# SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding

**Harish Tayyar Madabushi**[1], **Edward Gow-Smith**[1],
**Marcos Garcia**[2], **Carolina Scarton**[1],
**Marco Idiart**[3] and **Aline Villavicencio**[1]

[1] University of Sheffield, UK
[2] Universidade de Santiago de Compostela, Spain
[3] Federal University of Rio Grande do Sul, Brazil

{h.tayyarmadabushi, egow-smith1, c.scarton, a.villavicencio} @sheffield.ac.uk
marcos.garcia.gonzalez@usc.gal, marco.idiart@gmail.com

## Abstract

This paper presents the shared task on *Multilingual Idiomaticity Detection and Sentence Embedding*, which consists of two Subtasks: (a) a binary classification task aimed at identifying whether a sentence contains an idiomatic expression, and (b) a task based on semantic text similarity which requires the model to adequately represent potentially idiomatic expressions in context. Each Subtask includes different settings regarding the amount of training data. Besides the task description, this paper introduces the datasets in English, Portuguese, and Galician and their annotation procedure, the evaluation metrics, and a summary of the participant systems and their results. The task had close to 100 registered participants organised into twenty five teams making over 650 and 150 submissions in the practice and evaluation phases respectively.

## 1 Introduction

Multiword Expressions (MWEs) are a challenge for natural language processing (NLP), as their linguistic behaviour (e.g., syntactic, semantic) differs from that of generic word combinations (Baldwin and Kim, 2010; Ramisch and Villavicencio, 2018). Moreover, MWEs are pervasive in all domains (Biber et al., 1999), and it has been estimated that their size in a speaker's lexicon of any language is of the same order of magnitude as the number of single words (Jackendoff, 1997; Erman and Warren, 2000), thus being of crucial interest for language modelling and for the computational representation of linguistic expressions in general.

One distinctive aspect of MWEs is that they fall on a *continuum* of idiomaticity (Sag et al., 2002; Fazly et al., 2009; King and Cook, 2017), as their meaning may or may not be inferred from one of their constituents (e.g., *research project* being a type of 'project', vs. *brass ring* meaning a 'prize'). In this regard, obtaining a semantic representation of a sentence which contains potentially idiomatic expressions involves both the correct identification of the MWE itself, and an adequate representation of the meaning of that expression in that particular context. As an example, it is expected that the representation of the expression *big fish* will be similar to that of *important person* in an idiomatic context, but closer to the representation of *large fish* when conveying its literal meaning.

Classic approaches to representing MWEs obtain a compositional vector by combining the representations of their constituent words, but these operations tend to perform worse for the idiomatic cases. In fact, it has been shown that the degree of idiomaticity of a MWE can be estimated by measuring the distance between a compositional vector (obtained from the vectors of its components) and a single representation learnt from the distribution of the MWE in a large corpus (Cordeiro et al., 2019).

Recent approaches to identify and classify MWEs take advantage of the contextualised representations provided by neural language models. On the one hand, some studies suggest that pre-training based on masked language modeling does not properly encode idiomaticity in word representations (Nandakumar et al., 2019; Garcia et al., 2021b,a). However, as these embeddings encode contextual information, supervised approaches using these representations tend to obtain better results in different tasks dealing with (non-)compositional semantics (Shwartz and Dagan, 2019; Fakharian and Cook, 2021; Zeng and Bhat, 2021).

As such, this shared task[1,2] presents two Subtasks: i) Subtask A, to test a language model's

[1]Task website: https://sites.google.com/view/semeval2022task2idiomaticity
[2]GitHub:https://github.com/H-TayyarMadabushi/SemEval_2022_Task2-idiomaticity



*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107 - 121

ability to detect idiom usage, and ii) Subtask B, to test the effectiveness of a model in generating representations of sentences containing idioms. Each of these Subtasks are further presented in two *settings*: Subtask A in the Zero Shot and One Shot settings so as to evaluate models on their ability to detect previously unseen MWEs, and Subtask B in the Pre Train and the Fine Tune settings to evaluate models on their ability to capture idiomaticity both in the absence and presence of training data. Additionally, we provide strong baselines based on pre-trained transformer-based language models and release our codetr which participants can build upon.

## 2 Related Tasks

The computational treatment of MWEs has been of particular interest for the NLP community, and several shared tasks with different objectives and resources have been carried out.

The SIGLEX-MWE Section[3] has organised various shared tasks, starting with the exploratory *Ranking MWE Candidates* competition at the MWE 2008 Workshop, aimed at ranking MWE candidates in English, German and Czech.[4] More recently, together with the PARSEME community, they have conducted three editions of a shared task on the automatic identification of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018, 2020). In these cases, the objective is to identify both known and unseen verb-based MWEs in running text and to classify them under a set of predefined categories. Interestingly, these PARSEME shared tasks provide annotation guidelines and corpora for 14 languages, and include 6 categories (with additional subclasses) of verbal MWEs.

The *Detecting Minimal Semantic Units and their Meanings (DiMSUM 2016)* shared task (Schneider et al., 2016) consisted of the identification of *minimal semantic units* (including MWEs) in English, and labelling some of them according to a set of semantic classes (supersenses).

Focused on the interpretation of noun compounds, the *Free Paraphrases of Noun Compounds* shared task of SemEval 2013 (Hendrickx et al., 2013) proposed to generate a set of free paraphrases of English compounds. The paraphrases should be ranked by the participants, and the evaluation is performed comparing these ranks against a list of paraphrases provided by human annotators.

Similarly, the objective of the SemEval 2010 shared task on *The Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions* (Butnariu et al., 2010) was to rank verbs and prepositions which may paraphrase a noun compound adequately in English (e.g., *olive oil* as 'oil *extracted from* olive', or *flu shot* as 'shot *to prevent* flu').

Apart from these competitions, various studies have addressed different tasks on MWEs and their compositionality, such as: classifying verb-particle constructions (Cook and Stevenson, 2006), identifying light verb constructions and determining the literality of noun compounds (Shwartz and Dagan, 2019), identifying and classifying idioms in running text (Zeng and Bhat, 2021), as well as predicting the compositionality of several types of MWEs (Lin, 1999; McCarthy et al., 2003; Reddy et al., 2011; Schulte im Walde et al., 2013; Salehi et al., 2015).

## 3 Dataset Creation

The dataset used in this task extends that introduced by Tayyar Madabushi et al. (2021), also including Galician data along with Portuguese and English. Here we describe the four step process used in creating this dataset.

The first step was to compile a list of 50 MWEs across the three languages. We sourced the MWEs in English and Portuguese from the Noun Compound Senses dataset (consisting of adjective-noun or noun-noun compounds) (Garcia et al., 2021b), which extends the dataset by Reddy et al. (2011) and provides human-judgements for compositionality on a Likert scale from 0 (non-literal/idiomatic) to 5 (literal/compositional). To ensure that the test set is representative of different levels compositionality, we pick approximately 10 idioms at each level of compositionality (0-1, 1-2, ...). For Galician, we extracted noun-adjective compounds from the Wikipedia and the CC-100 corpora (Wenzek et al., 2020) using the following procedure: First, we identified those candidates with at least 50 occurrences in the corpus. They were randomly sorted, and a native speaker and language expert of Galician selected 50 compounds from the list. The language expert was asked to take into account both the compositionality of the compounds (including idiomatic, partly idiomatic, and literal expressions),

[3] https://multiword.org/

[4] http://multiword.sourceforge.net/mwe2008

and their ambiguity (trying to select potentially idiomatic examples, i.e. compounds which can be literal or idiomatic depending on the context).

