ID10M: Idiom Identification in 10 Languages

Simone Tedeschi[1,2], **Federico Martelli**[2] and **Roberto Navigli**[2]
[1]Babelscape, Italy
[2]Sapienza NLP Group, Sapienza University of Rome
tedeschi@babelscape.com, martelli@di.uniroma1.it,
navigli@diag.uniroma1.it

Abstract

Idioms are phrases which present a figurative meaning that cannot be (completely) derived by looking at the meaning of their individual components. Identifying and understanding idioms in context is a crucial goal and a key challenge in a wide range of Natural Language Understanding tasks. Although efforts have been undertaken in this direction, the automatic identification and understanding of idioms is still a largely under-investigated area, especially when operating in a multilingual scenario. In this paper, we address such limitations and put forward several new contributions: we propose a novel multilingual Transformer-based system for the identification of idioms; we produce a high-quality automatically-created training dataset in 10 languages, along with a novel manually-curated evaluation benchmark; finally, we carry out a thorough performance analysis and release our evaluation suite at https://github.com/Babelscape/ID10M.

1 Introduction

Idioms pertain to a wider family of linguistic phenomena referred to as multi-word expressions (MWEs). Broadly speaking, an MWE can be defined as a combination of two or more words, behaving as a complex lexical unit and showing idiosyncratic properties (Baldwin and Kim, 2010). Over the course of the last few years, several attempts have been made to classify MWEs based on specific dimensions such as polylexicality, fixedness, compositionality and idiomaticity (Sailer and Markantonatou, 2018). According to Sag et al. (2002), MWEs can be divided into lexicalized and institutionalized phrases. While the former show syntactic or semantic idiosyncrasies, e.g. *kingdom come* and *spill the beans*, the latter are compositional from a syntactic and semantic perspective, but statistically idiosyncratic, e.g. *traffic light* and *telephone booth*.

Among lexicalized phrases, idioms are of particular interest in that their meaning cannot be obtained by compositionally interpreting their word constituents. These include non-compositional phrases, e.g. *kick the bucket*, and partially-compositional phrases, e.g. *rain cats and dogs* (Nunberg et al., 1994).

Given their complex nature, idioms are hard to be automatically identified and pose a crucial challenge to the entire field of Natural Language Understanding (NLU). Although research in this field has recently achieved great advancements, the current formulation of many tasks tends to overlook the idiomatic usage of language. Instead, idioms ought to be playing an important role in NLU as they are a frequent phenomenon which can be observed in all languages. The correct identification of idioms in context is crucial for tasks such as Word Sense Disambiguation (Bevilacqua et al., 2021) and Entity Linking (Sevgili et al., 2020), but also for many downstream applications. For instance, in Question Answering or dialog, a system must be able to understand *"It was a piece of cake"* in relation to the question *"How was the test?"* (Jhamtani et al., 2021; Mishra and Jain, 2016). Similarly, if the idiom *kick the bucket* is identified, then a Text Summarization system would be able to summarize all its occurrences within a text with "die" (Chu and Wang, 2018; Gambhir and Gupta, 2017). Finally, once an idiom is identified, a Machine Translation system would then be able to avoid its compositional translation, and treat it as a whole (Anastasiou, 2010). Furthermore, idioms are widely studied in linguistics and psycholinguistics (Cacciari and Tabossi, 1988; Gibbs Jr, 1992; Nunberg et al., 1994; Cacciari and Tabossi, 2014; Liu, 2017), hence a system capable of effectively identifying idioms in texts would significantly improve many research areas, far beyond NLU.

Most of the past idiom extraction strategies focused on specific domains and on a limited number

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2715 - 2726

of languages. In our work, we tackle these shortcomings and, taking inspiration from the Named Entity Recognition (NER) task (Yadav and Bethard, 2018), we reformulate the identification of idioms as a sequence labeling task. Specifically, we propose the following new contributions:

- We design a novel multilingual Transformer-based system for the identification of idioms;
- We release a high-quality silver training dataset in 10 languages and a novel manually-curated evaluation benchmark in 4 languages;
- We measure the quality of the data produced and of our system design through an extensive evaluation.

We hope that this work will provide a stepping stone for further studies regarding idiomatic expressions and their applications, and encourage further work on the identification of idioms in multiple languages. We release the produced datasets and software at `https://github.com/Babelscape/ID10M`.

2 Related Work

Systems Over the course of the past two decades, several approaches have been put forward to address the idiom identification task. To this end, two main properties of idioms have been leveraged, namely their syntactic and semantic idiosyncrasies. While the former refers to the peculiar syntactic behaviour of idioms, the latter indicates the linguistic property in which the meaning of an idiomatic expression cannot be completely derived from the meaning of its individual components.

