NER4ID at SemEval-2022 Task 2: Named Entity Recognition for Idiomaticity Detection

Sai Rohith Pasham
George Mason University
1 PUBLICATION **0** CITATIONS

NER4ID at SemEval-2022 Task 2: Named Entity Recognition for Idiomaticity Detection

Srikanth Reddy Dubba
George Mason University
G01353043
sdubba@gmu.edu

Sai Rohith Pasham
George Mason University
G01348426
spasham@gmu.edu

Abstract

The detection and understanding of idiomatic expressions in natural language are crucial for various language understanding tasks. However, these expressions remain understudied despite their significance. SemEval-2022 Task 2 addresses this gap by focusing on multilingual idiomaticity detection and sentence embeddings. Using a Transformer-based dual-encoder architecture, we focus on the idiomaticity detection subtask and present a novel method for computing the semantic similarity between a potentially idiomatic expression and its context. Based on this, the system predicts idiomaticity. We demonstrate how Named Entity Recognition can be employed to enhance the performance of idiom identification systems by reducing confusion. The proposed model achieves an F1 score of 85.2 in the one-shot setting and shows strong robustness towards previously unseen idioms, achieving an F1 score of 73.7 in the zero-shot setting.

1 Introduction

This section provides an overview of the importance of identifying idiomatic expressions in natural language processing (NLP) and the need for more attention to this phenomenon in existing research. We note that idiomatic expressions, which are multi-word expressions with an established meaning unrelated to the meanings of the individual constituents, are prevalent in all languages and should play an essential role in NLP. The section highlights the SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding, which promotes research on idioms by adapting datasets and tasks from previous work. The task includes two subtasks, one of which is a binary classification task in which potentially-idiomatic expressions (PIEs) must be labeled as either ”Idiomatic” or ”Literal” based on the context they appear in. We propose a novel approach to identifying idiomatic expressions using a dual-encoder Transformer-based architecture and a Named Entity Recognition (NER) module in the classification pipeline. The NER module is used to identify specific words belonging to predefined semantic types, such as Person, Location, and Organization, to avoid errors in ambiguous cases. The section also discusses the limitations of prior work in identifying idiomatic expressions and the potential applications of idiomaticity detection in NLP tasks such as Word Sense Disambiguation, Semantic Role Labeling, Machine Translation, Question Answering, and Text Summarization.

1.1 Task / Research Question Description

This paper presents a system called NER4ID for SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. The task involves identifying idiomatic expressions directly from raw text without pre-identifying potentially idiomatic expressions and studying a broader set of languages. The paper proposes a two-step system that uses Named Entity Recognition (NER) to pre-identify non-idiomatic expressions. It exploits a novel Transformer-based dual-encoder architecture to compute the semantic similarities between the remaining potentially-idiomatic expressions and their contexts and predict idiomaticity. The base paper evaluates the system in both one-shot and zero-shot settings. They are trying to answer how to automatically identify and understand idioms, which is essential for a wide range of Natural Language Understanding tasks, and how to improve the performance of idiom identification systems.

1.2 Motivation and Limitations of existing work

Yes, others have tried to solve the same task of identifying idiomatic expressions in natural language text. Prior work has focused on identifying idiomatic expressions based on their syntactic and lexical fixedness, using measures such as Pointwise Mutual Information (PMI) and collocation statistics. However, these approaches have limitations in identifying idiomatic expressions that do not follow typical syntactic patterns or have low frequency in the corpus. More recent work has focused on using contextual embeddings and neural network models to identify idiomatic expressions. However, these models often require pre-identified potentially idiomatic expressions, a limitation in real-world applications where such words may not be known in advance. The paper presented in this document proposes a novel dual-encoder Transformer-based architecture that encodes the potentially idiomatic expression and its context and predicts idiomaticity based on their similarity. The base paper also includes an auxiliary Named Entity Recognition (NER) module in the pipeline to avoid errors where the individual constituents of a potentially idiomatic expression are unrelated to the context. The proposed approach outperforms strong baselines provided by the task organizers and shows strong robustness towards unseen idioms

1.3 Proposed Approach

The proposed approach seems promising in identifying idiomatic expressions in natural language text. A novel and effective approach uses a dual-encoder Transformer-based architecture that encodes the potentially idiomatic expression and its context, along with an auxiliary Named Entity Recognition (NER) module. The high performance of the proposed system in both one-shot and zero-shot settings and its robustness towards unseen idioms suggests that this approach could be useful in a wide range of Natural Language Understanding tasks.