In the second step of the dataset creation process, in English and Portuguese, annotators were instructed to obtain between 7 and 10 examples for each possible meaning of each MWE from news stories available on the web, thus giving between 20 and 30 total examples for each MWE. Each example consisted of three sentences: the target sentence containing the MWE and the two adjacent sentences. Annotators where explicitly instructed to select high quality examples, where neither of the two adjacent sentences were empty and, preferably, from the same paragraph. They were additionally required to flag examples containing novel meanings, so such new meanings of MWEs could be incorporated into the dataset. Sentences containing MWEs in Galician were directly obtained from the Wikipedia and the CC-100 corpora due to the sparsity of Galician data on the web. During this annotation step, we follow the method introduced by Tayyar Madabushi et al. (2021), and add two additional labels: 'Proper Noun' and 'Meta Usage'. 'Meta Usage' represents cases wherein a MWE is used literally, but within a metaphor (e.g. *life vest* in "Let the Word of God be our life vest to keep us afloat, so as not to drown.").

In the third phase, across all three languages, each possible meaning of each MWE was assigned a paraphrase by a language expert. For example, the compositional MWE *mailing list* had the associated paraphrase 'address list' added, whereas the idiomatic MWE *elbow room* had the associated paraphrases 'joint room', 'freedom' and 'space' added to correspond to each of its possible meanings. Language experts focused on ensuring that these paraphrases were as short as possible, so the resultant adversarial paraphrases could be used to evaluate the extent to which models capture nuanced differences in each of the meanings.

The final phase of the process involved the annotation of each example with the correct paraphrase of the relevant MWE. This was carried out by two annotators, and any disagreements were discussed (in the case of Galician, in the presence of a language expert) and cases where annotators were not able to agree were discarded.

### 3.1 The Competition Dataset

We use the training and development splits from Tayyar Madabushi et al. (2021) with the addition of Galician data, and use the test split released by them as the evaluation split during the initial practice phase of the competition. We create an independent test set consisting of examples with new MWEs, and this set was used to determine the teams' final rankings. The labels for the evaluation and test sets are not released. We note that the competition is still active (in the 'post-evaluation' phase), and open for submissions from anyone[5].

Since one of the goals of this task is to measure the ability of models to perform on previously unseen MWEs (Zero Shot) and on those for which they have very little training data (One Shot), we extract, where available, exactly one idiomatic and one compositional example associated with each MWE in the test data, which is released as associated One Shot training data.

The final dataset consisted of 8,683 entries and the breakdown of the dataset is shown in Table 1. For further details on the training, development and practice evaluation splits, we direct readers to the work by Tayyar Madabushi et al. (2021). It should be noted that this original dataset does not contain data from Galician and so the only training data available in Galician was the One Shot training data. This was to evaluate the ability of models to transfer their learning across languages, especially to one that is low resourced.

| | Language | | | |
|---|---|---|---|---|
| Split | English | Portuguese | Galician | All |
| train | 3487 | 1290 | 63 | 4840 |
| dev | 466 | 273 | 0 | 739 |
| eval | 483 | 279 | 0 | 762 |
| test | 916 | 713 | 713 | 2342 |
| All | 5352 | 2555 | 776 | 8683 |

Table 1: Breakdown of the full dataset by language and data split.

## 4 Task Description and Evaluation Metrics

SemEval-2022 Task 2 aims to stimulate research into a difficult area of NLP, that of handling non-compositional, or idiomatic, expressions. Since this is an area of difficulty for existing language

[5] https://competitions.codalab.org/competitions/34710

models, we introduce two Subtasks; the first Subtask relates to idiomaticity detection, whilst the second relates to idiomaticity representation, success in which will require models to correctly encode idiomaticity. It is hoped that these tasks will motivate the development of language models better able to handle idiomaticity. Since we wish to promote multilingual models, we require all participants to submit results across all three languages. Both Subtasks are available in two settings, and participants are given the flexibility to choose which settings they wish to take part in.

### 4.1 Subtask A: Idiomaticity Detection

The first Subtask is a binary classification task, where sentences must be correctly classified into 'idiomatic' (including 'Meta Usage') or 'non-idiomatic' / literal (including 'Proper Noun'). Each example consists of the target sentence and two context sentences (sourced from either side of the target sentence) along with the relevant MWE. Some examples from this Subtask are shown in Table 2.

This Subtask is available in two settings: Zero Shot and One Shot. In the Zero Shot setting, the MWEs in the training set are disjoint from those in the development and test sets. Success in this setting will require models to generalise to unseen MWEs at inference time. In the One Shot setting, we include in the training set one idiomatic and one non-idiomatic example for each MWE in the development and test sets. This breakdown is shown in Table 3.

We use macro F1 score between the gold labels and predictions as the evaluation metric for this Subtask, due to the imbalanced datasets.

### 4.2 Subtask B: Idiomaticity Representation

The second Subtask is a novel idiomatic semantic textual similarity (STS) task, introduced by Tayyar Madabushi et al. (2021), where, given two input sentences, models must return an STS score between 0 (least similar) and 1 (most similar), indicating the similarity of the sentences. This requires models to correctly encode the meaning of non-compositional MWEs (idioms) such that the encoding of a sentence containing an idiomatic phrase (e.g. "I initially feared that taking it would make me a **guinea pig**.") and the same sentence with the idiomatic phrase replaced by a (literal) paraphrase (e.g. "I initially feared that taking it would make me a **test subject**.") are semantically similar to each other. Notice also that these two sentences, which mean the same thing, must necessarily be equally similar to any other third sentence. We choose this third sentence to be the sentence with the idiomatic phrase replaced by an *incorrect* literal paraphrase (e.g. "I initially feared that taking it would make me a **pig**."). Such a sentence is the ideal adversarial example, and ensures that we test if models are making use of an incorrect meaning of the MWE in constructing a sentence representation.

