literal or Contextual. Knowing their ability of non-literal behaviour and the capacity to take multiple means relying on the context, they've proved as an emerging challenge in NLP systems. To address this challenge we detect whether or not the sentence used is idiomatic in its context or not. Previous attempts at this problem used contextual and non-contextual approaches. Taking a lesson from the ones we suggest a model which makes use of Contextual in addition to Static Embeddings for the Idiomatic Expressions Identification Task.

The method proposed uses a multi-stage neural network architecture to solve the problem. The network effectively fuses static and contextual meanings of each sentence and generates an embedding which is further given as input to the next layer. The final output obtained is the probability of the sentence being used in the literal or contextual form.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today is the age of communication and data. In this fast-changing world data and information play a very crucial role in every course of life and technology. Hence understanding and interpreting language is a crucial part of different systems. There are different models built to understand and interpret different languages that we use in our day-to-day life. English being the most commonly used and most popular one, we try to take a part of the problem from this language and find a solution to it. There are various models built to understand and interpret the English language, but some aspects of the English language are not being solved completely. Idioms are one of them. As they have contextual differences from the original sentence, interpreting them is a difficult task. We try to take this as a challenge and find a solution to it.

1.1 What are Idioms?

Idioms are often referred to as multi-word expressions. But the difference between the multi-word expression and phrase is that MWE's actual meaning is totally different from the actual meaning of the words present in MWE. For example, "you are the apple of my eye", one cannot interpret the meaning of the utterance by interpreting the meaning of each word in the sentence. We can clearly see that the meaning of the sentence considering the context and actual meaning of individual words are totally different.

1.2 Distinguishing between Idioms and other Phrases

To distinguish idioms from other phrases and sayings, the following properties can be considered: Conventionality, Inflexibility, Figuration, Proverbiality, Informality, Affect. By making use of these properties one can establish a meaningful understanding and can differentiate between idioms and normal phrases. Conventionality states that the entire meaning of an idiom can't entirely be predicted from the words, whereas figuration states the use of metaphors, hyperboles types of figures of speech objectifying figurative meaning and forming a connection between them.

1.3 Potentially Idiomatic Expressions(PIE)

By definition, a multiword expression(MWE) is idiomatic in the sense that its true meaning cannot be interpreted from just the meaning of its individual words. However, while sometimes a sequence of words corresponding to an MWE only has the idiomatic interpretation (e.g. for and up), there is often also a literal interpretation of the same sequence, resulting in ambiguity. Multi-word expressions are commonly referred to as potentially idiomatic expressions (hereafter, PIE) and the correct meaning of a PIE in the context proves crucial for many downstream activities, including sentiment analysis, automatic spelling correction, and machine translation. Following this idea, we examine the applicability of semantic compatibility to identify potentially idiomatic expressions in a sentence.

Chapter 2

Literature Review

Title of The Paper	Journal/Conference Details	Methodology Used	Proposed Idea	Advantages/Achieved Objectives in Paper	Disadvantages/Limitations	Accuracy
Multi-Task Learning for Metaphor Detection with Graph Convolutional Neural Networks and Word Sense Disambiguation[1]	The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)	The multi-task model for metaphor detection in this work consists of four modules: (i) the encoding module, (ii) the graph convolution module, (iii) the control module, and (iv)the multi-task learning module	The goal of the multi-task learning module in this section is to transfer the knowledge from the datasets for word sense disambiguation to improve the performance for our metaphor prediction task.	The aim is to achieve state-of-the-art performance on several benchmark datasets for metaphor detec-tion. The experimental results demonstrate the effectiveness of the components proposed in this work	Predicting the metaphor labels of the context words (as in the sequential labeling setting) is suggested by the previous work (Gao et al. 2018; Mao, Lin, and Guerin 2019) as the better way to do metaphor detection than the classification setting	83.2%
A Generalized Idiom Usage Recognition Model Based on Semantic Compatibility [2]	The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)	When the literal interpretation of a potential idiomatic ex-pression is not compatible with the context, it typically indi-cates that the idiom is used figuratively.	(1) the model is first trained on large raw text corpora (such as Wikipedia) with the aim of predict-ing the semantic compatibility between context and a single word. (2) the learned model can then be applied to deter-mine an idiom's intended usage by measuring the seman-tic compatibility between the idiom's literal sense and the context.	The aim is to develop a generalized model for idiom usage recognition based on semantic compatibility. Then it is shown how the model is used for adapting the CBOW along with semantic compatibility. Finally it is described how the model is used for idiom usage recognition.	1.] A lack of Sequential information 2.] Not all words are equal 3.] A paradox of transitivity	76%
A New Approach for Idiom Identification Using Meanings and the Web[3]	Recent Advances in Natural Language Processing (RANLP)	Idioms are sentences whose properties of individual words in a phrase differ from the properties of the phrase in itself.	The recreated definitions which are obtained are subtracted from the definitions of the individual words. If after subtraction the union or the intersection of the two sets is not null then the word is an idiom.	The paper proposes a new idiom identification technique, which is general, domain independent and unsupervised in the sense that it requires no labeled datasets of idioms.	In certain cases, not all words of all the idioms are joined together with either of the special characters and parts of idioms were tagged as MWEs. The method is useful only if we have a definition of each and every word. This was not the case in test cases where the model failed to recognize the idiom.	83.01%

