Contextualized Embeddings Encode Monolingual and Cross-lingual Knowledge of Idiomaticity

Samin Fakharian and **Paul Cook**
Faculty of Computer Science, University of New Brunswick
Fredericton, NB E3B 5A3 Canada
{samin.fakharian,paul.cook}@unb.ca

Abstract

Potentially idiomatic expressions (PIEs) are ambiguous between non-compositional idiomatic interpretations and transparent literal interpretations. For example, *hit the road* can have an idiomatic meaning corresponding to 'start a journey' or have a literal interpretation. In this paper we propose a supervised model based on contextualized embeddings for predicting whether usages of PIEs are idiomatic or literal. We consider monolingual experiments for English and Russian, and show that the proposed model outperforms previous approaches, including in the case that the model is tested on instances of PIE types that were not observed during training. We then consider cross-lingual experiments in which the model is trained on PIE instances in one language, English or Russian, and tested on the other language. We find that the model outperforms baselines in this setting. These findings suggest that contextualized embeddings are able to learn representations that encode knowledge of idiomaticity that is not restricted to specific expressions, nor to a specific language.

1 Introduction

Multiword expressions (MWEs) are lexicalized combinations of multiple words, which display some form of idiomaticity (Baldwin and Kim, 2010). In this paper we focus on potentially-idiomatic expressions (PIEs), i.e., expressions which are ambiguous between a semantically-opaque idiomatic interpretation, and a compositional literal meaning. In the following example, the English PIE *hit the road* has an idiomatic meaning corresponding roughly to 'start a journey':

1. The marchers had hit the road before 0500 hours and by midday they were limping back having achieved success on day one.

On the other hand, *hit the road*, can also be used literally, as in the example below:

2. Two climbers dislodged another huge block which hit the road within 18 inches of one of the estate's senior guides.[1]

PIEs occur across languages, with one particularly common class of PIE cross-lingually being verb–noun combinations (VNCs, Fazly et al., 2009) — i.e., PIEs consisting of a verb with a noun in its direct object position — such as *hit the road* in the example above. Although VNCs are common, PIEs also occur in other syntactic constructions, with English examples including combinations of a verb and prepositional phrase — e.g., *skating on thin ice* (which can be used idiomatically to mean roughly 'at risk') — and prepositional phrases — e.g., *off the hook* (with a potential idiomatic meaning of roughly 'out of danger'). Distinguishing between literal and idiomatic usages of PIEs could be particularly important for down-stream natural language processing applications such as machine translation (Isabelle et al., 2017).

Previous work has considered both unsupervised and supervised approaches to predicting the token-level idiomaticity of PIEs. However, annotated data to train supervised approaches is not available for all PIEs in all languages. This makes unsupervised approaches (e.g., Fazly et al., 2009; Haagsma et al., 2018; Liu and Hwa, 2018; Kurfalı and Östling, 2020), which do not have this resource requirement, appealing. On the other hand, supervised approaches (e.g., Salton et al., 2016; King and Cook, 2018) tend to outperform unsupervised approaches, but are restricted to languages and PIEs for which annotated training data is available.

In this paper we consider supervised approaches based on contextualized embeddings (Devlin et al., 2019; Liu et al., 2019; Kuratov and Arkhipov, 2019) to predicting usages of PIEs as idiomatic

[1] These example sentences are taken, with light editing, from the VNC-Tokens dataset (Cook et al., 2008).

Proceedings of the 17th Workshop on Multiword Expressions, pages 23–32
Bangkok, Thailand (online), August 6, 2021.

or literal; however, we measure the ability of these approaches to generalize to expressions that were not observed during training, and also to generalize across languages. We begin by considering monolingual experiments for English and Russian in which we train and test on instances of the same PIEs. For English, we focus on VNCs (Cook et al., 2008). For Russian, we consider a wider-range of types of PIEs (Aharodnik et al., 2018). We then consider a second monolingual setting in which we evaluate on PIEs, again either English or Russian, that were not observed during training. Finally, we consider cross-lingual detection of idiomaticity. Here we train on instances of PIEs in one language, English or Russian, and evaluate on instances of PIEs in the other language.