Data for this Subtask is generated in the following manner: MWEs in sentences are replaced by the literal paraphrase of one of its associated meanings. For example, the MWE 'guinea pig' in the sentence "I initially feared that taking it would make me a *guinea pig*." is replaced by one of the literal paraphrases 'test subject' or 'pig' (see Table 4). Crucially, these replacements can either be with the correct paraphrase, or one that is incorrect. As such, there are two cases:

- The MWE has been replaced by its correct paraphrase. In this case, the similarity should be 1.
  $sim(E, E_{\rightarrow \mathrm{c}}) = 1$
- The MWE has been replaced by its incorrect paraphrase. In this case, we require the model to give equivalent semantic similarities between this and the sentence where the MWE has been replaced by its correct paraphrase, and this and the original sentence.
  $sim(E, E_{\rightarrow \mathrm{i}}) = sim(E_{\rightarrow \mathrm{c}}, E_{\rightarrow \mathrm{i}})$

Importantly, the task requires models to be *consistent*. Concretely, the STS score for the similarity between a sentence containing an idiomatic MWE and that same sentence with the MWE replaced by the correct paraphrase must be equal to *one* as this would imply that the model has correctly interpreted the meaning of the MWE. In the case where we consider the incorrect paraphrase, we check for consistency by requiring that the STS between the sentence containing the MWE and a sentence where the MWE is replaced by the incorrect paraphrase is equal to the STS between the sentence where the MWE is replaced by the correct paraphrase and one where it is replaced by the incorrect one. Notice, that all this does, is to require the model to, once again, interpret the meaning of the MWE to be the same (or very similar) to the correct literal paraphrase of that MWE. More formally, we require models to output STS scores for each example $E$ such that:

| Language | MWE | Sentence | Label |
|---|---|---|---|
| English | old hat | Serve our favorite bourbon whiskeys in an **old hat** and we'd still probably take a sip or two. | 1 |
| English | old hat | But not all of the accouterments of power are **old hat** for the president. | 0 |
| Portuguese | força bruta | **Força Bruta** vai reunir alguns dos homens mais fortes do mundo. | 1 |
| Portuguese | força bruta | Gardner é conhecido por ser impulsivo e usar os poderes com grande impacto, de forma instintiva, com **força bruta**. | 0 |
| Galician | porta grande | Á esquerda da **porta grande**, en terra, observamos a tumba de "Don Manuel López Vizcaíno. | 1 |
| Galician | porta grande | Os dous dominadores da Copa Galicia 2017 regresaron pola **porta grande** ao certame autonómico na súa quinta xornada. | 0 |

Table 2: Examples for Subtask A. Note that the label 1 is assigned to non-idiomatic usage, which includes proper nouns, as in the Portuguese example.

| | | Language | | | |
|---|---|---|---|---|---|
| Train Split | MWEs | English | Portuguese | Galician | All |
| Zero Shot | 236 | 3327 | 1164 | 0 | 4491 |
| One Shot | 250 | 160 | 126 | 63 | 349 |
| Total | 486 | 3487 | 1290 | 63 | 4840 |

Table 3: Breakdown of the training data into zero shot and one shot. Note that the MWEs in the zero shot and one shot data are disjoint.

$$\forall_{i \in I} \Big( sim(E, E_{\rightarrow \mathrm{c}}) = 1; \\ sim(E, E_{\rightarrow \mathrm{i}}) = sim(E_{\rightarrow \mathrm{c}}, E_{\rightarrow \mathrm{i}}) \Big) \quad (1)$$

In Equation 1 above, $E_{\rightarrow \mathrm{c}}$ represents an example containing the MWE $E$, wherein that MWE is replaced by its *correct* contextual paraphrase. $E_{\rightarrow \mathrm{i}}$ on the other hand, represents the example wherein the MWE $E$ is replaced by one of its *incorrect* contextual paraphrases. Examples for this Subtask are shown in Table 4.

Since this task relies on models' ability to correctly assign STS scores for sentences with do not contain idiomatic MWEs, we additionally include standard STS data in our test data. This has the added benefit of preventing models from overfitting on the MWE dataset. We include this STS evaluation data from the STS Benchmark dataset (Cer et al., 2017) in English and the ASSIN2 STS dataset (Real et al., 2020) in Portuguese. There is no available STS data for Galician, so none is included. We use the Spearman's rank correlation coefficient between the two sets of STS scores generated by models as the evaluation metric in this Subtask. We do not use Pearson correlation as it has been shown to be a poor indicator of performance on STS tasks (Reimers et al., 2016).

This Subtask is also available in two settings: the Pre Train setting and the Fine Tune setting. In the Pre Train setting, we require that models are *not* trained on idiomatic STS data. However, models can be trained (including "fine-tuned") on any other training objective (such as during the pre-training of language models). The Fine Tune setting, on the other hand, allows all training regimes, including the fine-tuning on any idiomatic STS dataset.

### 4.3 Baselines

In order to generate baseline results, we used pre-trained transformer-based (Vaswani et al., 2017) language models. We use multilingual BERT (Devlin et al., 2019) to benefit from cross-lingual transfer. For both settings in Subtask A, we simply Fine Tune the pre-trained model on the training data provided. For the Zero Shot setting, we include the context sentences, whereas in the One Shot setting, we exclude the context sentences but add the MWE as a second sentence. This is based on the best-performing approaches found by Tayyar Madabushi et al. (2021).

For Subtask B Pre Train, we introduce single tokens for each MWE in the data. This is motivated by the 'idiom principle' (Sinclair and Sinclair, 1991), which hypothesises that humans process idioms by treating them as a single unit. Since BERT embeddings cannot be directly used for STS, we create a sentence transformer model (Reimers and Gurevych, 2019) using multilingual BERT with these added tokens, and train it on the English and Portuguese STS data. Importantly, the new tokens introduced for MWEs are randomly initialised and no continued pre-training is performed. As such, they serve to 'break compositionality' rather than to create more effective representations of MWEs. This breaking of compositionality has been shown to be effective by Tayyar Madabushi et al. (2021).

For the Fine Tune setting, the same approach is taken, although no training is done on the STS

| Sentence (E) | Correct Replacement ($E_{\text{MWE}\to\text{c}}$) | Wrong Replacement ($E_{\text{MWE}\to\text{i}}$) | Expected |
|---|---|---|---|
| And finally, the snow falls again, this time in a thick, **wet blanket** that encapsulates everything. | And finally, the snow falls again, this time in a thick, **damp blanket** that encapsulates everything. | And finally, the snow falls again, this time in a thick, **killjoy** that encapsulates everything. | $sim(E, E_{\to\text{c}}) = 1$<br>$sim(E, E_{\to\text{i}}) = sim(E_{\to\text{c}}, E_{\to\text{i}})$ |
| I initially feared that taking it would make me a **guinea pig**. | I initially feared that taking it would make me a **test subject**. | I initially feared that taking it would make me a **pig**. | $sim(E, E_{\to\text{c}}) = 1$<br>$sim(E, E_{\to\text{i}}) = sim(E_{\to\text{c}}, E_{\to\text{i}})$ |

Table 4: Examples for Subtask B. For brevity we only include examples in English.

data, and instead we Fine Tune on the training data provided. This lack of training on the STS data is intentional as we intend to establish the effectiveness of the MWE based training data, and are reflected by the comparatively lower scores on the STS subsection of the test data (Table 8).

It should be noted that these baseline methods that make use of multilingual BERT are particularly strong when compared to typical 'baselines'. This is intentional as we aim to promote the development of models that are comparable to the current state-of-the-art.

## 5 Participating Systems and Results

Twenty five teams in total participated, with the most participants to Subtask A Zero Shot (20). The results for the individual Subtasks are given in Table 5, Table 6, Table 7 and Table 8. Here we discuss the methods used by the best-performing teams as well as some interesting approaches. Full details of methods used by participants is given in Appendix A.