Initial studies regarding idiom identification focused on syntactic idiosyncrasy, concentrating on verb/noun idioms, e.g. *shoot the breeze* (Fazly and Stevenson, 2006; Cook et al., 2007; Diab and Bhutada, 2009), on verb/particle idioms, e.g. *call off* (Ramisch et al., 2008) or on idioms satisfying specific restrictions, i.e. subject/verb such as *tension mounted* and verb/direct-object, e.g. *break the ice* (Shutova et al., 2010).

Subsequent approaches exploited semantic idiosyncrasies. This property implies that idiomatic expressions often occur in contexts typically unrelated to the meaning of their individual constituents, thus providing a key feature to be exploited in an automatic approach. In particular, Muzny and Zettlemoyer (2013) introduced new lexical and graph-based features that use WordNet[1] and Wiktionary[2], and proposed a simple yet efficient binary Perceptron classifier to distinguish between idiomatic and non-idiomatic expressions by exploiting their components and dictionary definitions. A similar, but unsupervised approach was adopted by Verma and Vuppuluri (2015) which relied on the dictionary definitions of each component of a given idiom.

These latter methods have more recently been superseded by approaches making use of distributional similarity in the form of both static and contextualized word embeddings (Gharbieh et al., 2016; Ehren, 2017; Senaldi et al., 2019; Hashempour and Villavicencio, 2020; Fakharian, 2021; Garcia et al., 2021; Nedumpozhimana and Kelleher, 2021), while keeping the underlying assumption unchanged: the vector representation of the component words should be distant from the vector representation of the context or of the expression as a whole.

Notwithstanding the recent improvements, to the best of our knowledge, the identification of idiomatic expressions in multiple languages is largely under-investigated.

Datasets In the early 2000s, several datasets for idiom identification were created. For instance, Cook et al. (2008) and Sporleder et al. (2010) manually selected a limited number of idioms, and then extracted sentences containing such idioms from the British National Corpus (BNC, Consortium et al., 2007). Similarly, Sporleder and Li (2009) extracted a dataset from the Gigaword corpus (Graff and Christopher, 2003). Street et al. (2010), instead, used multiple annotators to validate sentences from the American National Corpus (ANC, Ide and Macleod, 2001). Additionally, Muzny and Zettlemoyer (2013) created a dataset by applying the aforementioned classifier on Wiktionary entries, more than doubling the number of idiomatic expressions in Wiktionary.

Furthermore, Korkontzelos et al. (2013) introduced Task 5b at SemEval-2013 regarding the detection of semantic compositionality in context. The authors selected idioms from Wiktionary, and extracted instances from the ukWaC corpus (Ferraresi et al., 2008). Schneider et al. (2016), instead, proposed the DiMSUM dataset for Task 10 at SemEval-2016, and extracted annotations from reviews, tweets and TED talks. However, this work

[1] https://wordnet.princeton.edu/
[2] https://www.wiktionary.org/

did not categorise MWEs into subtypes, making it difficult to quantify the number of idioms in the corpus.

Finally, Peng et al. (2015) expanded the dataset introduced by Cook et al. (2008) by retrieving further sentences from the BNC corpus, while more recently Gong et al. (2017) introduced a small-scale dataset derived from Google Books[3] for English and Chinese.

Unfortunately, almost all the aforementioned approaches focused on English. The first concrete attempt to scale to multiple languages was made by Madabushi et al. (2021) who also proposed a SemEval-2022 task on idiom identification. Nevertheless, their datasets are limited in size and they only cover three languages, namely English, Portuguese and Galician.

3 ID10M

In what follows, we first describe the creation process of our training datasets (Section 3.1) and the manually-curated test sets (Section 3.2). Then, we introduce our new task formulation and illustrate the architecture of our idiom identification system (Section 3.3).

3.1 Silver-Standard Data Creation

Automatic Annotation In order to create our training data, we exploit Wiktionary[4] as the main source, as it provides access to a large number of MWEs along with usage examples in multiple languages. However, since such examples are provided for a limited number of MWEs, we search for further textual contexts in a large external source, namely WikiMatrix[5] (Schwenk et al., 2021), a multilingual corpus that covers 83 languages and contains parallel sentences extracted from Wikipedia[6].

