Lexicons Gain the Upper Hand in Arabic MWE Identification

Najet Hadj Mohamed[1,2], **Agata Savary**[3], **Cherifa Ben Khelil**[1,4],
Jean-Yves Antoine[1,5], **Iskander Keskes**[2], **Lamia Belguith Hadrich**[2]

LIFAT - University of Tours[1]
MIRACL - University of Sfax [2]
LISN - University of Paris-Saclay [3]
EFREI Research Lab - University of Paris Panthéon Assas[4]
LIFO - University of Orleans[5]

najat.hadjmohamed@etu.univ-tours.fr, agata.savary@universite-paris-saclay.fr,
cherifa.ben-khelil@efrei.fr, jean-yves.antoine@univ-tours.fr, iskandarkeskes@fsegs.usf.tn,
lamia.belguith@fsegs.usf.tn

Abstract

This paper highlights the importance of integrating MWE identification with the development of syntactic MWE lexicons. It suggests that lexicons with minimal morphosyntactic information can amplify current MWE-annotated datasets and refine identification strategies. To our knowledge, this work represents the first attempt to focus on both seen and unseen of VMWEs for Arabic. It also deals with the challenge of differentiating between literal and figurative interpretations of idiomatic expressions. The approach involves a dual-phase procedure: first projecting a VMWE lexicon onto a corpus to identify candidate occurrences, then disambiguating these occurrences to distinguish idiomatic from literal instances. Experiments outlined in the paper aim to assess the efficacy of this technique, utilizing a lexicon known as LEXAR and the "parseme-ar" corpus. The findings suggest that lexicon-driven strategies have the potential to refine MWE identification, particularly for unseen occurrences.

Keywords: Multiword Expressions, Idiomatic Expressions, Literal vs. Figurative Meanings, Lexicon Augmentation, Arabic Language

1. Introduction

Multiword Expressions (MWEs) are a subject of interest across various fields related to language studies. They are part of each language's lexicon, distinct from literal words due to their non-compositional, preconstructed nature. Recently, the identification and analysis of MWEs have garnered significant attention in the field of Natural Language Processing (NLP), owing to their prevalence and nuanced semantic complexities. Despite considerable efforts in MWE identification, researchers have encountered challenges in addressing the issue of unseen MWE instances[1] (Taslimipoor et al., 2020; Pasquer et al., 2020b; Yirmibeşoğlu and Güngör, 2020; Kurfali, 2020). Savary et al. (2019) assert that to make substantial progress in MWE identification, it is imperative for the research community to integrate the identification process with the development of syntactic MWE lexicons. They advocate for lexicons that provide minimal morphosyntactic information, augmenting existing MWE-annotated corpora. This approach, they argue, complements traditional corpus-based methods with MWEs that occur rarely or never in MWE-annotated corpora. In this paper, we align ourselves with the same perspective, emphasizing the critical role of MWE lexicons in advancing MWE identification methodologies for Arabic language.

MWEs assume a unique and challenging role within this domain due to their non-compositionality and their ability to take on a figurative or literal meanings. For instance, the degree of transparency varies from one idiom to another. Thus, the following idiom is rather transparent سَلك الطَريق السَريع (salk al-ṭarīq al-sarīᶜ | lit. 'to take the fast road') 'to choose the easier way', i.e. it is easy to recover the motivation behind the image of taking a fast road. Conversely, in كَسَرَ السَيْف (kasara al-saif | lit. 'broke the sword') ' to triumph over an opponent or a difficult circumstance', the motivation for the image is unclear. Moreover, transparency can depend on the particular speaker's knowledge. For instance, the literal reading (e.g.'وَضَعَ يَداً عَلَى الجُرْح' (lit. 'to touch the wound') 'to evoke someone's weakness' is understandable for most speakers, while understanding the origin of the following idiom calls for historic and cultural knowledge: 'براءة الذئب من دم ابن يعقوب' ('brāᵓẗ al-ḏᵓib mn dm abn īᶜqūb' | lit. 'to have the innocence of the wolf from the Jacob's son blood') 'to be innocent'.[2]

[1] No other verbal multi-word expression containing the exact same set of lemmas has been annotated at least once in the training corpus.

[2] This idiom relates to the story of Jacob and his broth-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 88–97. May 25, 2024.