### 5.1 Subtask A Zero-Shot

Of the twenty teams that submitted to this setting, 12 reported using transformer-based approaches. The best-performing team (clay) used different masking strategies during pretraining, and performed finetuning with data augmentation (including back-translation, Edunov et al., 2018) as well as using soft-label finetuning (a knowledge distillation approach). The team in second (yxb) used a multilingual T5 model (Xue et al., 2021) with various data augmentation techniques including: back-translation; synonym replacement; random insertion, swap, and deletion. They also used an alternative loss function for unbalanced data, called focal loss (Lin et al., 2017). The third team (NER4ID; Tedeschi and Navigli, 2022) used a dual-encoder architecture to encode the MWE and its context, then predicted idiomaticity by looking at the similarity score. This approach has a precedent in previous work that hypothesises the semantic similarity between a MWE and its context to be a good indicator of idiomaticity (Liu and Hwa, 2018). They also implemented named entity recognition as an intermediate step which they found provided great improvements. Interestingly, two teams (UAlberta; Hauer et al., 2022, and Unimelb_AIP) used unsupervised approaches, i.e. not using any of the provided training data. UAlberta were able to beat the baseline using translation information from resources such as Open Multilingual Wordnet (Bond and Foster, 2013) and BabelNet (Navigli and Ponzetto, 2010). They hypothesised that for idiomatic MWEs, the individual words are less likely to share mult-synsets with their translations. They also used a POS tagger for identifying proper nouns.

### 5.2 Subtask A One Shot

The best-performing team (HIT; Chu et al., 2022) used XLM-R (Conneau et al., 2020), and added '[SEP]' tokens around the relevant MWE in the target sentence, unless it was capitalised, in which case they excluded these tokens. This is an alternative approach to that of Tayyar Madabushi et al. (2021), where the MWEs were added as a second sentence. They also used R-Drop (Wu et al., 2021) as a regularisation method. The second best team (kpfriends; Sik Oh, 2022) used an ensemble of checkpoints with soft-voting. They also started with XLM-RoBERTa (large) trained on CoNLL. Interestingly, this team had the largest difference in performance across the two settings of Subtask A (coming in 16th in the Zero Shot setting). The third best team (UAlberta; Hauer et al., 2022) used a transformer-based classifier with additional features of glosses for the individual words of the relevant MWE. They hypothesised that this would help for determining compositionality, since the meaning of compositional MWEs could be deduced from the glosses of the individual words. An interesting approach was taken by MaChAmp (van der Goot, 2022), who used multi-task learning across multiple SemEval tasks (2, 3, 4, 6, 10, 11, 12), pretrain-

| Ranking | Team | Language | | | All |
|---|---|---|---|---|---|
| | | English | Portuguese | Galician | |
| 1 | clay | 0.9016 | 0.8277 | 0.9278 | 0.8895 |
| 2 | yxb | 0.8948 | 0.8395 | 0.7524 | 0.8498 |
| 3 | NER4ID (Tedeschi and Navigli, 2022) | 0.8680 | 0.7039 | 0.6550 | 0.7740 |
| 4 | HIT (Chu et al., 2022) | 0.8242 | 0.7591 | 0.6866 | 0.7715 |
| 5 | Hitachi (Yamaguchi et al., 2022) | 0.7827 | 0.7607 | 0.6631 | 0.7466 |
| 6 | OCHADAI (Pereira and Kobayashi, 2022) | 0.7865 | 0.7700 | 0.6518 | 0.7457 |
| 7 | yjs | 0.8253 | 0.7424 | 0.6020 | 0.7409 |
| 8 | CardiffNLP-metaphors (Boisson et al., 2022) | 0.7637 | 0.7619 | 0.6591 | 0.7378 |
| 9 | Mirs | 0.7663 | 0.7617 | 0.6429 | 0.7338 |
| 10 | Amobee | 0.7597 | 0.7147 | 0.6768 | 0.7250 |
| 11 | HYU (Joung and Kim, 2022) | 0.7642 | 0.7282 | 0.6293 | 0.7227 |
| 12 | Zhichun Road (Cui et al., 2022) | 0.7489 | 0.6901 | 0.5104 | 0.6831 |
| 13 | 海鲛NLP | 0.7564 | 0.6933 | 0.5108 | 0.6776 |
| 14 | UAlberta (Hauer et al., 2022) | 0.7099 | 0.6558 | 0.5646 | 0.6647 |
| 15 | Helsinki-NLP (Itkonen et al., 2022) | 0.7523 | 0.6939 | 0.4987 | 0.6625 |
| 16 | daminglu123 (Lu, 2022) | 0.7070 | 0.6803 | 0.5065 | 0.6540 |
| | baseline (Tayyar Madabushi et al., 2021) | 0.7070 | 0.6803 | 0.5065 | 0.6540 |
| 17 | kpfriends (Sik Oh, 2022) | 0.7256 | 0.6739 | 0.4918 | 0.6488 |
| 18 | Unimelb_AIP | 0.7614 | 0.6251 | 0.5020 | 0.6436 |
| 19 | YNU-HPCC (Liu et al., 2022) | 0.7063 | 0.6509 | 0.4805 | 0.6369 |
| 20 | Ryan Wang | 0.5972 | 0.4943 | 0.4608 | 0.5331 |
| N/A | JARVix (Jakhotiya et al., 2022)[6] | 0.7869 | 0.7201 | 0.5588 | 0.7235 |

Table 5: Results for Subtask A Zero Shot. The evaluation metric is macro F1 score, and the ranking is based on the 'All' column.

ing a Rebalanced mBERT (RemBERT) (Chung et al., 2020) model across all of the tasks, then retraining a model for each specific task. Since for this task we do not allow the use of additional data, we do not include this team in the ranking, but their score is reported for reference.

### 5.3 Subtask B Pre Train

No teams reported using non-transformer-based approaches for this setting. The best-performing team (drsphelps; Phelps, 2022) used a modification of the baseline with BERT for Attentive Mimicking (BERTRAM) (Schick and Schütze, 2020) to generate embeddings as replacements for the randomly-initialised one token embeddings used by the baseline. This method takes both form and context into account, thus not assuming total non-compositionality as the one-token method does. It should be noted that every team in this setting improved upon the baseline result.

### 5.4 Subtask B Fine Tune

No teams reported using non-transformer-based approaches for this setting. The best-performing team (YNU-HPCC; Liu et al., 2022) used a pretrained Sentence-BERT (Reimers and Gurevych, 2019) model, then finetuned using multiple negatives ranking loss (Henderson et al., 2017) and triplet loss. The second best team (drsphelps; Phelps, 2022) used an identical approach to that in Subtask B Pre Train, using BERTRAM (Schick and Schütze, 2020), with additional finetuning on the training data provided. The third best team (Eat Fish) used a multilingual model pretrained with knowledge distillation, as well as data augmentation.

### 5.5 Overview of Submissions

In Figure 1 we show the models that were mentioned in the submissions.

The majority of participants used transformer-based approaches, although in both settings for Subtask A there were three teams using other approaches. In Subtask B, as mentioned previously, no non-transformer approaches were mentioned, which is expected since this task was designed for the pretrain-finetune paradigm.