We perform data extraction as follows. Let E_l be the set of MWEs available in Wiktionary in the language l, with $|E_l| = n$, and let us define the function $L(p)$ that, given a phrase p, outputs its lemma. Then, we apply a heuristic which allows us, for each expression $e_i \in E_l$, to search for a sentence in WikiMatrix such that there exists at least a span of tokens S_{k-j} starting at index k and ending at index j, where $e_i = S_{k-j} \vee e_i = L(S_{k-j}) \vee L(e_i) = S_{k-j} \vee L(e_i) = L(S_{k-j})$. By applying this heuristic, not only do we obtain a large set of sentences containing potentially idiomatic expressions (PIEs), but – thanks to the lemmatization step – we also collect several morphological variations of the original expressions in E_l, e.g. starting from '*kick the bucket*', we also obtain '*kick<u>ed</u> the bucket*' and '*kick<u>s</u> the bucket*'. In particular, if an MWE is marked as idiomatic in Wiktionary, we mark all its occurrences as idiomatic too. Similarly, if an MWE is not marked as idiomatic in Wiktionary, we mark all its occurrences as literal. However, this has a limitation: if an MWE is labeled as idiomatic (or literal) in Wiktionary, it will not necessarily always also be idiomatic (or literal) in the WikiMatrix sentences in which it appears.

We adopt the above-described procedure to create datasets in the following 10 languages: Chinese, Dutch, English, French, German, Italian, Japanese, Polish, Portuguese and Spanish.

Automatic Validation Since the data derived from Wiktionary and WikiMatrix may contain errors, we aim at automatically improving their quality. To achieve this goal, we exploit the semantic idiosyncrasy property of idiomatic expressions, and the consequent fact that the meaning of the individual constituents of idiomatic expressions are unrelated to the surrounding context. Specifically, following this intuition, and by taking inspiration from recent advances in the main disambiguation tasks (Blevins and Zettlemoyer, 2020; Botha et al., 2020; Tedeschi et al., 2021), we design a dual-encoder architecture (Figure 1) to produce a vector representation for both the expression and its context, and then, based on their cosine similarity, label the expression as idiomatic or literal.

More formally, let us define an expression encoder Ψ and a context encoder Ω. Then, given an expression-context pair $\langle e, c\rangle$, the output of the dual-encoder architecture Φ is defined as follows:

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

where $\Phi(e,c) = 1$ means that e is idiomatic in c, while $\Phi(e,c) = 0$ if e has a literal meaning

[3] https://books.google.com/

[4] We employ the Wiktextract library to collect the necessary data from Wiktionary. WikiExtract (https://pypi.org/project/wiktextract/) provides a preprocessed version of the Wiktionary dump together with useful APIs.

[5] https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix

[6] The encyclopedia-style prose of Wikipedia could have a lower idiom density compared to other textual sources, but the large size of WikiMatrix should balance this lower density.