Our findings evaluating on expressions that were observed during training are similar to those of (Kurfalı and Östling, 2020); we achieve strong improvements over baselines, and on English outperform previous approaches based on conventional word embeddings (King and Cook, 2018). In monolingual experiments evaluating on PIEs that were not observed during training, we again improve over baselines, and in the case of English, also over a strong linguistically-informed unsupervised baseline. In cross-lingual experiments, in which the model is evaluated on instances of PIEs in a language that was not observed during training, we again improve over baselines, and remarkably observe performance roughly on par with that of monolingual experiments evaluating on expressions not observed during training. These findings suggest that contextualized embeddings are able to learn representations that encode knowledge of idiomaticity that is not restricted to specific expressions, nor to a specific language.

2 Related Work

Previous work has considered unsupervised and supervised approaches to predicting the token-level idiomaticity of PIEs. Although unsupervised methods have been proposed to disambiguate a wide range of kinds of potentially-idiomatic expressions (Haagsma et al., 2018; Liu and Hwa, 2018; Kurfalı and Östling, 2020), and are not limited to languages and types of PIEs for which training data is available, these approaches tend to not perform as well as supervised approaches.

Focusing on specific languages and types of expressions can improve unsupervised approaches. For example, focusing on VNCs, the idiomatic interpretations of VNCs are typically lexico-syntactically fixed. Returning to the *hit the road* example from Section 1, the idiomatic interpretation is typically not accessible if the determiner is indefinite (e.g., *hit a road*), the noun is plural (e.g., *hit the roads*), or the voice is passive (e.g., *the road was hit*); in such cases typically only the literal interpretation is available. Fazly et al. (2009) propose an unsupervised statistical method based on the lexico-syntactic fixedness of VNCs to determine the canonical forms — with respect to the determiner, number of the noun, and voice of the verb — of VNCs. They observe that idiomatic usages of VNCs tend to occur in canonical forms, and that literal usages tend to occur in non-canonical forms. A strong, linguistically-informed unsupervised baseline for distinguishing literal from idiomatic VNC usages is therefore to label canonical form usages as idiomatic, and non-canonical form usages as literal.

Salton et al. (2016) propose a supervised approach to predicting the token-level idiomaticity of PIEs, focusing on English VNCs, based on training an SVM on skip-thoughts (Kiros et al., 2015) representations of sentences containing PIEs. King and Cook (2018) achieve better results using a simpler sentence representation based on average of word embeddings. Moreover, King and Cook show that adding a single binary feature to the sentence representation indicating whether the VNC occurs in a canonical form — based on the method of Fazly et al. (2009) — gives substantial improvements. Hashempour and Villavicencio (2020) propose a supervised approach in which PIE instances are treated as single units by fusing their lexicalized component words, and learning representations of these units using word and contextualized (Melamud et al., 2016; Devlin et al., 2019) embeddings. Hashempour and Villavicencio also focus on VNCs. Although they show improvements by treating VNC instances as fused units, they do not outperform King and Cook; they do, however, train their models on smaller corpora. Shwartz and Dagan (2019) use representations of spans of tokens based on contextualized embedding for predicting a range of MWE properties. Most closely related to our work, they consider light-verb construction and verb-particle construction classification, for both of which there is an ambiguity between MWE usages and similar-on-the-surface literal combina-

tions. Shwartz and Dagan do not, however, consider English VNCs or Russian idioms as we do.

Kurfalı and Östling (2020) propose a supervised approach to classifying instances of potentially-idiomatic expressions, as idiomatic or literal, based on contextualized embeddings. They represent MWE instances as the average of the contextual embeddings for the tokenized pieces of their lexicalized component words, which are lemmatized in a preprocessing step, and use a single-layer perceptron for classification. Their findings indicate that their approach improves over previous approaches on English and German PIEs. In this paper, similarly to Kurfalı and Östling, we consider an approach based on contextualized embeddings, but we consider experimental setups in which classifiers are evaluated on expressions, and also languages, that are unobserved during training.

3 Predicting PIE Idiomaticity with Contextualized Embeddings

Previous supervised approaches to identifying idiomatic instances of PIEs have represented PIE instances with sentence embeddings (Salton et al., 2016; King and Cook, 2018). We consider a similar approach here using contextualized embeddings from BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), RuBERT (Kuratov and Arkhipov, 2019), and mBERT (Devlin et al., 2019). Specifically we represent a PIE instance using the CLS (classification) token for the context in which it occurs.[2] For representing English PIEs we use the sentence in which the target expression occurs as the context. For representing Russian PIEs, the dataset we use (discussed in Section 4.1) does not include sentence segmentation, and so we instead use a context of up to 300 characters to the left and right of the target expression.[3]

Because we focus on VNCs for English experiments, following King and Cook (2018), for our monolingual experiments on English VNCs, we also consider whether incorporating information about lexico-syntactic fixedness of VNCs into our approach gives improvements. Specifically, we concatenate a single binary feature indicating whether a VNC usage is in a canonical form, referred to as CF, with the representation of the CLS token.