Significant research has been dedicated to detecting metaphors and understanding idiomatic expressions. Metaphors are deliberately constructed to convey figurative meanings, while idiomatic expressions can be interpreted either literally or figuratively, depending on the context of use (Shutova, 2010; Mason, 2004; Liu and Hwa, 2017). The accurate processing of idiomaticity within textual sequences is fundamental in NLP, given that idiomatic expressions constitute a significant aspect of linguistic communication. Attaining high performance in this task holds the potential to enhance various downstream applications, including sentiment analysis, information retrieval, and machine translation (Hashempour and Villavicencio, 2020; Mohamed et al., 2023). In this paper, our main focus is on identifying MWEs using an Arabic lexicon, with the goal of capturing unseen expressions more effectively and reducing the ambiguity of literal interpretations. Thus, we are also interested in the challenge of distinguishing between these two interpretations, which is complicated by the fact that idioms often do not follow easily identifiable linguistic patterns, especially for the Arabic language, given that is characterized by a fairly flexible word order (Hadj Mohamed et al., 2022). While our research primarily focuses on Arabic, we have also tested our model for the binary disambiguation of Potential Idiomatic Expression (PIE) task (see Section 2 on English and German languages. The paper is organized as follows: Section 2 provides a thorough review of existing literature on MWE identification. Section 3 focuses on MWE identification in Arabic. Following that, Section 4 elaborates on our methodology for MWE identification in Arabic, emphasizing the integration of lexicons and the disambiguation process, while Section 5 details the data used in our experiments. Finally, in Section 6, we present and analyze our experimental results.

2. Related work

A considerable amount of research has focused on MWE-specific tasks. In this paper we are primarily concerned with **MWE identification**, which consists in automatically annotating MWE occurrences in running text (Constant et al., 2017). Most approaches to this task are supervised, i.e. trained on manually annotated datasets, such as STREUSLE (Schneider and Smith, 2015) or PARSEME (Savary et al., 2018). Shared tasks such as DiMSUM (Schneider et al., 2016) and PARSEME (Ramisch et al., 2020) boosted the development of such tools. MWE identifiers are then trained and evaluated on these corpora. For instance, two approaches to MWE identification within a transition system were compared in (Al Saied et al., 2019): one based on a multilayer perceptron and the second on a linear SVM. Both approaches utilize only lemmas and morphosyntactic annotations from the corpus and were trained and tested on PARSEME Shared Task 1.1 data (Ramisch et al., 2018). The approach in (Kurfali, 2020) leverages feature-independent models with standard BERT embeddings. mBERT was also tested, but with lower results. An LSTM-CRF architecture combined with a rich set of features: word embedding, its POS tag, dependency relation, and its head word is proposed in (Yirmibeşoğlu and Güngör, 2020). The main focus of PARSEME Shared Task 1.2 was the detection of the unseen Verbal Multiword Expressions (VMWEs) which is more challenging compared to the identification of seen VMWEs (Ramisch et al., 2018). Several systems participated in the shared task, including MTLB-STRUCT (Taslimipoor et al., 2020), TRAVIS-mono and TRAVIS-multi developed by Kurfali (2020), Seen2Unseen developed by Pasquer et al. (2020a), ERMI by Yirmibeşoğlu and Güngör (2020) and others. Notably, the MTLB-STRUCT system, which leverages multilingual BERT fine-tuned for joint parsing and MWE identification, achieved the top cross-lingual macro-average in the open track for both the identification of VMWES and the subtask of identifying unseen VMWEs.

Since unseen VMWEs prove critically hard to identify, a natural idea would be to leverage the advances of **MWE discovery**, which consists finding new MWEs (types) in text corpora, and storing them for future use in a lexicon (Constant et al., 2017). Very many different approaches were devised for this task in the past, based on statistical association measures (Evert, 2005), parsing data (Seretan et al., 2011), lexico-syntactic constraints (Broda et al., 2008), possibly combined with the use of neural network (Pecina, 2010), etc.

An alternative approach in addressing unseen data, and the scarceness of MWE-annotated corpora in general, is to use existing **MWE lexicons**, extracted for instance from classical human-readable dictionaries (Kanclerz and Piasecki, 2022) or Wiktionary (Muzny and Zettlemoyer, 2013), possibly with example sentences contained therein (Tedeschi et al., 2022). Such a lexicon can be straightforwardly projected on a corpus by form/lemma matching. Each resulting word co-occurrence is then considered as a *potential idiomatic expression* (PIE), in the sense that it can be true idiomatic occurrence of a MWE, or just a literal/coincidental co-occurrence of the MWE component words.