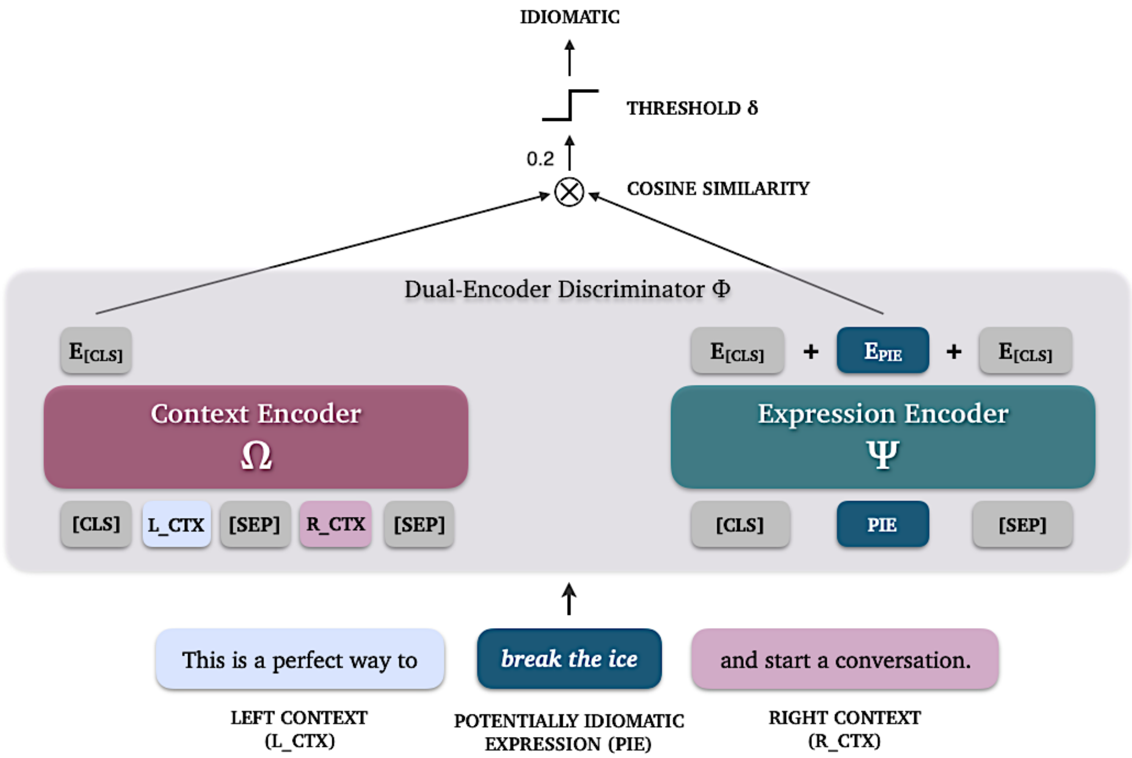

Figure 1: Graphical representation of the dual-encoder architecture given as input an example sentence. “E" stands for Embedding. A potentially idiomatic expression e is labeled as idiomatic when the cosine similarity score between the representations $\Omega(c)$ and $\Psi(e)$, where c is the surrounding context, is lower than the threshold δ.

in c. δ is a manually-tuned threshold. Both encoders are `bert-base-multilingual-cased` architectures that take as input the tokenized versions of expressions and their contexts, respectively, surrounded by the special tokens [CLS] and [SEP]. To encode an expression, we take the sum of the individual representations of all its subwords. Instead, for the representation of the context we take the representation of the [CLS] token. We evaluate the quality of our dual encoder in Section 4.3.

Additionally, to further improve the quality of the annotations produced, we follow the recent findings of Tedeschi and Navigli (2022) which demonstrated how NER can be exploited to better discriminate between idiomatic and literal usages of potentially idiomatic expressions.

3.2 Gold-Standard Data Creation

To evaluate the performance of our idiom identification system, we manually create a novel evaluation benchmark in 4 languages, i.e. English, German, Italian and Spanish. As explained in Section 3.1, we start by producing a set of sentences containing PIEs. Then, to properly label the expressions, depending on the context in which they occur, we ask professional annotators[7] to perform the following binary classification task: given a context-expression pair $\langle e, c\rangle$, the goal is to tag this pair with a label $y \in \{Idiomatic, Literal\}$. In order to make our gold standard more challenging, and better evaluate the system performance, we also ask annotators to include unseen idioms, i.e. idioms that do not appear in the training set.

3.3 Idiom Identification

Task Formulation Current and past approaches to idiom identification typically take expressions-context pairs $\langle e, c\rangle$ as input and limit themselves to determining whether e is used with a figurative meaning or not in c (Madabushi et al., 2021; Muzny and Zettlemoyer, 2013). However, this formulation has a major drawback: potentially idiomatic expressions need to be pre-identified. Importantly, we drop this requirement and reformulate the task as a sequence-labeling task, by employing the well-known BIO tagging scheme[8].

[7] We hired a mother-tongue professional annotator for each language.

[8] The BIO tagging scheme (short for Beginning, Intermediate, Out) is a popular tagging scheme where the B label indicates that the corresponding token is the first token of a

	Language	# Sentences	# Tokens	# Idioms	# B	# I	# O	# Seen	# Unseen	# Literal
Silver Data	Chinese (ZH)	9543	244422	1301	5272	3823	235327	-	-	3918
	Dutch (NL)	20935	548872	189	4530	10543	533799	-	-	16366
	English (EN)	37919	1199492	4568	10102	19884	1169506	-	-	27408
	French (FR)	35588	939161	188	12112	25248	901801	-	-	23238
	German (DE)	26963	722109	819	8311	11500	702298	-	-	18488
	Italian (IT)	29523	813445	452	8768	12353	792324	-	-	20506
	Japanese (JA)	6388	211437	165	2534	1662	207241	-	-	3852
	Polish (PL)	36333	862265	648	12971	14364	834930	-	-	22467
	Portuguese (PT)	30942	764017	559	5824	8871	749322	-	-	24816
	Spanish (ES)	28647	648776	1229	9994	13927	624855	-	-	17851
Gold Data	English (EN)	200	3287	142	159	373	2755	62	80	41
	German (DE)	200	4529	111	181	377	3971	71	40	19
	Italian (IT)	200	5043	139	155	271	4617	87	52	48
	Spanish (ES)	200	2240	78	133	348	1759	19	59	66

Table 1: Statistics concerning the automatically-created (Silver Data) training sets and our manually-curated test sets (Gold Data). "# Seen" represents the number of expressions in the test set already encountered in the training set, whereas "# Unseen" is the number of expressions never encountered. In the count of individual idioms (# Idioms), morphological variations of a certain idiom are mapped to the same idiom.

More formally, given as input a raw text sequence X of n tokens $x_1, \ldots, x_n$, each x_i must be labeled by the system with a tag $y_i \in \{B, I, O\}$ for each $i \in [1, n]$. This formulation also allows us to easily handle multiple idiomatic expressions within the same text.