1.4 Likely challenges and mitigations

Identifying idiomatic expressions in natural language text is a challenging task because idiomatic expressions can have a wide range of syntactic and semantic variations, and their meaning is often unrelated to the meanings of their constituents. Additionally, idiomatic expressions can be highly context-dependent, making it easier to identify them if considering their surrounding context. If the reproduction of the proposed approach turns out to be more complicated than expected or experiments go differently than planned, some contingency plans could include: exploring alternative approaches to identifying idiomatic expressions, such as using unsupervised learning techniques or incorporating additional contextual information and collaborating with other researchers to share knowledge and expertise and identify potential solutions to challenges. Seeking feedback and guidance from someone who is an expert in the field to identify possible solutions to the challenges faced.

2 Related Work

Research on identifying idiomatic expressions in natural language processing (NLP) and how it differs from the proposed approach. One of the initial studies focused on specific syntactic constructions, such as verb/noun idioms, and used the Pointwise Mutual Information (PMI) measure to quantify the degree of lexical, syntactic, and overall fixedness of a given verb+noun combination [Fazly and Stevenson, 2006]. [Cook et al., 2007] and [Diab and Bhutada, 2009] also focused on verb/noun idioms using similar strategies. [Shutova et al., 2010] focused on idioms satisfying specific restrictions, such as subject/verb and verb/direct object. They proposed a method combining lexical, syntactic, and semantic features to identify idiomatic expressions. [Tedeschi et al., 2022] proposed a Transformer-based dual-encoder architecture to compute the semantic similarity between a potentially-idiomatic expression and its context and, based on this, predict idiomaticity. They also showed how Named Entity Recognition could be exploited to reduce confusion in idiom identification systems and improve performance. The proposed approach differs from prior work by exploiting the semantic idiosyncrasy property of idiomatic expressions and by including a NER module to avoid errors in ambiguous cases.

Additionally, the proposed approach aims to identify idioms directly on raw texts without pre-identified potentially idiomatic expressions and study a broader set of languages. We proposed a multilingual approach that can handle idiomatic expressions in multiple languages. It uses a combination of syntactic and semantic features to identify idiomatic expressions, which has yet to be explored in previous studies. We evaluate the proposed approach on multiple datasets in various languages, illustrating its effectiveness and generalizability.

3 NER4ID

NER stands for Named Entity Recognition, which is a subtask of natural language processing that involves identifying and classifying named entities in text into predefined categories such as person names, organization names, locations, dates, and so on. The goal of NER is to extract structured information from unstructured text data. In the context of the paper, the NER module is used as an auxiliary component in the idiomaticity detection pipeline to manage ambiguous cases. Specifically, the NER module is used to pre-identify non-idiomatic expressions that are part of named entities, which can be unrelated to the context in which they are used. By identifying these expressions, the NER module helps to avoid errors in the idiomaticity detection process. The NER module in the paper inputs a raw text sequence containing a potentially idiomatic expression and predicts all the entities in the text sequence. The module uses predefined semantic types such as Person, Location, and Organization to identify specific words as belonging to these types. Overall, NER is a useful tool for extracting structured information from unstructured text data, and in the context of the paper, it helps to improve the accuracy of the idiomaticity detection system by managing ambiguous cases.

3.1 The Dual-Encoder Architecture for Idiomaticity Detection

The dual-encoder architecture is a novel approach to idiomaticity detection that encodes the potentially idiomatic expression (PIE) and its context and predicts idiomaticity based on similarity. To better understand, let's consider two examples where a single idiomatic expression is used:

a) The actor wished his fellow performer to break a leg before going on stage.

b) During the rehearsal, one actor accidentally fell and broke his leg.

In case a, the idiomatic expression "break a leg" conveys a message of good luck or best wishes to the fellow performer before their performance. The words "break" and "leg" in this context have no direct connection to the surrounding context of the stage performance but carry the idiomatic meaning of wishing success. In the case of b, the expression "broke his leg" literally describes a physical injury where the actor's leg was fractured or damaged due to the fall. The words "broke" and "leg" in this context directly connect to the surrounding context of the rehearsal incident and convey the actual injury.