Enhancing Metaphor Detection by Gloss-based Interpretations [4]	Findings of the Association for Computational Linguistic	MDGI-Joint employs two encoders to generate the contextual representation and the gloss representation for a target word. The probability distribution over all candidate glosses is computed by an attention mechanism. The gloss with the highest probability is predicted as the interpretation of the target word.	For the target word a set of candidate glosses are collected from an existing dictionary. The gloss with the highest probability is selected as the interpretation for the word. The weighted sum of these gloss representations undergo concatenation of the contextual representation.	The proposed model achieves similar results comparable with BEM. MDGI-Joint-S is even slightly superior to BEM. The implemented BEM model does not share parameters between the two encoders. Therefore, it can be confirmed that the improved performance of MDGI-Joint-S comes from capturing the interaction between the two tasks	This model addresses the metaphor detection task only and neglects the metaphor interpretation task. It detects metaphors based on the contextual representation of words only.	81.3%
Idiom Token Classification using Sentential Distributed Semantics [5]	Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics	The paper explores the possibility whether the representations generated by Sent2Vec encodes features that are useful for idiom token classific	The encoder takes an input sentence and maps it into a distributed representation. Sent2Vec, is used to "predict" the sentences surrounding the input sentence. Sent2Vec encoder learns (among other things) to encode information about the context of an input sentence without the need of explicit access to it.	The Model needs less contextual information than the state-of-the-art method and achieves competitive results, making it an important contribution to a range of applications that do not have access to a full discourse context.	The general classification model has low performance compared to the other classification models.	88%
Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions [6]	Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 72–80.	Contextualized word embeddings are used to directly detect idiomaticity in potentially idiomatic expressions. MWE is treated as a token during training and testing the models.	The model uses Word2Vec as non-CWEs and leverages the Context2Vec and BERT as CWEs in combination with the Idiom Principle to detect idiomaticity in potentially idiomatic expressions.	Context2Vec and BERT outperform the non-CWE, i.e. Word2Vec, up to 7% higher average F-score. Both models achieved higher F-scores in detecting literal sense of MWEs. As for Context2Vec, the results improved by 6% on average and up to 10% in detecting literal sense of MWEs.	The Model is trained on a small corpus. There exists an inability of the model to provide quality embeddings for rare words. as treating each MWE as a single token turns it into a rare word for which the models need to learn an embedding	77%