The task of **binary disambiguation of PIEs** has been addressed by a number of works. Sporleder

ers, shared by the Jewish, Christian and Muslim religions.

and Li (2009) propose a generalized method utilizing cohesion graphs, hypothesizing that a PIE is used figuratively if its removal improves cohesion. Liu and Hwa (2018) introduce a ”literal usage metric” quantifying the literalness of a PIE, computed as the average similarity between words in the sentence and a literal usage representation. Ehren et al. used a 2-layer LSTM network to get latent representations for the verbal idiom tokens. These were then used in a fully connected layer to predict the class using softmax. They used pretrained static and contextualized word embeddings as an input for their model. In recent years, several shared tasks have been organized to advance research in binary PIE disambiguation. Notably, the Multilingual Idiomaticity Detection and Sentence Embedding shared task (Madabushi et al., 2022) has gained attention. It comprises two subtasks: (a) binary disambiguation of PIEs, and (b) semantic text similarity detection, including sentences with and without MWEs.

3. Arabic and MWEs processing

The ”Arabic language” includes Modern Standard Arabic (MSA) and diverse Arabic dialects. MSA is used in religious texts, poetry, and formal writing, while dialects are spoken in everyday conversation. In this section, we provide an overview of MSA's distinctive characteristics and review previous research on the automatic processing of MWEs in Arabic, with a specific focus on MSA rather than dialectal forms.

In MSA, capitalization is absent, and the usage of punctuation marks is infrequent in contemporary Arabic texts. Additionally, this language commonly features long, complex sentences with right-to-left writing, often resulting in paragraphs that lack punctuation. Furthermore, as a Semitic language, Arabic exhibits a complex morphology. It uses *concatenative morphology (agglutinated or compound words)*, where words are formed via a sequential concatenation process[3]. For example, the sentence *'then they will write it'* is presented in Arabic as one word فسيكتبونها. Moreover, Arabic includes words that can be altered with diacritical marks, either above or below them, creating new words with distinct pronunciations and meanings, often similar to the original word. Consequently, texts lacking diacritical marks are prone to ambiguity.

In Arabic, as in German, the word order is flexible, allowing specific words in a sentence to be rearranged without altering its meaning. This adaptability is achieved through the language's use of case markers, particles, and other linguistic mechanisms to clarify word relationships, resulting in a more versatile syntax compared to languages with a more rigid word order. These unique features make Arabic a challenging language for NLP tasks.

Several studies and research have been conducted on Arabic Multiword Expressions (AMWEs). Attia (2006) explored AMWEs using a finite-state machinery and Lexical Functional Grammar (LFG). During processing, fixed and adjacent semi-fixed MWEs were scrutinized using lexical transducers, deconstructing one-word phrases into segments and integrating MWEs into spaced words. Syntactically flexible MWEs were handled by grammar rules as syntactically compositional but semantically non-compositional due to lexical selection rules. Attia et al. (2010) introduced a linguistic method based on regular expressions for extracting AMWEs from texts, with a specific focus on nominal AMWEs. Hawwari et al. (2014) compiled an AMWE list from 5,000 expressions extracted from dictionaries.(Al-Badrashiny et al., 2016) employed a paradigm detection method on the Arabic Treebank and Arabic Gigawords corpus, resulting in the autonomous extraction of 1,884 AMWEs, each displaying various forms due to morphological variations. Recently, as part of the PARSEME framework (Savary et al., 2023), Hadj Mohamed et al. (2022) manually constructed a corpus comprising 4,700 instances of Verbal AMWEs.

4. Method

Our ultimate goal is to address the task of identifying VMWEs in Arabic. However, within this paper, we specifically concentrate on the critical challenge of detecting unseen instances, which represents a significant frontier in the field. Our approach relies on a lexicon and minimizes noise by filtering out literal interpretations. In contrast to numerous existing methods for VMWE identification, we choose not to rely on a VMWE-annotated corpus, opting instead for a carefully curated VMWE list. This decision stems from the limited representation of MWEs with literal and figurative meanings in resources such as Arabic Wiktionary, leading us to manually extract VMWEs from an exhaustive paper dictionary. Given this VMWE lexicon, our methodology unfolds in two phases: the first is the identification of VMWE candidates, while the second involves the disambiguation of these candidate occurrences, as outlined by Algorithm (1). We start by aligning the VMWE lexicon with the test corpus to identify potential VMWE candidates within the text. This process involves comparing the lexicon entries with the content of the

[3]Agglutination is the process, common in Arabic, of adjoining clitics from simple word forms to create more complex forms.

test corpus in order to detect instances where VMWEs may occur. Then, we apply a binary PIE disambiguation method to distinguish between idiomatic and literal instances among these candidates. VMWEs are identified from idiomatic occurrences, while literal instances are retained for further analysis as supplementary data.