The architecture comprises two BERT-based encoders: an expression encoder (Ψ) and a context encoder (Ω). To encode an expression, the architecture takes the sum of the individual representations of all its subwords. For context, the architecture takes the representation of the [CLS] token. The output of the dual-encoder architecture (Φ) is defined as follows:

$$\Phi(e,c) = \begin{cases} 0, & if \frac{\Psi(e)^T \Omega(c)}{||\Psi(e)||||\Omega(c)||} \leq \delta \\ 1, & otherwise \end{cases}$$

Where δ is a manually-tuned threshold. If the cosine similarity score between the representations of the PIE and its context is lower than the threshold δ, the PIE is labeled as idiomatic. Otherwise, it is labeled as literal. The dual-encoder architecture exploits the semantic idiosyncrasy property of idiomatic expressions to discriminate between idiomatic and literal expressions.

However, there are cases where the individual constituents of a potentially idiomatic expression are unrelated to the context, but the expression used in that particular context is not idiomatic. Many of these cases correspond to named entities. To address this issue, the architecture includes an auxiliary Named Entity Recognition (NER) module in the idiomaticity detection pipeline to avoid errors. Overall, the dual-encoder architecture is a powerful tool for idiomaticity detection that can greatly benefit from including a NER module in the classification pipeline to manage ambiguous cases.

As mentioned earlier, the distinguishing factor between idiomatic and literal expressions lies in their semantic idiosyncrasy. However, there are instances where the individual components of a potentially idiomatic expression have no direct rel-

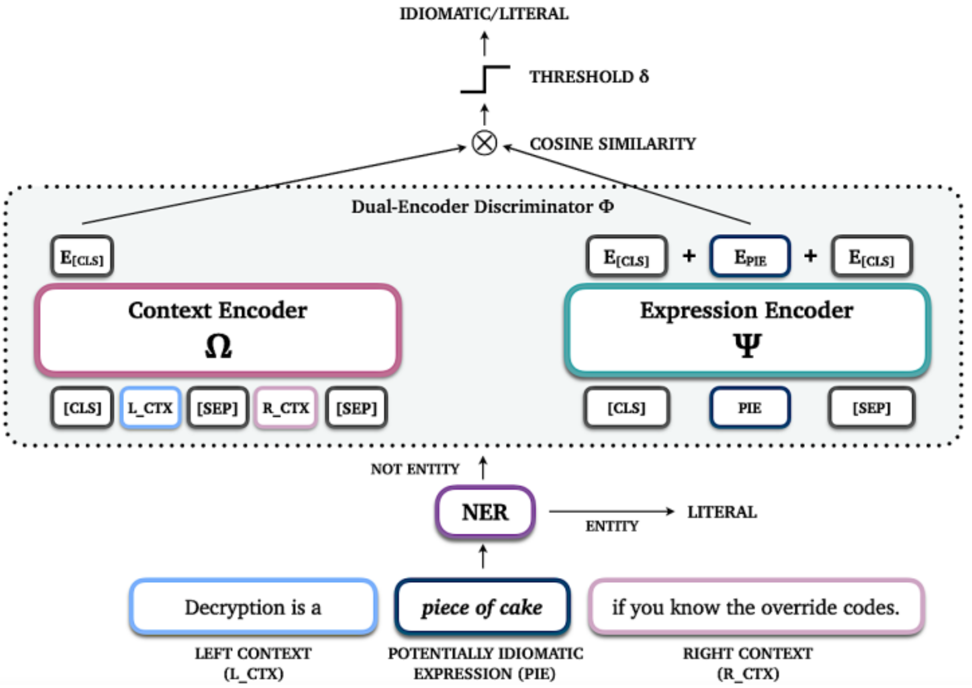

Figure 1: Graphical representation of our architecture for idiomaticity detection. "E" stands for Embedding. A potentially idiomatic expression e is labeled as idiomatic when: i) e is not an entity, and ii) the cosine similarity score between the representations $\Omega(c)$ and $\Psi(e)$, where c is the surrounding context, is lower than the threshold δ

Table 1: Examples of sentences containing potentially idiomatic expressions (PIEs).