In order to use our new formulation, we convert all the datasets constructed in Section 3.1 and Section 3.2 in BIO format. Table 2 shows an example of instance labeled using the BIO tagging scheme.

Token	Label
After	O
some	O
reflection	O
,	O
he	O
decided	O
to	O
bite	B-IDIOM
the	I-IDIOM
bullet	I-IDIOM
.	O

Table 2: Example of instance labeled according to the BIO tagging scheme.

Our System Our model for idiom identification is inspired by the BERT-based neural architecture of Mueller et al. (2020) used for Named Entity Recognition, however, rather than encoding a word with the first contextualized subword representation as indicated by Devlin et al. (2019), we take the mean of its subwords, as suggested by recent literature (Ács et al., 2021). Then, the resulting vectors are passed through a multi-layer sentence-level BiLSTM network, whose logits are finally fed into a CRF model, trained to maximize the log-likelihood of the span-based gold label sequences (Huang et al., 2015).

4 Experiments

In this Section, we describe our experimental setup (Section 4.1), the datasets we use to train and evaluate our idiom identification system (Section 4.2), and the results obtained (Section 4.3).

4.1 Experimental Setup

We implement our idiom identification system (Section 3.3) and our dual-encoder discriminator (Section 3.1) with PyTorch (Paszke et al., 2019), using the Transformers library (Wolf et al., 2019) to load the weights of `BERT-base-multilingual-cased` (mBERT). We fine-tune our idiom identification system for 30 epochs with a Cross-Entropy loss criterion, adopting an early stopping strategy with a patience value of 5, Adam (Kingma and Ba, 2015) optimizer and a learning rate of 10^{-5}, as standard when fine-tuning the weights of a pretrained language model. For our dual-encoder discriminator, instead, we use mBERT as feature extractor since no training data for the task were available.

span, in this case an idiomatic expression, the I label denotes an intermediate token of a span, and O means out of a span.

Hyperparameter name	Value
number of Bi-LSTM layers	2
LSTM hidden size	256
gradient accumulation steps	4
batch size	32
learning rate	0.00001
dropout	0.5
gradient clipping	1.0
adam β_1	0.9
adam β_2	0.999
adam ϵ	1e-8

Table 3: Hyperparameter values of the reference idiom identification system used for our experiments.

Language	Accuracy
English	84.12
German	81.98
Italian	82.74
Spanish	82.55
Avg.	**82.85**

Table 4: Accuracy of the annotations produced by our automatic system compared to those provided by the human annotators on the 4 languages covered by our gold-standard test sets.

The entire model training is carried out on a NVIDIA GeForce RTX 3090. Each training (i.e. for each language) requires ~8min/epoch on average, for a mean of ~20 epochs. Table 3 shows the full list of hyperparameters.

4.2 Training, Validation and Test Data

The training and validation sets that we use in our experiments are those obtained by applying the methodology described in Section 3.1, with $\delta = 0.4$[9]. Although we automatically produce training data in 10 languages, we report results only on the 4 languages for which manually-curated test sets are available (see Section 3.2). However, since the training data has been created with the same procedure for each of the 10 languages, similar results are expected on non-tested languages. Statistics are provided in Table 1.

[9]We use the English validation set to manually search for the best value of δ by choosing from the following set of possible values: $\delta = \{0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7\}$

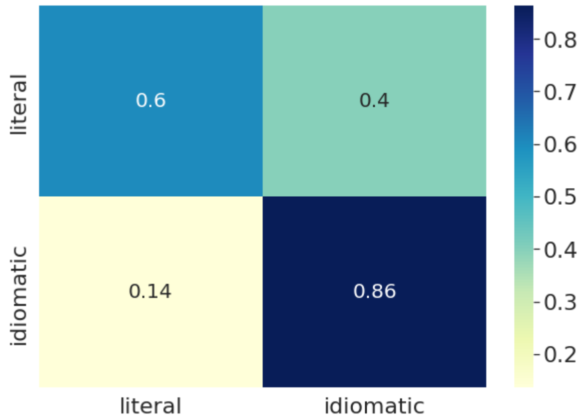

Figure 2: Confusion matrix of the predictions of our automatic system (X-axis) compared to the corresponding ground truth values (Y-axis). Results are averaged over the 4 languages covered by the test sets.

4.3 Results

In what follows, we measure both the quality of our automatic annotation methodology (Section 3.1) and of our idiom identification system (Section 3.3) by means of accuracy and token-level macro F_1-score metrics, respectively. In the latter case, we rely on the macro-F_1 metric due to the high class imbalance in the datasets, i.e. the number of O tags is much higher than the sum of the number of B and I tags, see Table 1.