MICE: Mining Idioms with Contextual Embeddings [7]	International Conference on Software Engineering and Service Science (ICSESS) 2022	Two state-of-the-art deep neural network approaches to contextual embeddings, ELMo and BERT, followed by the proposed neural network architectures for identification of IEs and their Bayesian ensemble.	In the proposed approach, MICE, ELMo and BERT embeddings are used as an input to a neural network and show that using them as the first layer of neural networks improves results compared to existing approaches. This is the first attempt to combine BERT and ELMo embeddings using an ensemble approach for idiom detection	1.]The first system to successfully recognize IEs not present in the training set. 2.] The first system to successfully analyze both sentence-level and token-level IE detection. 3.] The first successful cross-lingual approach for detection of IEs. 4.] The first Bayesian ensemble approach to combine ELMo and BERT-based models.	Only used embeddings that were pre trained on general text and were not fine-tuned for the specific task of detecting idiomatic language. Several authors have shown that specializing embeddings for specific tasks can improve results on a variety of NLP tasks	76%
Unsupervised Type and Token Identification of Idiomatic Expressions [8]	Computational Linguistics, 2008	We use the observed connection between idiomaticity and flexibility to devise statistical measures for automatically distinguishing idiomatic verb+noun combinations from literal phrases. More specifically, we aim to identify verb–noun pairs such as keep, word as having an associated idiomatic expression and also distinguish these from verb–noun pairs which do not have an idiomatic interpretation	Unsupervised methods that rely on automatically acquired knowledge about idiom types to identify their token occurrences as idiomatic or literal	Lexical and syntactic fixedness are good indicators of idiomaticity, better than a simple measure of collocation such as PMI. On TESTall, Model performs only slightly better than the random baseline (8% error reduction), reflecting that a position-based fixedness measure is not sufficient for identifying idiomatic combinations	For expressions with a relatively high proportion of literal usages, automatically acquired canonical forms are less accurate and also have low predictive value	74%
Mining Idioms in the Wild [9]	DeepAI,2021	A new canonicalized tree representation based on inexpensive dataflow to mine semantic idioms. The approach takes as input a code corpus, generates a tree for each method augmented with dataflow information and uses Bayesian learning methods to mine idiomatic patterns.	Mining trees at method level are used to capture refactoring idioms, it is sufficient to keep the control flow structure and collapse control-free sequence code to a region. The proposed model goes directly from code ASTs to pTSG, without any tree augmentation	Model first scrape edits in a given time interval and API from Facebook's code repository and construct a dataset for mining refactoring patterns using before versions of the edits. **Then the model collects code edits whose API keyword occurrence in after version is higher compared to before.**	The major drawback is that edits collected using this approach would, most likely, contain excessive noise, i.e., changes that are not relevant to the refactoring changes	62%

A New Approach to Instruction-Idioms Detection in a Retargetable Decompiler [10]	Computer Science and Information Systems, 2014	LLVM IR is a set of low-level instructions similar to assembly instructions. Moreover, LLVM IR is platform-independent and strongly typed, which meets our requirements. Therefore, machine instructions from different architectures can be easily mapped to sequences of LLVM IR instructions. This brings an ability to implement platform independent instruction-idioms analysis.	Every instruction that follows has to use the expected operands and results. If an instruction idiom is found the inspection continues after the last instruction belonging to this detected idiom	All the three mentioned architectures are described as instruction accurate models in the ISAC language in order to automatically generate our retargetable decompiler.	The front end approach was not optimal for complex programs and architectures.	
An ensemble model for classifying idioms and literal texts using BERT and RoBERTa [11]	ELSEVIER,2022	The predictions of the two models BERT and RoBERTa are combined and the weighted average method used to predict the final class labels, namely, idioms and literals. The model is also tested with our in house dataset.	The paper proposes a predictive ensemble model to classify idioms and literals using BERT and RoBERTa, fine-tuned with the TroFi dataset.	The model can be used for text classification tasks such as sentimental analysis and movie reviews, and enhanced to act as a multi classifier for text classification tasks demanding multiclass identification. Thus, robust applications can be built atop this classifier.	Models like BERT and RoBERTa have not been exclusively used for idiom and literal classification. This is the first attempt in doing so.	90%
Accurate Design Pattern Detection Based on Idiomatic Implementation Matching in Java Language Context [12]	IEEE,2019	The proposed approach is composed of three sub-processes: (i) Ontology Generation, (ii) Knowledge Inference, and (iii) Pattern Template Matching.	A practical approach for Design Pattern(DP) detection from source code, which exploits idiomatic implementation in the context of Java language. Based on static analysis and inference techniques, the approach enables flexible search strategies integrating structural, behavioral and semantic aspects of DPs for the detection	The model obtains a precision of 85.7% and a recall of 93.8% on average. In all the cases, the precision is greater than 80.6% and the recall is greater than 90.6%.	For the DPs, aggregating all systems, the most time-consuming pattern is Adapter. The main reason is that its pattern template composes a group of implementation variants to offset the lack of unique characteristics, which raises the search cost.	89.5%

Chapter 3

Research Gaps and Problem Statement

Idioms are used to improve effectiveness and are used to convey ideas effectively and artistically when used in everyday speech. These sentences are difficult for reasons that include their non-compositionality (semantic idiomaticity), as well as their actual or contextual meaning depending on the context. These phrases are called Potentially Idiomatic Expressions (PIE).