The following sections provide more detailed descriptions of these two phases.

Algorithm 1 : Procedure for extracting and filtering sentences containing MWEs from the corpus

```
procedure ExtractAndFilter(C, L, model)
    literal ← []
    idiomatic ← []
    for mwe ∈ L do
        for sentence ∈ C do
            if mwe occurs in sentence then
                class ← PIEC⁴(mwe, sentence)
                if classification is "literal" then
                    literal.append(sentence)
                else
                    idiomatic.append(sentence)
                end if
            end if
        end for
    end for
    return literal, idiomatic
end procedure
```

4.1. Identifying VMWE candidates

During this phase, VMWE candidates are identified based on the lemmas associated with each MWE in the lexicon. The use of multisets allows for the identification of candidates in any order, regardless of the syntactic dependency between them. For example, consider the first VMWE seen in the lexicon (**L**) in Figure 1: وضع يده (ūḍᶜ īdh | lit. ‘ put hand+his’) ‘put one’s hand’.

In sentences **(1)** and **(2)** from the *parseme-ar* corpus, the three lemmas ”وضع” (*’to put’*), ”يد” (*’hand’*), and ”ه” (*’his’*) are present, resulting in their extraction as VMWE candidates. However, sentence **(2)** contains no VMWEs but rather a coincidental occurrence. In contrast, the candidate identified from sentence **(4)** represents a literal occurrence for the third VMWE طار غرابه (tar ġurab-h | lit. ‘his crow flew off’) ‘to get old’” in **L**. The choice of using a forward step of filtering is a matter of balance between precision and recall. The expected noise present in the identification phase results in good recall (R= 0.79) but low precision (P=0.41). Addressing this challenge, the second filtering phase (4.2) aims to enhance precision. We achieve this through the implementation of subtask (A) of the SemEval shared task (Madabushi et al., 2022).

4.2. Disambiguating candidate VMWE occurrences

As previously stated, we proceed with our filtering phase by employing the same subtask (A) from the SemEval shared task. The aim here is to distinguish between the compositional (literal) and non-compositional (idiomatic) uses of PIE within a given context. This is different from the task of MWE extraction, which focuses on identifying MWEs within a corpus. Namely, our method takes a set of sentences containing a target PIE as input. We handle the disambiguation of PIEs in a manner similar to word sense disambiguation. Our fundamental assumption is that the context in which PIEs are used literally and figuratively differs significantly enough to justify distinct contextual representations. Figure 2 outlines an overview of the architecture, which is built upon the contextual language model used in our experiments, namely BERT.

Firstly, we aim to leverage the semantic idiosyncrasy characteristic of idiomatic expressions, highlighting that the meanings of the components within idiomatic expressions are related to the context in which they appear. To achieve this, we start by tokenizing the input, which consists of the sequence S and the target PIE. Following this, contextualized embeddings are generated using BERT and produce a vector representation for both the expression (PIE) and its context (S). Then, we add a Bidirectional LSTM (BiLSTM) layer for each embedding sequence to extract initial features from the raw embeddings. This results in in $h^{(S)} = \text{BiLSTM}(e^{(S)})$ and $h^{(PIE)} = \text{BiLSTM}(e^{(PIE)})$.

The attention flow layer integrates and combines information from both the context word sequence and the query word sequence (Seo et al., 2017). This process generates query-aware vector representations of the context words and propagates the word embeddings from the preceding layer. Similarly, in our specific task, the attention flow layer merges details from two embedding sequences that encode diverse types of information. We fused $h^{(S)}$ and $h^{(PIE)}$ into an attention layer to obtain an enhanced contextualized representations for both the sentence and the PIE. This results in a unified representation that integrates information from both the entire sentence and the PIE. Finally, we introduce a MaxPooling layer to reduce spatial dimensions in neural network architectures while preserving the most important features by selecting the maximum value from each feature map. Following this, the fused representation is passed through a series of Dense layers for classification.