In Figure 2 we show the methods mentioned in more than one submission. Data augmentation approaches were popular, the most frequently-mentioned being back-translation (Edunov et al., 2018). Equally as popular were approaches using

[6]Not ranked due to only submitting to the 'post-evaluation' phase.

| Ranking | Team | Language | | | All |
|---|---|---|---|---|---|
| | | English | Portuguese | Galician | |
| 1 | HIT (Chu et al., 2022) | 0.9639 | 0.8944 | 0.9369 | 0.9385 |
| 2 | kpfriends (Sik Oh, 2022) | 0.9606 | 0.8993 | 0.9215 | 0.9346 |
| 3 | UAlberta (Hauer et al., 2022) | 0.9453 | 0.8918 | 0.9120 | 0.9243 |
| 4 | Zhichun Road (Cui et al., 2022) | 0.9344 | 0.8559 | 0.8927 | 0.9033 |
| 5 | clay | 0.9181 | 0.8423 | 0.9313 | 0.9022 |
| 6 | YNU-HPCC (Liu et al., 2022) | 0.9179 | 0.8633 | 0.8781 | 0.8948 |
| 7 | CardiffNLP-metaphors (Boisson et al., 2022) | 0.9464 | 0.8385 | 0.8545 | 0.8934 |
| 8 | yxb | 0.8995 | 0.8266 | 0.8781 | 0.8779 |
| 9 | NER4ID (Tedeschi and Navigli, 2022) | 0.9079 | 0.8179 | 0.8695 | 0.8771 |
| 10 | HYU (Joung and Kim, 2022) | 0.9159 | 0.8457 | 0.8287 | 0.8750 |
| 11 | yjs | 0.9199 | 0.8365 | 0.8294 | 0.8747 |
| | baseline (Tayyar Madabushi et al., 2021) | 0.8862 | 0.8637 | 0.8162 | 0.8646 |
| 12 | Mirs | 0.7570 | 0.7549 | 0.6712 | 0.7367 |
| 13 | daminglu123 (Lu, 2022) | 0.7486 | 0.7085 | 0.6004 | 0.7040 |
| 14 | 海鲛NLP | 0.7649 | 0.7156 | 0.5134 | 0.6851 |
| 15 | OCHADAI (Pereira and Kobayashi, 2022) | 0.7069 | 0.6445 | 0.5235 | 0.6573 |
| 16 | Ryan Wang | 0.3314 | 0.4058 | 0.3779 | 0.4044 |
| N/A | MaChAmp (van der Goot, 2022)[7] | 0.7204 | 0.6247 | 0.5532 | 0.6607 |
| N/A | JARVix (Jakhotiya et al., 2022)[8] | 0.8410 | 0.8162 | 0.7918 | 0.8243 |

Table 6: Results for Subtask A One Shot. The evaluation metric is macro F1 score, and the ranking is based on the 'All' column.

| Ranking | Team | Subset | | All |
|---|---|---|---|---|
| | | Idiom Only | STS Only | |
| 1 | drsphelps (Phelps, 2022) | 0.4030 | 0.8641 | 0.6402 |
| 2 | colorful | 0.4290 | 0.8880 | 0.6262 |
| 3 | Mirs | 0.3750 | 0.8623 | 0.6038 |
| 4 | Zhichun Road (Cui et al., 2022) | 0.2826 | 0.8359 | 0.5632 |
| 5 | YNU-HPCC (Liu et al., 2022) | 0.2872 | 0.7125 | 0.5577 |
| 6 | ALTA | 0.2154 | 0.8608 | 0.5379 |
| | baseline (Tayyar Madabushi et al., 2021) | 0.2263 | 0.8311 | 0.4810 |

Table 7: Results for Subtask B Pre Train. The evaluation metric is Spearman correlation, and the ranking is based on the 'All' column.

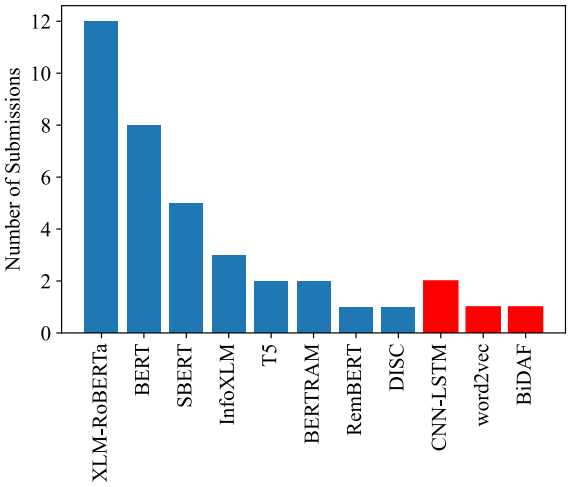


Figure 1: Models mentioned in the submissions. In blue are models that use transformers either wholly or partially, whilst in red are alternative models.

alternative loss functions.

## 6 Methods

The primary goal of this shared task was to provide a platform for the evaluation of a variety of methods for the identification and represention of MWEs. This section gives an overview of the methods that have been successful in each of the Subtasks. In particular, we attempt to identify the combination of methods across submissions that have significant potential for future development.

### 6.1 Subtask A

Subtask A, the identification of MWEs, comprised two settings: Zero Shot and One Shot. Crucially, the results from the task show that *methods that are successful in the Zero Shot setting, fail to be*

[7]Not ranked due to using a multi-task learning approach.

[8]Not ranked due to only submitting to the 'post-evaluation' phase.

| Ranking | Team | Subset | | All |
|---|---|---|---|---|
| | | Idiom Only | STS Only | |
| 1 | YNU-HPCC (Liu et al., 2022) | 0.4277 | 0.6637 | 0.6648 |
| 2 | drsphelps (Phelps, 2022) | 0.4124 | 0.8188 | 0.6504 |
| 3 | Eat Fish | 0.3688 | 0.8660 | 0.6475 |
| 4 | Zhichun Road (Cui et al., 2022) | 0.3956 | 0.5615 | 0.6401 |
| | baseline (Tayyar Madabushi et al., 2021) | 0.3990 | 0.5961 | 0.5951 |
| 5 | ALTA | 0.2566 | 0.6156 | 0.5755 |

Table 8: Results for Subtask B Fine Tune. The evaluation metric is Spearman correlation, and the ranking is based on the 'All' column.

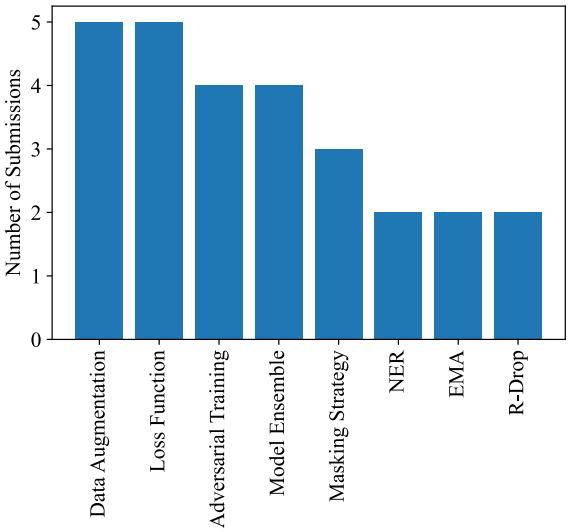


Figure 2: Methods mentioned in more than one submission.