3.1 PIE and MWE Processing

There are two types of classification in which the idioms are classified. They are Idiom type and Idiom Token Classifications. In the first classification i.e Idiom type, we decide if a sentence can be used as an idiom even without taking into consideration its context. In Idiom Token Classification we decide if a PIE is used with literal intention or there is a hidden contextual meaning to it.

PIEs are actually special types of MWE(Multi-World Expressions). For Multi-word Expression identification a text corpus is given as input and each token is classified whether it is an MWE or not. To do so a tree-based approach is followed in which a syntactic tree is constructed and traversed along with subset of the previously identified MWEs to identify MWEs. This approach even considers the non-compositional phrases whereas we aim to identify only Idiomatic Expressions(IE) from sentences containing PIEs.

3.2 Problem Statement

Idioms are an inevitable part of our everyday communication. As the language improves, new idioms emerge and are added to the huge corpus of the English language. Because idioms take on literal and figurative meanings considering their context in which they have been used, they have proved to be an emerging challenge for NLP systems. To address this, we take care of detecting if an idiomatic expression is present within a sentence and attempt to locate it when it appears figuratively. The model proposed by us will be trained and tested on the MAGPIE dataset which is the largest collection of Idiomatic Expressions.

Chapter 4

Proposed Methodology/ Solution

The 3 different phases of model are Embedding Phase, Attention Phase, Prediction Phase. In the embedding phase, the input sentence is first tokenized. Then the contextual and static word embeddings are generated. Also POS tagged embeddings are generated to get the syntactic information of the input tokens. The attention phase has two attention layers each having a specific purpose. Pytorch framework is used to build the model.

4.1 Embedding Phase

The Model uses two types of Embeddings, Contextual and Non-Contextual.

4.1.1 Contextual Embeddings

BERT

BERT is a bidirectional transformer for pre-training a large amount of unlabeled text data to learn a language representation that can be used to optimize specific machine learning tasks. In the masked BERT language model, only masked tokens (15%) are predicted.

XL-NET

It uses larger data and high computation power. This results in better prediction than the BERT model. To improve the training, XLNet introduces permutation language modeling, where all tokens are predicted but in random order. We used xlnet-large-cased Model of XLNet.

4.1.2 Non-Contextual Embeddings

Character Embedding

All the characters are represented using numbers.

Glove Embedding

Full form of GloVe is global vectors for word representation. We have used the glove.840B.300d model for non contextual embedding. It is the largest glove model.

POS Tag Embedding

Using NLTK, POS of every word is stored in a list. The syntactic information of the sentence is produced by the POS tags of the sentence.

As Mentioned Embedding phase first goes through the process of tokenization. The tokenization takes place as inputs are required for the training of corresponding model. In contextual embedding tokens are generated in accordance with BERT And XL-NET model. For the non-contextual embedding tokens are generated from the glove dictionary. First of all, sentences are tokenized and these tokens are passed to all the embeddings. The output of BERT and XL-NET is passed to the attention flow layer through Linear and Bi-LSTM layers separately. The output of Character embedding and Glove embedding is merged using Highway Network and then passed to the attention flow layer through Bi-LSTM. Similarly, output of Part of Speech embedding is passed to the attention flow layer through Bi-LSTM.