Silver-Data Quality Evaluation We first aim at providing an empirical evaluation of the effectiveness of the proposed automatic strategy for producing idiom-related[10] sentences. To do so, for each language, we apply our dual-encoder discriminator $\Phi(e, c)$ to the expression-context pairs $\langle e, c \rangle$ available in our manually-curated test set, and we measure the accuracy score by comparing the predictions produced by the system with the human annotations in the gold-standard test sets. The accuracy results obtained are reported in Table 4.

With this being a binary-classification task, we can observe that the performance achieved by our dual encoder is much higher than the 50% baseline of a random classifier, hence implying that the system is able to distinguish between idiomatic and literal usages of PIEs based on the surrounding contexts.

However, the accuracy is not sufficient for us to determine the strengths and the weaknesses of our system. Therefore, we group both the predic-

[10]With the term "idiom-related sentences" we refer to sentences containing potentially idiomatic expressions.

	Tag	P	R	F1	% Seen
EN	B	84.2	53.5	65.4	-
	I	91.1	57.4	70.4	-
	O	92.5	99.1	95.7	-
	ALL	**89.2**	**70.0**	**77.1**	43.7%
DE	B	87.6	70.2	77.9	-
	I	90.7	72.7	80.7	-
	O	96.2	98.9	97.5	-
	ALL	**91.5**	**80.6**	**85.4**	64.0%
IT	B	72.2	61.9	66.7	-
	I	76.7	62.0	68.6	-
	O	96.7	98.2	97.4	-
	ALL	**81.8**	**74.0**	**77.6**	62.6%
ES	B	47.2	51.1	49.1	-
	I	67.4	45.1	54.0	-
	O	87.7	92.8	90.2	-
	ALL	**67.4**	**63.0**	**64.4**	24.4%

Table 5: Results of the idiom identification system in terms of Precision (P), Recall (R) and Macro-F1 (F1) scores on the four test languages. "% Seen" represents the percentage of idioms already encountered in the training set. Morphological variations of the same idiom are considered as a unique idiom.

	Seen			**Unseen**		
Language	P	R	F1	P	R	F1
EN	91.9	71.5	79.1	87.2	68.8	75.6
DE	98.7	95.5	97.1	64.5	49.1	53.8
IT	96.5	91.3	93.8	55.9	49.7	52.2
ES	94.9	96.9	95.9	60.2	55.6	56.9

Table 6: Results on the "Seen" and "Unseen" test set subsets in terms of token-level Precision (P), Recall (R) and Macro-F1 (F1) scores.

	Dual Encoder?	F1	Δ
EN	Yes	**77.1**	-
	No	73.6	+ 3.5
DE	Yes	**85.4**	-
	No	81.9	+ 3.5
IT	Yes	**77.6**	-
	No	73.4	+ 4.2
ES	Yes	**64.4**	-
	No	58.3	+ 6.1

Table 7: Comparison of the results obtained by training the system on the silver-standard data validated by our dual encoder (Yes) and non validated ones (No).

tions and the labels coming from the 4 languages, and construct a confusion matrix in order to better analyze the system behavior. From the confusion matrix in Figure 2, we can observe that the system is able to (almost always) identify idiomatic expressions as such, mainly thanks to their semantic distance from the meaning of the surrounding words. On the other hand, when dealing with literal expressions, the system again correctly predicts the majority of these, but it makes more errors. We attribute this to the fact that the context is often not sufficiently rich to find a strong similarity (i.e. higher than the threshold δ) with the meaning of the individual constituents of the idiomatic expression, and hence to label the expression as literal. Indeed, the lower the value of δ, the higher the number of literal expressions discovered, but the system inevitably classifies more idiomatic expressions as literal.

Multilingual Idiomatic Expression Identification In the previous paragraph we evaluated the performance of our dual-encoder architecture on the binary *literal or idiomatic* classification task, where the PIE was pre-identified. In this paragraph, instead, we use the refined silver-data produced by the aforementioned dual encoder, and measure the identification capabilities of our idiom identification system on the sequence-labeling task we introduced (Section 3.3) by comparing the BIO tags produced with the corresponding gold labels. The results obtained are reported in Table 5 (further results are provided in Appendix A).

The first thing that catches the eye is that the performances on the O tags are much higher than those on the B and I tags, on all tested languages. However, this is not surprising, owing to the fact that there is a high class imbalance. An interesting result, instead, is that the system achieves an average score of about 76 F1 points, while the percentage of seen entities[11] is only 48.7% on average. This implies that the system is able to generalize, and consequently also to correctly predict unseen idioms. This phenomenon is particularly evident on English and Spanish, where the percentage of seen idioms is very low.