PIE	Context
Break the ice	He told a joke to **break the ice** and make everyone feel more comfortable.
Walking on thin ice	After missing several deadlines, she knew she was **walking on thin ice** with her boss and needed to significantly improve.
The ball is in your court	I've given you all the necessary information; now **the ball is in your court**.
Call it a day	We've been working for hours; let's **call it a day** and continue tomorrow.
blood bath	Deborah Loomis is an actress known for Hercules in New York (1970), Foreplay (1975) and **Blood Bath** (1976)
fine line	**Fine** Line received generally positive reviews from music critics, particularly towards its production and stylistic influences.
monkey business	**Monkey Business** is an Action, Adventure, Comedy, Crime movie that was released in 1998 and has a run time of 1 hr 29 min.
rocket science	After finishing "Confrontation", the band shifted to "**Rocket Science**".
night owl	Andrew Gonzalez, owner, **Night Owl** Cookies: "Nobody believed in me except for Deco Drive.","They got me on air very quickly!"
silver spoon	Not only is it endorsed by the UK's biggest food brands – Weetabix, Shredded Wheat, **Silver Spoon**, Carling lager, Marriage's flour – but being Red Tractor also means you can supply different retailers without lots of different requirements.

evance to the context, yet the expression itself, when used in that specific context, does not possess an idiomatic meaning. In other words, there are cases where the literal interpretation of the expression aligns with its contextual usage, even though the individual words may not have an apparent connection to the surrounding context.

‘

4 Experiments

4.1 Datasets

The base paper provided the datasets we used, and they were easily accessible. The datasets are publicly available in their GitHub repository. The datasets were actually provided in the SemEval2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. It was already in the same format as required by the tasks later. There were separate datasets available for one-shot and zero-shot settings. The train data contains around 4492 rows in the zero-shot settings and 210 rows in the one-shot settings. There are eight columns in the dataset: ID, language, setting, multi-word embedding, previous, target, next, and label. ID is the id number for that row; language specifies which language the text is in; setting mentions whether it is one-shot setting or zero-shot setting; multi-word embedding(MWE) consists of the potential idiom; previous column contains the context building up to the targeted sentence; target column has the sentence that potentially can be idiomatic expression which we need to detect; next column consists of the context following the potential idiomatic expression. In the zero-shot setting, potentially idiomatic expressions in the training set are completely disjoint from those in the validation and test sets. Potentially idiomatic terms in the training set are wholly distinct from those in the validation and test sets in the zero-shot setting. The test and validation sets for each MWE are instead comprised of one positive and one negative example in the one-shot context.

4.2 Implementation

The Implementation task requires categorizing potentially idiomatic expressions (PIEs) as either ”Idiomatic” or ”Literal” depending on the context in which they are used. By introducing a two-step system, we are able to successfully complete the idiom identification task. The system uses Named Entity Recognition (NER) to pre-identify non-idiomatic expressions and a novel Transformer-based dual-encoder architecture to compute the semantic similarity between the remaining potentially idiomatic expressions and their contexts and, based on these, predict idiomaticity. Finally, we thoroughly assess both one-shot and zero-shot settings for our system. Two distinct settings are offered to more effectively test the generalization abilities of models: one-shot and zero-shots. We used the code which the author in the GitHub repository provided. We made some changes to reproduce it. The link to our reproduced work is as follows.

https://github.com/sairohithpasham/nlp_project

4.3 Results

Table 2: Results of our reproduced system compared with competitive baseline and base paper authors’ results using the Macro-F1 score metric.

System	Result
Competitive Baseline (zero-shot)	65.4
Basepapers Result (zero-shot)	77.4
Our Reproduced work (zero-shot)	**73.7**
Competitive Baseline (one-shot)	86.5
Basepapers Result (one-shot)	92.1
Our Reproduced work (one-shot)	**85.2**

4.4 Discussion

We encountered a number of problems while trying to replicate this project, such as the need to modify the provided code because it was creating errors. Additionally, it required a few days to get the author’s consent before we could submit the test predictions. With a Mean Squared Error loss criterion, an early stopping strategy, an Adam optimizer, and a 10^{-5} learning rate. We fine-tuned our idiom identification system over the course of 100 epochs.

We set $\delta = 0.4$, utilized a batch size of 32, and accumulated the gradient in 4 steps. Our results differed by around 3 in zero-shot and 7 in one-shot settings compared to the author’s reported results.

This discrepancy could be due to the author running 1000 epochs while we ran around 50 epochs. We attempted to mitigate this difference by changing the random seeds and hyperparameters, reporting the highest accuracy achieved. However, our accuracy still differed slightly even with the same random seed as the author. We performed a sensitivity analysis by changing hyperparameters such as learning rate and batch size. The resultant micro F1-score decreased significantly. The best results were obtained with a learning rate of 0.0001, batch size of 32, 50 epochs, gradient accumulation steps of 1, and a patience of 5.