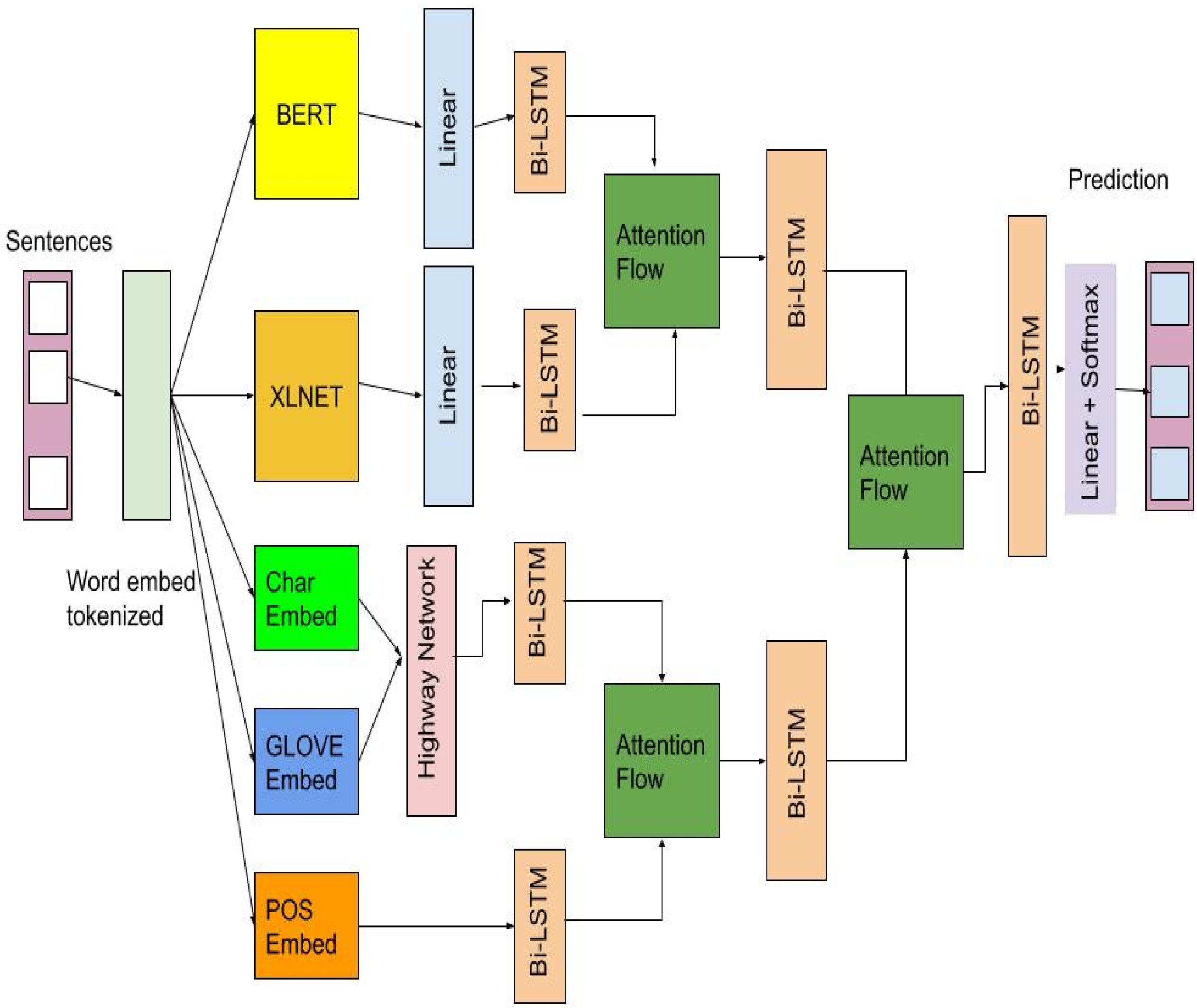
4.2 Attention Phase

The main purpose of the attention is to focus on idiomatic words present in the sentence. It combines the results of previous phases (non-contextual embedding and POS embeddings) to produce literal representations of words present. Now, output of contextual and static embedding is combined using a third attention flow layer by passing through Bi-LSTM.

4.3 Prediction Phase

A BiLSTM layer followed by a linear layer present in this phase. The output of the Attention phase is again passed through Bi-LSTM and Linear layer with Softmax function for the final output. The final output is the probability of each token if it is idiomatic of literal.

Figure 4.1: Architecture of the Model



Chapter 5

Dataset and Experimentation Setup

5.1 Dataset

Here we are using the MAGPIE dataset to train and test our model.. MAGPIE is the largest up-to-date dataset which contains 56,000 instances of sentences each labelled as literal or idiom. The data of the MAGPIE is drawn from various genres of resources such as news, science and also from British National Corpus. There are four major labels which are, literal, idiom, other and unclear. There are 436 instances with other label and only 7 instances with unclear label. The majority being idiom and literal we find it promising for training and testing our model.