The final output is produced by a sigmoid-

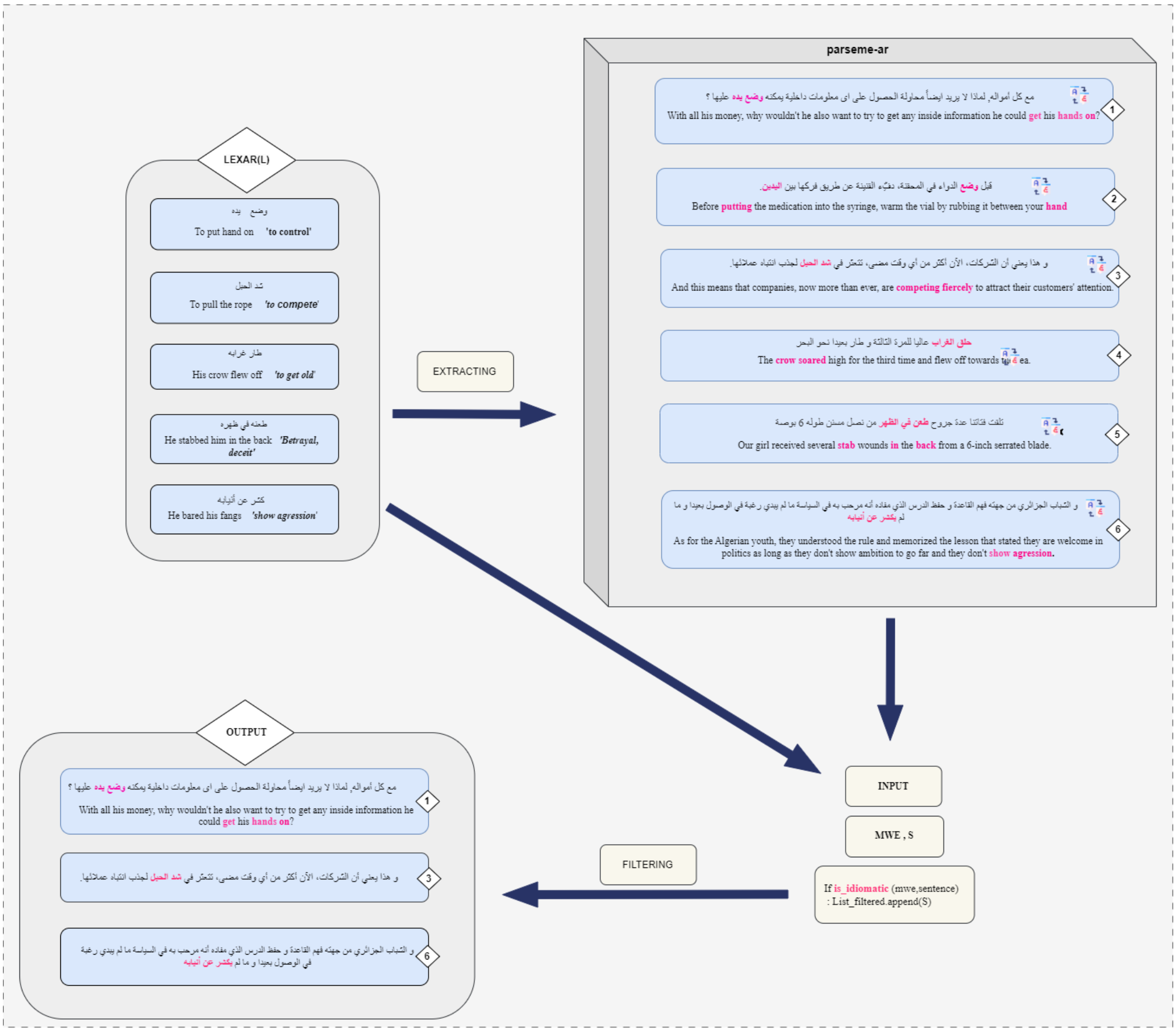

Figure 1: Overview of the method.