To better highlight the capability of the system to go beyond idioms already seen during training, we also analyze the system performance on the "seen" and "unseen" subsets independently, and report the results in Table 6. As we can observe, the

[11] Seen entities are entities in the test set which have already been encountered in the training set.

	Type	Prediction
DE	Correct ✔	Ich bin nur der Typ, der ***ihr die Stange hält***.
	Correct ✔	Wir haben dieses Geschäft ***von Grund auf*** aufgebaut.
	Wrong ✘	***Durch den Wind*** wurden 27 Häuser in der Region zerstört.
	Wrong ✘	Sei nicht so'ne beleidigte Leberwurst!
EN	Correct ✔	The old horse finally ***kicked the bucket***.
	Correct ✔	Written tests are his ***Achilles' heel***...
	Wrong ✘	Her aunt is a great cook, do you want a ***piece of cake***?
	Wrong ✘	It is difficult, but possible to quit smoking cold turkey.
IT	Correct ✔	Mi sono ***cavato gli occhi*** dopo aver decifrato la grafia farraginosa.
	Correct ✔	Invece di ***decidere su due piedi***, diedi disposizioni a Tom Donilon perché convocasse i delegati...
	Wrong ✘	A quel punto Smith lanciò ***a terra*** un bicchiere.
	Wrong ✘	Non era affatto scontato che Romney rientrasse nei ranghi, visti i suoi rapporti burrascosi con Trump.
ES	Correct ✔	A la inaguración fueron ***cuatro gatos***.
	Correct ✔	El gobierno sigue ***metiendo el dedo en la llaga***.
	Wrong ✘	¿Has visto alguna vez a tu gato ***meter la pata*** en su bebedero?
	Wrong ✘	El agente tiene vista de lince.

Table 8: Examples of idioms correctly and wrongly identified by our idiom identification system. Underline represents the target idiomatic expression (if any), while bold + italic represents the predicted idiomatic expression.

system is able to correctly predict the majority of unseen idioms on all tested languages, achieving an F1 score of 59.6 points, on average. Moreover, on seen idioms, the system behaves almost perfectly reaching an average score of 91.5 points. We underline that morphological variations of idioms encountered in the training sets are considered as seen idioms. Table 1 provides dimensions of the "seen" and "unseen" subsets, for each language.

Then, to further demonstrate the effectiveness of our dual-encoder architecture (Section 3.1), we compare the results obtained by training the system on the data produced with and without the validation performed by our dual encoder. The results reported in Table 7 highlight an average gap of 4.3 F1-score points between the refined version of the data and the original one, showing how the validation step is fundamental for improving the quality of the annotations, consequently leading the system to a better understanding of idioms.

Finally, the high results in Table 5 also prove that, thanks to our renewed task formulation (Section 3.3), common sequence-labeling architectures (e.g. those used for NER) can be successfully imported into the idiom identification task, thus enabling knowledge transfer from other research areas.

5 Qualitative Analysis

Together with the quantitative evaluation provided in Section 4.3, we now perform a qualitative analysis of our system. More specifically, in Table 8, we provide 4 examples of system predictions (2 correct and 2 wrong) for each tested language. Although our system proves to be robust over literal PIEs (see Figure 2), its most common mistake consists in classifying a PIE used with its literal meaning as idiomatic. This is mainly due to the system bias towards the labels associated to such PIEs during training, e.g. if more than 90% of occurrences of a certain PIE are labeled as idiomatic in the training set, then the system will tend to classify as idiomatic any other of its occurrences in the test set. This result suggests that improvements over the distribution of labels of PIEs are possible. In Table 8 we provide an example of one such wrongly labeled PIE for each language. Another commonly observed error, again highlighted in Table 8, is that in which unseen idiomatic expressions are not identified by the system. However, as previously demonstrated in Table 6, the system is nevertheless able to correctly handle the majority of such cases.

On the other hand, we observe that the system is able to correctly identify both lemmatized and inflected forms of idiomatic expressions, for both seen and unseen ones.

6 Conclusions and Future work

In this work, we introduced ID10M, an innovative framework for idiom identification consisting of i) a new multilingual Transformer-based architecture, ii) a novel automatic annotation pipeline for creating high-quality silver-data in 10 languages, and iii) a challenging manually-curated benchmark in 4 languages. Moreover, while the majority of current approaches to idiom identification need pre-identified potentially idiomatic expressions, we, instead, dropped this requirement and proposed a new formulation for the idiom identification task that lets systems be directly applicable to raw texts. Finally, our experiments showed that our system is able to generalize beyond idioms seen during training, hence achieving up to 85.4 macro F1-score on the idiom identification task.