We fine tune pre-trained BERT, RoBERTa, RuBERT, and mBERT models for binary classification of PIE token instances as idiomatic or literal. We use two fully-connected layers on top of the contextualized embedding model. The first layer has the same dimensionality as the representation of the VNC (i.e., 768 dimensions, the hidden layer size of each of the contextualized embedding models considered, and an additional dimension when the CF feature is used) and uses the ReLU activation function. The second layer has 512 dimensions and uses the softmax activation function.

4 Materials and Methods

In this section we describe our datasets (Section 4.1), experimental setups and evaluation metric (Section 4.2), and then the implementation of our models and the parameter settings used (Section 4.3).

4.1 Datasets

Following Salton et al. (2016), King and Cook (2018), and Hashempour and Villavicencio (2020), for English, we use the VNC-Tokens dataset (Cook et al., 2008), which consists of English VNC usages extracted from the British National Corpus (Burnard, 2000) annotated as literal or idiomatic.[4] VNC-Tokens includes DEV (development) and TEST sets — referred to here as EN-DEV and EN-TEST to distinguish them from the Russian dataset introduced below — which each include roughly 600 instances of 14 VNC types. The expressions in EN-DEV and EN-TEST do not overlap. Each of EN-DEV and EN-TEST is roughly balanced with respect to idiomatic and literal instances. We use EN-DEV for hyper-parameter tuning, and carry out no such tuning on EN-TEST.

For Russian, we use the dataset of Aharodnik et al. (2018) which consists of instances of Russian PIEs annotated at the token level as literal or idiomatic. Unlike the English dataset, this dataset is not restricted to VNCs. It includes id-

[2] In preliminary experiments we also considered representations of English VNC instances formed by averaging and concatenating contextualized representations of the verb and noun components of a target VNC (where the verb and noun representations are themselves averages of the representations of the word pieces they are segmented into). We found these approaches to perform roughly on par with representing VNC instances using the CLS token, and so only consider this approach here.

[3] We did not attempt to tune this context window size, although there is scope to do so in future work.

[4] Following Salton et al. (2016), King and Cook (2018), and Hashempour and Villavicencio (2020), we ignore instances labelled as unknown in VNC-Tokens.

Dataset	# expressions	# tokens	% idiomatic
EN-DEV	14	594	60.9
EN-TEST	14	613	63.3
RUSSIAN	37	775	54.3

Table 1: The number of PIE types and tokens, and the percentage of idiomatic tokens, in each dataset.

ioms with a range of syntactic constructions including preposition+noun, preposition+adj+noun, and VNCs. The dataset consists of three sections containing classical prose, modern prose, and text from Russian Wikipedia. We consider only the Russian Wikipedia portion because classical prose is substantially older than the text in the English VNC-Tokens dataset (which is from the British National corpus, which primarily includes texts from the late twentieth century), and the modern prose portion is relatively small compared to the Russian Wikipedia portion, which includes roughly 500M tokens. Each instance is accompanied by a context window of up to three paragraphs. Meta-data for this dataset indicating the location of the target expression in the context unfortunately does not appear to be available. We therefore restrict our experiments to the subset of this dataset for which there is an exact match between the target expression and a token sequence in the context. This gives a dataset consisting of 37 expressions and 775 token instances.[5] The dataset is again roughly balanced between idiomatic and literal usages with 54.3% being idiomatic. In contrast to the English dataset, we do not split this Russian dataset at the type level into separate DEV and TEST datasets because we carry out no hyper-parameter tuning on this dataset. We refer to this dataset as RUSSIAN.

Statistics for the number of PIE types and tokens, and the percentage of idiomatic tokens, in each dataset, are given in Table 1.

4.2 Experimental Setups and Evaluation

We first consider an experimental setup similar to King and Cook (2018) and Kurfalı and Östling (2020), referred to here as "all expressions". In this monolingual experimental setup we train and test on instances of the same PIEs in the same language. For each of EN-DEV, EN-TEST, and RUSSIAN, we randomly partition the instances into training (roughly 75%) and testing (roughly 25%) sets, keeping the ratio of idiomatic to literal usages of each expression balanced across the training and testing sets. We repeat this random partitioning 10 times. For EN-DEV and EN-TEST we use the same partitions as King and Cook.