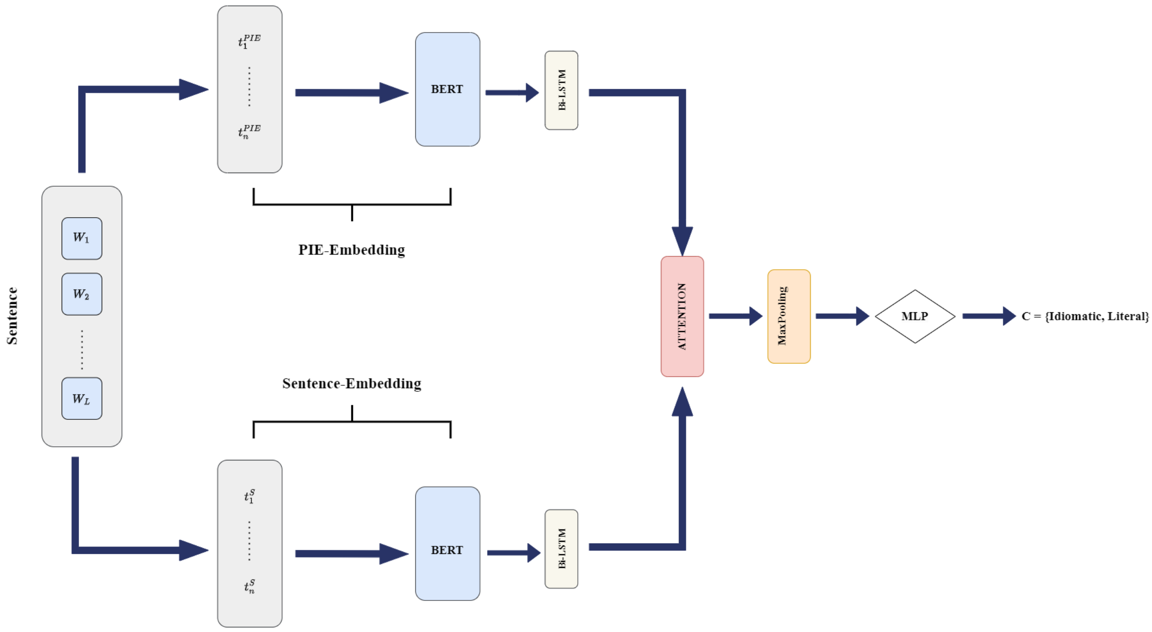

Figure 2: Overview of the PIEC model

activated Dense layer, providing a binary classification result (idiomatic or literal). Table 1 shows the hyper-parameters use with this architecture.

Parameter	Value
Sequence Length	128
Training Batch Size	256
Epoch number	30
Learning Rate	0.00001
Optimizer	Adam

Table 1: Model Training Parameters

5. Data

Assessing the efficacy of our MWE identification method necessitates both a VMWE lexicon and a corpus. As for the corpus, we used the "parseme-ar" corpus from PARSEME 1.3 (Hadj Mohamed et al., 2022; Savary et al., 2023), which contains 4,7000 VMWEs within 7,500 sentences extracted from PADT belonging to the UD collection (Hajic et al., 2009). In our experiments, our focus was on two categories of VMWEs outlined in the parseme-ar corpus: LVC (Light Verb Construction) and VID (Verbal Idiom). We excluded the IAV (In Inherently Adpositional Verb) category, as it is considered optional. Following this, we manually created a lexicon named LEXAR[5], referenced as (**L**) in Figure 1. We meticulously extracted and compiled idiomatic expressions from "Contextual Dictionary of Idiomatic Expressions" by Elsini (1998). Following the PARSEME annotation guidelines[6], we identified a total of 1504 Arabic VMWEs, and each expression in LEXAR underwent categorization by assigning a part-of-speech (POS) tag and determining its type as either LVC or VID. The annotation process, which took between 1-2 days and overlapped almost 70% of VMWEs with PARSEME-AR, ensured a comprehensive coverage of VMWEs in our corpus. We evaluated the performance of our idiomatic expression classifier, *PIEC*, by conducting evaluations with specialized datasets tailored to measure its accuracy in classifying sentences with idiomatic expressions. These evaluations encompassed datasets in Arabic, German, and English languages. Table 2 provides a summary of the data used to evaluate the secondary task. For Arabic, we trained the *PIEC* on a dataset included 34 idiomatic expressions. Each expression accompanied by sentences from the corpus of the shared task ConLL[7] encompassing both idiomatic and literal meanings. The 34 expressions were crafted manually by two native Arabic speakers. For instances lacking literal examples, we used ChatGPT to generate them, followed by manual verification. The MAGPIE corpus (Haagsma et al., 2020) provided the English dataset. It offers a collection of 1,756 PIEs, each representing different syntactic patterns, along with their associated sentences, totaling 56,622 annotated data instances with an average of 32.24 instances per PIE. For German we used the COLF-VID dataset (COrpus of Literal and Figurative meanings of Verbal IDioms) (Ehren et al., 2020). It contains 6,985 sentences sourced from newspaper articles, with annotations for 34 German VID types. Each MWE in the dataset is tagged with one of four labels: IDIOMATIC, LITERAL, UNDECIDABLE, or BOTH.