As future work, we plan to scale our system to a greater number of languages and textual sources, but, most importantly, investigate the benefits derived from our work in key tasks such as Word Sense Disambiguation, Machine Translation and Question Answering.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme and the support of the ELEXIS project No. 731015 under the European Union's Horizon 2020

References

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Dimitra Anastasiou. 2010. *Idiom treatment experiments in machine translation*. Cambridge Scholars Publishing.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of memory and language*, 27(6):668–683.

Cristina Cacciari and Patrizia Tabossi. 2014. *Idioms: Processing, structure, and interpretation*. Psychology Press.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

BNC Consortium et al. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. 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, Prague, Czech Republic. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

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

Rafael Ehren. 2017. Literal or idiomatic? identifying the reading of single occurrences of German multiword expressions using word embeddings. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–112, Valencia, Spain. Association for Computational Linguistics.

Samin Fakharian. 2021. *Contextualized embeddings encode knowledge of English verb-noun combination idiomaticity*. Ph.D. thesis, University of New Brunswick.

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

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 112–118, Berlin, Germany. Association for Computational Linguistics.

Raymond W Gibbs Jr. 1992. What do idioms really mean? *Journal of Memory and language*, 31(4):485–506.

Hongyu Gong, Suma Bhat, and Pramod Viswanath. 2017. Geometry of compositionality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

David Graff and Cieri Christopher. 2003. English gigaword ldc2003t05. web download. In *Philadelphia: Linguistic Data Consortium, 2003*.

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.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Nancy Ide and Catherine Macleod. 2001. The american national corpus: A standardized resource of american english. In *Proceedings of corpus linguistics*, volume 3, pages 1–7. Citeseer.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Dilin Liu. 2017. *Idioms: Description, comprehension, acquisition, and pedagogy*. Routledge.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. Astitchinlanguagemodels: Dataset and methods for the exploration of idiomaticity in pre-trained language models.

Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

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.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding bert's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jing Peng, Anna Feldman, and Hamza Jazmati. 2015. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the*

International Conference Recent Advances in Natural Language Processing, pages 507–511, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Carlos Ramisch, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 49–56, Manchester, England. Coling 2008 Organizing Committee.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Manfred Sailer and Stella Markantonatou. 2018. *Multiword Expressions: Insights from a multi-lingual perspective*. Language Science Press.

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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Marco Silvio Giuseppe Senaldi, Yuri Bizzoni, and Alessandro Lenci. 2019. What do neural networks actually learn, when they learn to identify idioms? *Proceedings of the Society for Computation in Linguistics*, 2(1):310–313.

Ozge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2020. Neural entity linking: A survey of models based on deep learning. *arXiv preprint arXiv:2006.00575*.

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

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, Athens, Greece. Association for Computational Linguistics.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Laura Street, Nathan Michalov, Rachel Silverstein, Michael Reynolds, Lurdes Ruela, Felicia Flowers, Angela Talucci, Priscilla Pereira, Gabriella Morgon, Samantha Siegel, Marci Barousse, Antequa Anderson, Tashom Carroll, and Anna Feldman. 2010. Like finding a needle in a haystack: Annotating the American national corpus for idiomatic expressions. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021. Named Entity Recognition for Entity Linking: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2584–2596, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. NER4ID at SemEval-2022 Task 2: Named Entity Recognition for Idiomaticity Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 681–687, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Additional Results

Since we reformulated the idiom identification task as a sequence-labeling task (Section 3.3), all previous approaches (that required pre-identified potentially idiomatic expressions) cannot be compared directly. Nonetheless, in order to select a

System	F1
Bi-LSTM	69.5
Bi-LSTM + CRF	70.9
mBERT	74.8
mBERT + Bi-LSTM	75.4
mBERT + Bi-LSTM + CRF	**76.1**
XLM-R	74.3
XLM-R + Bi-LSTM	75.4
XLM-R + Bi-LSTM + CRF	75.9

Table 9: Token-level macro F1 scores of different sequence tagging alternatives computed on our test set. Results are averaged over the four languages.

robust architecture for the idiom identification task, we compared various sequence tagging architectures. Specifically, we evaluated the performance of several alternative systems: Bidirectional LSTM (Bi-LSTM), Bi-LSTM + CRF, Multilingual BERT (mBERT), mBERT + Bi-LSTM, mBERT + Bi-LSTM + CRF, XLM-RoBERTa (XLM-R, Conneau et al., 2020), XLM-R + Bi-LSTM, XLM-R + Bi-LSTM + CRF. Results are reported in Table 9. Surprisingly, mBERT achieved performance slightly higher than XLM-R. Moreover, the addition of Bi-LSTM and CRF modules provided further improvements.