*successful in the One Shot setting and vice versa.* The two problems seem to require capabilities that are quite distinct. This seems intuitive when translated into the kind of thinking that one might use in identifying idioms: When one hears an idiom for the first time, we are likely to recognise that it sounds 'idiom-like' based on our prior understanding of idioms, whereas when we come across an idiom that we are familiar with, we link our existing knowledge of that idiom with the current instance of it.

This seems to play out in the successful models in this Subtask, as the general trend amongst the methods that were successful in the Zero Shot setting, with one exception, is the generalisation of models using regularisation, data augmentation or dropout. While regularisation did feature in the top performing model in the One Shot setting, it seems to have been less important to generalise models when they had access to as little as one training example associated with each model. The best performing linguistically motivated method – which compares the semantic similarity between the MWE span and that of the surrounding context – ranked third in the Zero Shot setting, although it performed 11 points below the best performing method. This is of particular interest as this method has previously been shown to be extremely powerful in detecting idiomaticity in non-contextual models.

Models successful in the One Shot setting, again with one exception, seem to be those which are more powerful at extracting cues from the minimal training examples and tended to be larger, ensembled or trained to a larger extent using adversarial training. The best performing method which incorporated elements based on linguistic theory also ranked third in this setting and incorporated the gloss of each individual word in the target MWE to aid in models' ability to detect compositionality.

Interestingly, the use of the idiom principle in creating single token representations for MWEs is absent amongst the methods used for this Subtask. While such a comparison would have been interesting, it is hardly surprising that this method is not amongst those used, given that the cost of pre-training with new MWE tokens is rather high.

### 6.2 Subtask B

Subtask B, the novel task of creating contextual representations of MWEs which are consistent with the paraphrased version of that MWE as measured by Spearman's rank correlation, coefficient also had two settings: the first without associated training examples (Pre Train) and the second with (Fine Tune). Since the sentence embeddings generated by pre-trained language models cannot be directly compared for similarity, such models must be altered so as to be used for this Subtask. Additionally, as pointed out by Tayyar Madabushi et al. (2021), the MWEs contained within sentences can be represented using single tokens even without pre-training, as the 'breaking' of compositionality

itself produces more accurate representations of sentences containing MWEs.

As such, models that perform the best on the Pre Train setting focus on the creation of more accurate single token representations of MWEs, while the top performing models on the Fine Tune setting, in general, focus on optimising sentence similarity. This seems to be consistent with the observation by Tayyar Madabushi et al. (2021) that fine-tuning is indeed a reasonable way of learning the representation of MWEs. It should be noted that these trends are less certain since there are fewer participants on this Subtask, some of whom do not share their methods, and the one team that we know used a method of learning new representations of MWEs is ranked first in the Pre Train setting but ranked second in the Fine Tune setting.

## 7 Conclusions and Future work

We present, in this paper, 'SemEval 2022 Task2: Multilingual Idiomaticity Detection and Sentence Embedding', consisting of two Subtasks: i) Subtask A, to test a language model's ability to detect idiom usage, and ii) Subtask B, to test a model's ability to generate representations of sentences containing idioms. This task, aimed at boosting research into the detection and representation of idiomatic expressions, had submissions from 25 teams consisting of close to 100 participants.

We additionally provide an overview and analysis of the methods used by participants, which we believe will help future research in this field. In particular, we highlight the need for distinct methods when detecting MWEs that have been previously seen and when detecting ones that have not. In representing idiomatic expressions, we show, through the novel idiomatic STS task presented here, that models are rather effective when they have training data available, but, as demonstrated in the Pre Train setting, more methods of encoding MWEs are required when training data is not available.

While the top performing methods across this task have been driven by deep neural models independent of linguistic features, we highlight that this does not imply that the addition of linguistically motivated features does not lead to improvements on the task. Instead, it points to the possibility of integrating these methods into the more powerful neural models in future work where an ablation study might shed more light on the impact of each feature.


### Acknowledgements

This work was partially supported by the UK EPSRC grant EP/T02450X/1, by the CDT in Speech and Language Technologies and their Applications (UKRI grant number EP/S023062/1), by a *Ramón y Cajal* grant (RYC2019-028473-I), and by the grant ED431F 2021/01 (Galician Government).


## References


Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Ltd.

Joanne Boisson, Jose Camacho-Collados, and Luis Espinosa-Anke. 2022. Cardiffnlp-metaphor at semeval-2022 task 2: Targeted fine-tuning of transformer-based language models for idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. SemEval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 39–44, Uppsala, Sweden. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Zheng Chu, Ziqing Yang, Yiming Cui, Zhigang Chen, and Ming Liu. 2022. Hit at semeval-2022 task 2: Pre-trained language model for idioms detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

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.

Paul Cook and Suzanne Stevenson. 2006. Classifying particle semantics in English verb-particle constructions. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 45–53, Sydney, Australia. Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

Xuange Cui, Wei Xiong, and Songlin Wang. 2022. Zhichunroad at semeval-2022 task 2: Adversarial training and contrastive learning for multiword representations. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Britt Erman and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1):29–62.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. 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.

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, and Grzegorz Kondrak. 2022. Ualberta at semeval 2022 task 2: Leveraging glosses and translations for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Sami Itkonen, Jörg Tiedemann, and Mathias Creutz. 2022. Helsinki-nlp at semeval-2022 task 2: A feature-based approach to multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ray Jackendoff. 1997. Twistin' the Night Away. *Language*, 73:534–559.

Yash Jakhotiya, Vaibhav Kumar, Ashwin Pathak, and Raj Shah. 2022. Jarvix at semeval-2022 task 2: It takes one to know one? idiomaticity detection using zero and one shot learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Youngju Joung and Taeuk Kim. 2022. Hyu at semeval-2022 task 2: Effective idiomaticity detection with consideration at different levels of contextualization. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Milton King and Paul Cook. 2017. Supervised and unsupervised approaches to measuring usage similarity. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 47–52, Valencia, Spain. Association for Computational Linguistics.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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.

Kuanghong Liu, Jin Wang, and Xuejie Zhang. 2022. Ynu-hpcc at semeval-2022 task 2: Representing multilingual idiomaticity based on contrastive learning. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Daming Lu. 2022. daminglu123 at semeval-2022 task 2: Using bert and lstm to do text classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan. Association for Computational Linguistics.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

Lis Pereira and Ichiro Kobayashi. 2022. Ochadai at semeval-2022 task 2: Adversarial training for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Dylan Phelps. 2022. drsphelps at SemEval-2022 Task 2: Learning idiom representations using BERTRAM. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

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, online. Association for Computational Linguistics.

