Token-Level Idiom Identification in Turkish and Italian: A Transformer-Based Approach

Efe Can KIRBIYIK and **Berke Kurt**
Department of Artificial Intelligence and Data Engineering
Istanbul Technical University
kirbiyike22@itu.edu.tr, kurtbe21@itu.edu.tr

Abstract

Multi-word expressions (MWEs) pose significant challenges in Natural Language Processing due to their semantic and syntactic complexities. This paper addresses the task of token-level idiom identification in Turkish and Italian sentences, specifically focusing on distinguishing idiomatic usage from literal meaning. First, we present a comprehensive literature review highlighting recent advancements in transformer-based architectures. Building upon these insights, we propose an innovative methodology. Our approach aims to improve upon existing baselines by systematically integrating proven techniques from recent studies.

1 Introduction

Multi-word expressions (MWEs) are combinations of words that work as a unified meaning unit, frequently breaking the semantic and syntactic norms that define their individual components. Although they occur quite frequently in languages, they are difficult to detect using conventional methods. Therefore, one of the most important objectives and significant challenges in the Natural Language Processing is recognizing and comprehending MWEs in context (Tedeschi et al., 2022). Recent studies have explored innovative approaches to MWE identification and found that transformer based architectures are currently one of the most successive tools for MWE detection.

In this paper, we will present the results of our extensive literature review and we will show state of the art in MWE identification techniques. Then we will present our own methodology to tackle the specific challenge of identifying idiom-related tokens within sentences in the ITU NLP Course's competition: https://www.codabench.org/competitions/5986/

2 Related Work

- **ID10M Idiom Identification in 10 Languages (Tedeschi et al., 2022):** This paper introduces a novel multilingual Transformer-based system for idiom identification by reformulating the task as a sequence labeling problem using the BIO tagging scheme. The authors create training datasets by extracting MWEs from Wiktionary for 10 languages, including Italian. A pre-trained multilingual BERT (mBERT) is used to generate token embeddings, which are then fed into a multi-layer BiLSTM network to capture sequence-level information. Finally, a Conditional Random Field (CRF) layer is applied to produce the final sequence of BIO tags. They also classified the MWEs as literal and idiomatic like our case. The dataset does not contain Turkish examples. However, since mBERT supports Turkish in its pre-training, it can be fine-tuned to predict Turkish idioms. We aim to use the methods and insights presented in this paper to solve our problem. We will use their model as a baseline to compare our work.

- **MICE: Mining Idioms with Contextual Embeddings (Škvorc et al., 2022):** This paper presents a novel approach for idiom detection by leveraging contextual embeddings. An ELMo model and two BERT variants (mBERT and CroSloEngual BERT) are used to generate token embeddings that capture rich semantic nuances. These embeddings are fed into a bidirectional GRU layer for sequence modeling, with a softmax output layer to classify tokens as idiomatic or literal. Additionally, a Bayesian ensemble method combines predictions from all models to enhance accuracy. This approach demonstrates strong performance on both token and sentence-level tasks. Although focused on Slovene, the use of multilingual embeddings suggests that with suitable training data, the method can be extended to languages such as Italian and Turkish. We believe using an ensemble of different Transformer variants, like in this paper, will help us. Also, the authors published many test results for different languages. We believe the insight and information in this article will help us.

- **The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (Savary et al., 2017):** This paper introduces a multilingual shared task focused on advancing the automatic identification of verbal multiword expressions (VMWEs). The authors introduces a comprehensive multilingual dataset specifically designed for VMWEs, featuring detailed linguistic annotations across various languages including Turkish and Italic. They also detail various evaluation metrics for the MWE classification task. We plan to use these metrics and the comprehensive dataset to fine-tune our model.

- **An Ensemble Model for Classifying Idioms and Literal Texts using BERT and RoBERTa (Briskilal and Subalalitha, 2022):** This paper proposes an ensemble framework that combines two state-of-the-art Transformer models, BERT and RoBERTa, to classify idiomatic expressions versus literal texts. Each model is fine-tuned on a specially curated dataset designed to capture the nuanced differences in figurative language. The ensemble integrates predictions using methods such as weighted voting, leading to improved performance in distinguishing between idiomatic and literal usages. The approach highlights the benefits of combining diverse Transformer architectures. We might use this cross-domain application.

3 Solution Idea

Our proposed solution is based on recent improvements in transformer-based architectures, especially taking advantage of BERT variants such as mBERT, XLM-RoBERTa, and DistilBERT Gamage et al., 2022, which have demonstrated state-of-the-art performance in idiom identification tasks. Initially, we will replicate the methodology proposed by Tedeschi et al., 2022 in their ID10M framework as our baseline. This approach provides a robust baseline due to its proven effectiveness across multiple languages. To enhance this baseline, we plan to explore several innovative modifications:

3.1 Ensemble of Transformer Models:

Inspired by the work of Škvorc et al., 2022, we plan to use an ensemble approach combining different transformer models such as mBERT, XLM-RoBERTa, and possibly language-specific models like Turkish BERT (Kesgin et al., 2023) and Italian BERT.

3.2 Exploring GRU vs LSTM:

While Tedeschi et al. utilized BiLSTM layers effectively, we intend to investigate whether using a bidirectional GRU layer can yield additional performance gains.

3.3 Pre-Finetuning with Domain-Specific Data:

To further enhance our model's performance, we plan to pre-finetune our transformer models on several datasets including the PARSEME multilingual corpus introduced by Savary et al., 2017 and the ID10M dataset introduced by Tedeschi et al., 2022.

By systematically combining these strategies, we aim to achieve state-of-the-art performance in token-level idiom identification tasks for Turkish and Italian languages while ensuring generalizability across unseen idiomatic expressions.

4 Limitations

Firstly, our biggest constraint is computational power. We think we can apply UHEM to train models, but we don't know whether our application will be accepted or not. Also, we have no idea how long it will take to train the model even if we use UHEM, which might constrain our ability to try different architectures. Secondly, we believe the labels in the training and test datasets are not good, so the model might not be able to learn them.

References

J. Briskilal and C.N. Subalalitha. 2022. An ensemble model for classifying idioms and literal texts using bert and roberta. *Information Processing & Management*, 59(1):102756.

Gihan Gamage, Daswin De Silva, Achini Adikari, and Damminda Alahakoon. 2022. A bert-based idiom detection model. In *Proceedings of the 15th International Conference on Human System Interaction (HSI)*, pages 1–8. IEEE.

Himmet Toprak Kesgin, Muzaffer Kaan Yüce, and Mehmet Fatih Amasyali. 2023. Developing and evaluating tiny to medium-sized turkish bert models. `https://arxiv.org/abs/2307.14134`. ArXiv preprint.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. Id10m: Idiom identification in 10 languages. *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.