4.5 Resources

The paper that we have chosen is competition-based. It took us some time to reproduce the paper. Two group members invested their time and expertise in trying to duplicate it. To access the test label, we had to ask permission from the author in CodaLab. The author provided a link to the code repository for the NER4ID system, which can be used to reproduce the results presented in the paper. Producing the results would require significant computational resources and development effort. The system uses a novel Transformer-based dual-encoder architecture involving Named Entity Recognition and semantic similarity computation.

4.6 Error Analysis

Upon performing error analysis, we found a couple of cases where the model fails to identify idiomatic expressions correctly. Two such cases are ”take the cake,” which was labeled as literal despite being an idiomatic expression, and ”in the red,” which is labeled as idiomatic despite being used in a literal sense.

4.7 Robustness

Testing for robustness is the process of evaluating how well a system or model performs under unexpected or challenging conditions, such as changes in the input data, changes in the environment, or adversarial attacks. The purpose of testing for robustness in our system is to ensure that our system or model can handle a wide range of scenarios and that its performance is consistent and reliable. We have referred to the paper [Ribeiro et al., 2020] Beyond Accuracy: Behavioral Testing of NLP Models with CheckList Marco Tulio Ribeiro Tongshuang Wu Carlos Guestrin Sameer Singh to test for robustness in our model. The paper specifically focuses on testing for robustness in models that perform named entity recognition (NER), which involves identifying and classifying named entities (e.g. people, organizations, locations) in text. The authors argue that testing for robustness in NER models is particularly important because these models are often used in real-world applications, such as information extraction and question-answering systems. The authors propose a framework for testing the robustness of NER models, which involves perturbing the input text in various ways (e.g. changing the order of words, replacing words with synonyms) and evaluating the impact of these perturbations on the model's performance. So, we will discuss in detail all the perturbation experiments that we have performed in the following sections:

4.7.1 Character level perturbation:

Character-level perturbation is a technique used in natural language processing (NLP) to test the robustness of machine learning models. In this technique, the text is perturbed by changing or deleting individual characters in the text. This can help identify the robustness of the model to character-level variations in the input data. We manually took the one-shot training data and performed character-level perturbation. We have performed at least 50 changes to the data. Some of the examples are shown in the table below. After predicting the labels on this data, we got a macro f1 score of 66.5. The score decreased from 73.7, which we reached without the data perturbation. This suggests that the model is fairly robust since the score didn't drop significantly. It also suggests that the model performs reasonably well when faced with real-world scenarios where the input data may not be clean or well-formed. View Table 3

4.7.2 Word level perturbation:

In word-level perturbation, we modified the original text data by replacing some words with synonyms, adding or deleting words, or swapping the positions of words within a sentence or document. This perturbation is done with the intention of assessing how well the NLP model can still perform its task when the input data is slightly different from what it was trained on. Word level perturbation is particularly useful for evaluating the generalization ability of NLP models, as it allows for testing the models on data that is not part of the training data. We manually took the one-shot training data and performed character-level perturbation. We have performed at least 50 changes to the data. Some of the examples are shown in the table below. Our score on this data is almost similar to the character level perturbation of 64. However, it's important to note that this slight decrease in score is not necessarily an indication of poor performance. View Table 4.

4.7.3 Sentence level perturbation and all perturbations:

Sentence-level perturbation involves modifying entire sentences in a text corpus by replacing words, swapping clauses, or rephrasing the sentence while maintaining the original meaning. Sentence level perturbation can help in detecting overfitting, where the model has memorized specific sentences in the training data and fails to generalize to new sentences. We performed at least 50 sentence perturbations to the data and got a score of 66. Next, we have performed all the perturbations that we have discussed in the same dataset and got a score of 65.

The goal of testing for robustness is to ensure that the model can handle different types of variations in the data, and a small decrease in performance could be acceptable as long as the model still performs well overall. View Table 5.