Carlos Ramisch and Aline Villavicencio. 2018. Computational Treatment of Multiword Expressions. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, 2nd edition. Oxford University Press.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 1–15, Mexico City, Mexico. Springer, Berlin, Heidelberg.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

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

Timo Schick and Hinrich Schütze. 2020. BERTRAM: Improved word embeddings have big impact on contextualized model performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online. Association for Computational Linguistics.

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.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Min Sik Oh. 2022. kpfriends at semeval-2022 task 2: Neamer - named entity augmented multi-word expression recognizer. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

John Sinclair and Les Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press, USA.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477.

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.

Rob van der Goot. 2022. Machamp at semeval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task multilingual learning for a pre-selected set of semantic datasets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Atsuki Yamaguchi, Gaku Morio, Hiroaki Ozaki, and Yasuhiro Sogawa. 2022. Hitachi at semeval-2022 task 2: On the effectiveness of span-based classification approaches for multilingual idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

# A Full Breakdown of Methods

All participants were invited to submit a short description of their methods, as well as to submit a paper. In Table 9, Table 10, Table 11, and Table 12 we give all the method descriptions that were submitted.

| Ranking | Team | Method |
|---|---|---|
| 1 | clay | "domain pretraining with different masking strategies finetuning with data augmentation such as back-translation finetuning with soft label from former checkpoint" |
| 2 | yxb | "use mT5-Base use Easy Data Augmentation techniques include back-translation, synonym replacement, random insertion, random swap, random deletion include label unbalanced loss function：focal loss use model ensemble" |
| 3 | NER4ID | "Dual-encoder (Transformer-based) architecture that encodes both the potentially idiomatic expression and its context, and predicts idiomaticity by looking at their similarity score: high similarity -> compositional, low similarity -> idiomatic. Another core contribution of our method is the use of Named Entity Recognition as an intermediate step to pre-identify some non-idiomatic expressions; this provides great improvements." |
| 4 | HIT | "1. we use the big pre-trained model, XLM-R-large. Compared with multilingual-BERT and XLM-R-base, XLM-R-large is obviously improved. 2. Separate the exact same phrases as MWE in the target sentence with the sep token. If the phrase in the sentence is capitalized, It is more likely to be named entities that the model can distinguish, so the sep tokens are not added around the capitalized phrases. 3. Using Regularized Dropout(r-drop) as regularization." |
| 5 | Hitachi | "Our approach is built on top of multilingual pre-trained language models, which include InfoXLM and XLM-R. We solve the task of multilingual idiomaticity detection as a binary classification task and follow the standard fine-tuning method except not using a special [CLS] representation for classification. Instead, we first take an average over MWE's span representations and subsequently feed the averaged representation into a linear layer for classification." |
| 6 | OCHADAI | "our model relies on pre-trained contextual representations from different multilingual state-of-the-art transformer-based language models (i.e., multilingual BERT and XLM-RoBERTa), and on adversarial training, a training method for further enhancing model generalization and robustness." |
| 7 | yjs | "For each input sentence in the training set, if the MWE is idiomatic then its corresponding tokens are labeled as "idiomatic" and the remaining tokens are labeled as "literal"; if the MWE is literal then all the tokens in the sequence are labeled as "literal". Method 1: We apply a Bi-Directional Attention Flow (BiDAF) network (Seo et al., 2017), while we use mBERT as the contextualised embedding, and we use pos tag embedding as its query input." |
| 8 | CardiffNLP-m | "CardiffNLP-metaphors submitted the results of two methods in total, applied both for Task A Zero Shot and one-shot. The first method uses xlm-roberta-large and the second uses several monolingual bert language models for English, Portuguese and Galician. For the Zero Shot settings, bert-multilingual-base is used to label the Galician sentences, because no Galician examples were included in the training set. The embedding of the three sentences and the embeddings of the isolated target are input of the models. We optimized the models over different training parameters on the development set." |
| 9 | Mirs | - |
| 10 | Amobee | - |
| 11 | HYU | "We devise four features ((i), (ii), (iii), and (iv) in the following) as input for a simple yet effective idiomaticity classifier that is a multi-layer perceptron with one hidden layer.<br>First, to consider the contextualized semantics of a target sentence when influenced by its surrounding context, we concatenate the target sentence with its (i) previous and (ii) next sentences respectively and inject the two chunks into our feature extractor (XLM-R; a bidirectional multilingual language model) independently to generate two distinct ((i) and (ii)) features.<br>While constructing the aforementioned features, we also introduce two techniques to clarify the presence of a MWE in the sequence: the first highlights the location of the MWE with a new, dedicated positional encoding, and the second appends the MWE once again at the end of the sequence.<br>In addition, we focus on the way of better utilizing the information existing solely in the target sentence, regarding a MWE and its context (i.e., phrases in the target sentence except for the MWE) as separate ones.<br>Specifically, we derive (iii) the "context-only" representation of the target sentence by using a variant of the target sentence where the MWE is masked, while we compute (iv) the "MWE-only" representation, which corresponds to the intrinsic meaning of the MWE irrespective of context, by inserting only the MWE into the feature extractor." |
| 12 | Zhichun Road | "1. We use InfoXLM-Base as text encoder. (performance: infoxlm > XLM-R > Mbert) 2.We use exponential moving average (EMA) method. 3.We use adversarial attack strategy(performance: Smart > freeLb > PGD = FGM). Finally，our approach ranked 12th." |
| 13 | 海鲛NLP | - |
| 14 | UAlberta | "Our unsupervised translation-based approach leverages translation information in multilingual resources such as OMW and BabelNet. The hypothesis is that the translations of idiomatic MWEs tend to be non-compositional, and therefore the individual words of an MWE are less likely to share mult-synsets with their translations. In addition, since MWEs that are named entities are usually literal, we use a part-of-speech tagger to identify proper nouns." |
| 15 | Helsinki-NLP | "The system utilizes linguistically motivated features that typically characterize idiomatic expressions: non-substitutability, non-compositionality and affectiveness. This feature model is based on pre-trained models and classification pipelines that have been integrated into the transformers library provided by HuggingFace. The final classification combines the feature model with either sentence-transformers or a base BERT model. The system also adds a back-translation feature and applies simple post-correction rules based on boolean features." |
| 16 | daminglu123 | "We used the same model as baseline but added one more LSTM layer at last." |
| 17 | kpfriends | "We experimented with various inductive training methods only using Zero Shot data provided. We are still experimenting various schemes, including novel MWE ideas. We will share the findings in our paper." |
| 18 | Unimelb_AIP | "We tackled this task in an unsupervised way (i.e. without using any portion of the training data). First, we trained a standard CBOW word2vec model on unlabelled data and used it to predict the top-500 words that would fit into the surrounding context of the target MWE (as performed during the training of the CBOW model). Then, we calculated the maximum cosine similarities between the predicted words and each MWE component word, and regarded the MWE as "literal" ("non-idiomatic") if they are higher than the mean cosine similarity between the component words and their 500 closest words. Finally, we ensembled five CBOW models trained with different window sizes (5, 10, 15, 20, and 30) to incorporate different levels of contextual information. One limitation of this approach is that it often classifies proper-noun and idiomatic usages into the same class ("non-literal"; as their surrounding contexts differ a lot from the literal usage ones), and to mitigate this problem, we always regarded MWEs as "non-idiomatic" if they contained any capital letter." |
| 19 | YNU-HPCC | "As for methods of the best submission results, we added a linear layer so as to choose effective information from all of output layer that were extracted by pre-trained model, XLM-RoBERTa, and then fine-tuned it to classify." |
| 20 | Ryan Wang | "CNN-bidirectional LSTM classifier with jointly trained word embeddings trained on full passages (target and context) from Zero Shot data" |
| N/A | JARVix | "we fine-tune a pretrained XLNet on the task dataset (after evaluating multiple large language models and their majority-voting ensemble)." |