6. Results

The main goal of this study is to identify VMWEs, with a particular emphasis on unseen instances. Accordingly, we employed evaluation metrics aligned with the criteria of the shared task (Savary et al., 2017): These metrics include **MWE-based** metrics, which encompass precision, recall, and F1 scores for accurately detecting entire VMWEs, as well as precision, recall, and F1 measures for all VMWEs, including those that are unseen (**unseen MWE-based**). In Table 3, we compare the performance of our approach against MTLB-STRUCT.

On the multilingual level, MTLB-STRUCT achieved an MWE-based F1 score of 34.24 on unseen VMWEs and a global MWE-based F1 score of 56.27. Note that these results were obtained by re-training MTLB-STRUCT on the parseme-ar without the IAV category. However, even with the improvement in scores generated by the AraBert-based model (F1= 0.62 on the dev), Arabic is still one of the languages with the lowest performance score for global MWE-based and unseen-based scores. Although the F1 scores for unseen MWEs are still not optimal, our approach outperforms MTLB-STRUCT in terms of MWE-based F1 score by 7% and for unseen MWEs by 9%. Among the 278 unseen VMWEs assessed, our approach detected 125, whereas MTLB-STRUCT identified 104 out of the total.

For our experiments on the **binary disambiguation of PIEs** task (Figure 2), we focused only on the IDIOMATIC and LITERAL labels. Table 4 presents the results of our experiments on the TEST set. As baseline, we used a conventional SVM (Support Vector Machine) with MUSE (Multilingual Unsupervised and Supervised Embeddings) (Conneau et al., 2018) features. Em-

[5] We plan to release the lexicon upon acceptance of this paper

[6] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2

[7] https://lindat.mff.cuni.cz/

Lang	Literal	Figurative	Total
AR-train	103	202	305
AR-dev	16	30	46
AR-test	29	57	86
COLF-VID-train	1,172	5,705	6,902
COLF-VID-dev	264	1,214	1,488
COLF-VID-test	265	1,238	1,511
MAGPIE-train	2,676	12,676	15,352
MAGPIE-dev	595	2719	3314
MAGPIE-test	635	3339	3974

Table 2: Literal and idiomatic occurrences of PIEs in Arabic (AR), German (DE) (we excluded both the types of BOTH and UNDECIDABLE, which accounts for the disparity in the count between literal and idiomatic expressions compared to the total) and English(EN)

beddings were independently generated for both the PIE instances and sentences using the MUSE library. Notably, PIEC demonstrates better performance compared to the baseline MUSE-SVM. Including semantic information regarding both the context and the PIE significantly enhances the classifier's performance. It performs highly better on both literal and figurative class across all languages, even when dealing with unbalanced data in German and English. For instance, in the literal class, the F-score exhibited significant improvements: in Arabic from 0.44% to 0.89%, in English from 0.39% to 0.86%, and in German from 0.54% to 0.78%. Hence, the consistency of the PIEC classifier's performance with BERT embeddings implies that accurate disambiguation of PIEs across numerous languages can be achieved with good precision, necessitating only a small set of annotated sentences.

7. Conclusion

This paper introduces a simple yet impactful strategy for improving the identification of VMWE through the integration of lexicons, with our lexicon named LEXAR. Specifically focusing on the Arabic language, we demonstrate that our approach outperformed neural architectures like MTLB-STRUCT. Additionally, our method effectively adresses the challenge of binary disambiguation by employing contextual embeddings, which differentiate between various uses of the same lexical units and assign appropriate representations. Although detecting unseen MWEs proves to be a challenging task in our experiments, we achieve promising results using lexicons, surpassing the previous state-of-the-art. Moreover, our proposed model for the **binary disambiguation of PIEs** task shows significant potential for extension to multiple languages, facilitated by multilingual contextual embeddings.

Acknowledgement

We would like to express our sincere gratitude to Rafael Ehren and Laura Kallmeyer for graciously accepting me (Najet Hadj Mohamed) to undertake a short-term scientific mission at Heinrich Heine University Düsseldorf. We especially thank Rafael Ehren for providing the preprocessed English data, which significantly contributed to the completion of this research. Additionally, we extend our appreciation to UniDive, the CA21167 COST Action[8]: Universality, diversity and idiosyncrasy in language technology for their support and funding, which facilitated this study.