Table 3: Character level perturbation

Original	Perturbed
Scripps also built two cottages on the property, one for her library and the other for visitors.	Scripps also buiilt twwo cottages on the propertty, one for herr librrary and thee otherr for visitors.
In the MCU, Old Man Steve Rogers probably doesn't have much to do aside from playing chess down at the retirement home, so perhaps he gets involved in the superhero business once more by taking on a leadership role a la Nick Fury.	In the MCU, Old Man Steve Rogers probably doesn't have much to do aside from playing chess down at the retirement home, so perhaps he gets involved in the superhero business once more by taking on a leadership role a la Nick Fury.

Table 4: Word level perturbation

Original	Perturbed
Scripps also built two cottages on the property, one for her library and the other for visitors.	Scripps also constructed a pair of cottages on the premises, one for her collection of books and the other for guests.
In the MCU, Old Man Steve Rogers probably doesn't have much to do aside from playing chess down at the retirement home, so perhaps he gets involved in the superhero business once more by taking on a leadership role a la Nick Fury.	In the MCU, Old Man Steve Rogers probably doesn't have much to do besides playing chess down at the retirement home, so perhaps he gets involved in the superhero biz once more by taking on a leadership position like Nick Fury.

4.7.4 Testing on real-world data:

We collected 20 rows of data which contain English, Portuguese, and Galician, and annotated them with the help of our classmates. Then, we passed this real-world data to our model and predicted the labels. We got a score of 80 percent. This suggests that the model generalizes fairly well and is relatively robust.

4.8 Multilinguality

Testing for multilinguality in NLP involves evaluating the performance of NLP models on tasks that involve multiple languages. This involves testing the ability of models to process text in different languages, including languages that they were not trained on. It is important because in a globalized world, NLP models need to be able to handle and process text in different languages to be truly useful. By being able to process text in different languages, NLP models can help overcome language barriers and facilitate cross-cultural communication and information access. Moreover, many NLP applications, such as machine translation and sentiment analysis, require multilingual support to be effective. For our model, we tested for multilinguality by creating a new dataset in spanish. It has 20 rows, and it is made by referencing the internet. We have already used the bert multilingual model on 3 languages: English, Portuguese, and Galician. We were getting around 73 score with the bert multilingual model. So, We tried changing the model from Bert multilingual to XLM-

Table 5: Sentence level perturbation and all perturbations

Original	Perturbed
Scripps also built two cottages on the property, one for her library and the other for visitors.	The property also included two cottages, one for her extensive library and the other for accommodating visitors, both of which were built by Scripps herself.
In the MCU, Old Man Steve Rogers probably doesn't have much to do aside from playing chess down at the retirement home, so perhaps he gets involved in the superhero business once more by taking on a leadership role a la Nick Fury.	Maybe in the MCU, Old Man Steve Rogers doesn't have a lot to do other than play chess at the retirement home, so he may get back into the superhero business by taking on a leadership role similar to Nick Fury.

Roberta. We found a slight increase in score from 73 to 76. Also, we applied this model on the Spanish dataset that we created and got a score of 65. This suggests that the model might be performing fairly well on the languages it hasn't been trained on.

5 Conclusion

We feel that this paper is very much reproducible. The author has given good instructions in the readme file and given out comments. The code was well documented. We presented our NER4ID submission, which aimed to detect idiomatic expressions in the Multilingual Idiomaticity Detection subtask. Our novel dual-encoder Transformer-based architecture encoded the potentially idiomatic expression and its context and predicted idiomaticity based on similarity. Through manual error analysis, we identified critical cases where PIEs were part of named entities and introduced an auxiliary NER module to improve the pipeline's performance. We performed testing for Robustness and from the results, we found that the model is fairly robust. As the score only changed slightly. We performed testing for Multilinguality on the Spanish dataset and found that the model performs reasonably well on languages the model has not been trained on. In future work, we plan to explore identifying idioms directly on raw texts and extending our analysis to a wider range of languages.

References

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48. Association for Computational Linguistics, 2007.

Mona Diab and Pravin Bhutada. Verb noun construction mwe token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE 2009)*, pages 17–22, 2009.

Afsaneh Fazly and Suzanne Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1002–1010. Coling 2010 Organizing Committee, 2010.

Simone Tedeschi and Roberto Navigli. NER4ID at SemEval-2022 task 2: Named entity recognition for idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 204–210, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.25. URL https://aclanthology.org/2022.semeval-1.25.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. Id10m: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, Washington, 2022. Association for Computational Linguistics.

[Ribeiro et al., 2020] [Tedeschi and Navigli, 2022]