Table 9: Methods used in Subtask A Zero Shot. Note: CardiffNLP-m is short for CardiffNLP-metaphors.

| Ranking | Team | Method |
|---|---|---|
| 1 | HIT | "Mostly the same as the zero-shot. We train the One Shot model initialized from the best Zero Shot checkpoint. We additionally post-processed the predictions based on the distribution of the labels in the One Shot train file." |
| 2 | kpfriends | "More than 10 checkpoints were created per "English" and "Spanish / Galician" and inferred separately, later ensembled using soft-voting. To stabilize training of xlm-roberta-large, we started with pre-trained models provided by Huggingface which were xlm-roberta-large trained on CoNLL. We also had some good results with xlm-roberta-base. We will deep dive into methodology and interesting observations / error analysis in our paper." |
| 3 | UAlberta | "Our method uses a transformer-based sequence classifier that takes as an input the context sentence and the glosses of each individual word in the target multi-word expression. The intuition is that the addition of the glosses to the input might help the classifier to detect if the meaning of the target multi-word expression can be deduced from the definitions of the individual words, i.e., if it is compositional. Note that this method is applicable to both settings." |
| 4 | Zhichun Road | "1. We use InfoXLM-Base as text encoder. (performance: infoxlm > XLM-R > Mbert) 2.We use exponential moving average (EMA) method. 3.We use adversarial attack strategy(performance: freeLB > Smart > PGD = FGM). Finally，our approach ranked 4th." |
| 5 | clay | "same as Zero Shot setting, but with more data include Zero Shot data and One Shot data" |
| 6 | YNU-HPCC | "As for methods of the best submission results, we simply concated sentence and MWE and input into pre-trained model, XLM-RoBERTa. CLS from last layer was extracted to classify." |
| 7 | CardiffNLP-m | "CardiffNLP-metaphors submitted the results of two methods in total, applied both for Task A Zero Shot and one-shot. The first method uses xlm-roberta-large and the second uses several monolingual bert language models for English, Portuguese and Galician. For the Zero Shot settings, bert-multilingual-base is used to label the Galician sentences, because no Galician examples were included in the training set. The embedding of the three sentences and the embeddings of the isolated target are input of the models. We optimized the models over different training parameters on the development set. xlm-roberta-large significantly ouperforms the monolingual experimental settings on the one shot track. " |
| 8 | yxb | "use mT5-Base use Easy Data Augmentation techniques include back-translation, synonym replacement, random insertion, random swap, random deletion include label unbalanced loss function：focal loss use model ensemble" |
| 9 | NER4ID | "Same as zero-shot" |
| 10 | HYU | "In One Shot setting, we used the same method as in a Zero Shot setting." |
| 11 | yjs | "Method 2: We used the BiDAF-based DISC architecture by (Zeng and Bhat, 2021). DISC firstly combine GLOVE embeddings and POS embeddings with a BiDAF layer, which is then infused with mBERT by another BiDAF layer. We use both methods in the two settings, Method 1 performs better than Method 2. In the submissions, the different results is caused by different random seeds, with/without previous and next sentences, and with/without MWE." |
| 12 | Mirs | - |
| 13 | daminglu123 | "We used the same model as baseline but added one more LSTM layer at last." |
| 14 | 海鲛NLP | - |
| 15 | OCHADAI | "our model relies on pre-trained contextual representations from different multilingual state-of-the-art transformer-based language models (i.e., multilingual BERT and XLM-RoBERTa), and on adversarial training, a training method for further enhancing model generalization and robustness." |
| 16 | Ryan Wang | "CNN-bidirectional LSTM classifier with jointly trained word embeddings trained on full passages (target and context) from zero- and One Shot data" |
| N/A | MaChAmp | "Multi-task learning across SemEval tasks (2, 3, 4, 6, 10, 11, and 12). First we Pre Train a RemBERT multi-task model across all the tasks. Then we re-train a model for each task specifically. We used the default hyperparameters of MaChAmp v0.3 for all settings, which were finetuned on the GLUE benchmark and UD_English-EWT." |
| N/A | JARVix | "we use a relation network (Sung, et. al 2018) to find a similarity (or a dissimilarity) score between a query and it's same MWE support set, and assign a label accordingly. For this, we also evaluate a siamese network with a similar inference methodology." |

Table 10: Methods for Subtask A One Shot. Note: CardiffNLP-m is short for CardiffNLP-metaphors.

| Ranking | Team | Method |
|---|---|---|
| 1 | drsphelps | "Our model is a modification of the baseline system with the randomly initialised word embeddings for the one token MWEs replaced with embeddings created using Schick and Schutze's BERT for Attentive Mimicking (BERTRAM). BERTRAM models are trained for Portuguese and Galician alongside the provided English model, and examples use to create the MWE emebddings are taken from the common crawl corpora for English, Portuguese, and Galician. Further pretraining (up to 45 epochs) is done on the sentence transformers." |
| 2 | colorful | - |
| 3 | Mirs | - |
| 4 | Zhichun Road | "1.We add CrossAttention-Module at the top of the Sentence-Bert. ( Including train and evaluate). 2.We add an extra Contrastive Loss. Finally, our approach ranked 4th." |
| 5 | YNU-HPCC | "As for methods of the best submission results, we extracted first-last-average vector and used an optimized method called CoSENT to train model. In comparison to SBERT, it could solve the problem of difference in process of training and prediction and get a better results." |
| 6 | ALTA | - |

Table 11: Methods for Subtask B Pre Train.

| Ranking | Team | Method |
|---|---|---|
| 1 | YNU-HPCC | "As for methods of the best submission results, both multiple-negatives-ranking-loss and triplet-loss function combined with pre-trained model, distiluse-base-multilingual-cased-v1, were used to fine-tune. " |
| 2 | drsphelps | "Using the models trained for the Pre Train setting, fine-tuning is performed using the provided training data, just as in the baseline system. The best overall performance is found after fine tuning for one epoch, however training for up to 50 epochs can drastically increase Spearman Rank scores for the idiom only data, while causing much less performance drop on the general STS data." |
| 3 | Eat Fish | "Multilingual model which was pretrained by using knowledge distillation Data augmentation Extract multiword from exist multiword package Two state training trick" |
| 4 | Zhichun Road | "1.We add CrossAttention-Module at the top of the Sentence-Bert. ( Including train and evaluate). 2.We add an extra Contrastive Loss. Finally, our approach ranked 4th." |
| 5 | ALTA | - |

Table 12: Methods for Subtask B Fine Tune.