5.2 Experimentation Setup

To test the effectiveness of the model we have split the dataset as follows : 32162 for training, 4030 for validation and 4030 for testing. We trained our model for 40 epochs. Batch size is 56 and Learning rate is 1e - 4. Maximum words in a sentence are assumed to be 512. In contextual embedding, we used the “bert-base-uncased” model of BERT and “xlnet-large-cased” model of XL-NET. LSTM dropout rate was set to 0.3

5.3 Language and Libraries

Language

Python v3

Libraries

Pytorch, NLTK, numpy, sklearn, transformers

Chapter 6

Results and Discussion

6.0.1 Results

Training Accuracy after 40 epochs was 89.3% and validation accuracy was 87%

Testing accuracy: 91.9%

Testing precision: 95.6%

Testing recall: 93.7%

Testing f1 score: 94.6%

6.0.2 Discussion

There is scope to improve the accuracy further by training the model for more epochs.

Chapter 7

Conclusion

During the project, we came across combining different models together to improve the overall accuracy. Here we combined contextual embedding models with non-contextual embedding. In the first part we focused on the meanings of the words to form the correct meaning of the sentence. In the second part of the non-contextual embedding we focused on the structure of the sentence. That's how all the static and contextual information of the sentence helped to correctly identify the presence of the idiom in the given sentence. The proposed model achieved the expected results in terms of identifying idiomatic expressions when given an input. Study on the interpretation of the true meaning of language and other figurative/literal constructions such as metaphors, categories that were underrepresented in the data set considered in this study.

Chapter 8

Bibliography

8.1 Dataset

https://github.com/hslh/magpie-corpus

8.2 References

[1]Duong Minh Le, My Thai, Thien Huu Nguyen; ”Multi-Task Learning for Metaphor Detection with Graph Convolutional Neural Networks and Word Sense Disambiguation”; The Thirty-Fourth AAAI Conference on Artificial Intelligence,2020

[2]Changsheng Liu, Rebecca Hwa; ”A Generalized Idiom Usage Recognition Model Based on Semantic Compatibility”; The Thirty-Third AAAI Conference on Artificial Intelligence,2019

[3]Rakesh Verma,Vasanthi Vuppuluri; ”A New Approach for Idiom Identification Using Meanings and the Web”; Recent Advances in Natural Language Processing (RANLP),2015

[4]Hai Wan, Jinxia Lin, Jianfeng Du, Dawei Shen1, Manrong Zhang; ”Enhancing Metaphor Detection by Gloss-based Interpretations”; Findings of the Association for Computational Linguistic, 2021

[5]Giancarlo D. Salton, Robert J. Ross,John D. Kelleher;”Idiom Token Clas-

sification using Sentential Distributed Semantics"; Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016

[6]Reyhaneh Hashempour, Aline Villavicencio; "Leveraging Contextual Embeddings and Idiom Principle for Detecting Idiomaticity in Potentially Idiomatic Expressions";Proceedings of the Workshop on Cognitive Aspects of the Lexicon, Barcelona, Spain, 2020

[7]Tadej Škvorc, Polona Gantar, Marko Robnik-Šikonja; "MICE: Mining Idioms with Contextual Embeddings"; International Conference on Software Engineering and Service Science (ICSESS), 2022

[8]Afsaneh Fazly, Paul Cook, Suzanne Stevenson; "Unsupervised Type and Token Identification of Idiomatic Expressions"; Computational Linguistics, 2008

[9]Aishwarya Sivaraman, Rui Abreu, Andrew Scott, Tobi; "Mining Idioms in the Wild"; DeepAI,2021

[10]Jakubroustek, Fridol; "A New Approach to Instruction-Idioms Detection in a Retargetable Decompiler"; Computer Science and Information Systems, 2014

[11]J Briskilal, C.N. Subalalitha; "An ensemble model for classifying idioms and literal texts using BERT and RoBERTa"; ELSEVIER, 2022

[12]Renhao Xiong, Bixin Li; "Accurate Design Pattern Detection Based on Idiomatic Implementation Matching in Java Language Context"; IEEE,2019