8. Bibliographical References

Mohamed Al-Badrashiny, Abdelati Hawwari, Mahmoud Ghoneim, and Mona Diab. 2016. SAMER: a semi-automatically created lexical resource for Arabic verbal multiword expressions tokens paradigm and their morphosyntactic features. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 113–122.

Hazem Al Saied, Marie Candito, and Mathieu Constant. 2019. Comparing linear and neural models for competitive mwe identification. In *The 22nd Nordic Conference on Computational Linguistics*.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef Van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 19–27.

Mohammed A. Attia. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In

[8]https://www.cost.eu/actions/CA21167/

Lang	Our approach						MTLB-STRUCT					
	MWE-based			*unseen MWE-based*			*MWE-based*			*unseen MWE-based*		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Arabic	64.87	61.91	63.36	44.88	41.67	43.21	55.07	57.35	56.27	37.77	31.47	34.24

Table 3: Comparing our approach performance with MTLB-STRUCT on MWE-based and unseen MWE-based metrics.

Lang	SVM-MUSE						PIEC					
	Literal			*Figurative*			*Literal*			*Figurative*		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Arabic	0.40	0.50	0.44	0.82	0.75	0.78	0.90	0.88	0.89	0.96	0.96	0.96
English	0.81	0.26	0.39	0.84	0.98	0.91	0.92	0.82	0.86	0.96	0.98	0.97
German	0.79	0.41	0.54	0.89	0.98	0.93	0.80	0.77	0.78	0.95	0.96	0.95

Table 4: Comparing SVM-MUSE and PIEC performance across 3 languages in term of Precision (P), Recall (R), and F-measure (F1).

International Conference on Natural Language Processing (in Finland), pages 87–98. Springer.

Bartosz Broda, Maciej Piasecki, and Stanislaw Szpakowicz. 2008. Sense-based clustering of polish nouns in the extraction of semantic relatedness. In *2008 International Multiconference on Computer Science and Information Technology*, pages 83–89. IEEE.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of german verbal idioms with a bilstm architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220.

Mahmoud Ismail Elsini. 1998. *Contextual dictionary of idiomatic expressions*. Lebanon Library Publishers.

Stefan Evert. 2005. *The statistics of word cooccurrences*. Ph.D. thesis, Dissertation, Stuttgart University.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskandar Keskes, Jean-Yves Antoine, and Lamia Belguith Hadrich. 2022. Annotating Verbal Multiword Expressions in Arabic: Assessing the Validity of a Multilingual Annotation Procedure. In *13th Conference on Language Resources and Evaluation (LREC 2022)*, Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1839–1848, Marseille, France.

Jan Hajic, Otakar Smrz, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2009. Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, volume 1.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80.

Abdelati Hawwari, Mohammed Attia, and Mona Diab. 2014. A framework for the classification and annotation of multiword expressions in dialectal Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 48–56, Doha, Qatar. Association for Computational Linguistics.

Kamil Kanclerz and Maciej Piasecki. 2022. Deep neural representations for multiword expressions detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 444–453, Dublin, Ireland. Association for Computational Linguistics.

Murathan Kurfali. 2020. Travis at parseme shared task 2020: How good is (m) bert at seeing the unseen? In *International Conference on Computational Linguistics (COLING), Barcelona, Spain (Online), December 13, 2020*, pages 136–141.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. Semeval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. *arXiv preprint arXiv:2204.10050*.

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.

Najet Hadj Mohamed, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. Alphamwe-arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020a. Seen2unseen at parseme shared task 2020: All roads do not lead to unseen verb-noun vmwes. In *Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LZX 2020)*.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020b. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44:137–158.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al. 2018. Parseme multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Agata Savary, Silvio Ricardo Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, et al. 2023. Parseme corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

Agata Savary, Carlos Ramisch, Silvio Ricardo Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemi Zadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al.

2017. The parseme shared task on automatic identification of verbal multiword expressions. In *The 13th Workshop on Multiword Expression at EACL*, pages 31–47.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension iclr. *arXiv preprint arXiv:1611.01603*.

Violeta Seretan et al. 2011. *Syntax-based collocation extraction*, volume 44. Springer Dordrecht.

Ekaterina Shutova. 2010. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. Mtlb-struct@ parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. *arXiv preprint arXiv:2011.02541*.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. Ermi at parseme shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135.