We do not expect to have annotated instances of all PIE types, limiting the applicability of models developed for the all expressions experimental setup. We are therefore particularly interested in determining whether a supervised model is able to generalize to expressions that were unseen during training. Here we consider a second monolingual experimental setup proposed by Gharbieh et al. (2016), referred to here as "unseen expressions". In these experiments we hold out all instances of one PIE type for testing, and train on all instances of the remaining types (within either EN-DEV, EN-TEST, or RUSSIAN). We repeat this fourteen times for each of EN-DEV and EN-TEST, and 37 times for RUSSIAN, holding out each PIE type once for testing.

For both experimental setups — i.e., all expressions and unseen expressions — we train and test models on EN-DEV for preliminary experiments and setting parameters. We then report final results by training and testing models on EN-TEST and RUSSIAN.

Just as we do not expect to have annotated instances of all PIE types for a given language, we also do not expect to have annotated instances of PIEs for all languages. We therefore consider an extension of the monolingual unseen expressions experimental setup in which we evaluate on instances of PIEs in a language that was not observed during training, referred to as "cross-lingual". In these experiments we train on either English or Russian, and evaluate on the other language. In particular, we train on either EN-DEV or EN-TEST and evaluate on RUSSIAN, and also train on RUSSIAN and evaluate on each of EN-DEV and EN-TEST.

The idiomatic and literal classes for both the English and Russian datasets are roughly balanced (Table 1). We therefore evaluate using accuracy. For the all expressions experimental setup, we report average accuracy across the 10 runs. In the unseen expressions experimental setup, we repeatedly hold out each expression until all instances of each expression (within either EN-DEV, EN-TEST, or RUSSIAN) have been classified, and then compute accuracy. For the cross-lingual experiments,

[5] The entire Russian Wikipedia portion of the dataset consists of 40 expressions and 799 token instances. Restricting the dataset to instances that have an exact match with the target expression therefore still retains the majority of the data.

we simply calculate accuracy over all instances in the dataset used for testing.

4.3 Implementation and Parameter Settings

We use Huggingface (Wolf et al., 2020) implementations of BERT, RoBERTa, mBERT, and RuBERT. Specifically we use bert-base-uncased, roberta-base, bert-base-multilingual-cased, and rubert-base-cased. All models have 12 layers and a hidden layer size of 768. The number of parameters for BERT, RoBERTa, mBERT, and RuBERT, is 125M, 125M, 179M, and 180M, respectively. BERT and RoBERTa are trained on uncased and cased English text, respectively. mBERT is trained on text from 104 languages. RuBERT is trained on Russian Wikipedia and Russian news data. We use BERT, RoBERTa, and mBERT for monolingual English experiments; RuBERT and mBERT for monolingual Russian experiments; and mBERT for cross-lingual experiments.

We train our models using Adam optimizer (Kingma and Ba, 2015) to minimize the cross-entropy loss. We use the default dropout of 0.5 for the network layers which are on top of BERT, RoBERTa, mBERT, or RuBERT. For fine-tuning, Devlin et al. (2019) recommend the following parameter settings: batch size of 8, 16, or 32; epochs between 2 and 4; and learning rate of 2e-5, 3e-5, or 5e-5.

We perform grid search over these parameter settings on EN-DEV for the monolingual all expressions and unseen expressions experimental setups. We report results for the best parameter settings on EN-DEV, and then use only these parameter settings for experiments on EN-TEST and RUSSIAN. For the cross-lingual experiments, we do no further parameter tuning, and report results for the best parameter settings for the unseen expressions experimental setup for EN-DEV. We repeat the experiments 10 times with different random seeds, and report the mean accuracy and standard deviation over the runs.

5 Monolingual Results

In this section, we present results for the unseen and all expressions experimental setups, for monolingual experiments on English (Section 5.1) and Russian (Section 5.2). In Section 6 we present results for cross-lingual experiments.

5.1 English

For English, we compare against three baselines: a most-frequent class (MFC) baseline, the unsupervised approach of Fazly et al. (2009, CForm) based on canonical forms, and the supervised approach of King and Cook (2018).

We begin by considering results for the all expressions experimental setup. Results are shown in the top panel of Table 2 (labelled "All"). On each dataset, both BERT and RoBERTa outperform all baselines, including King and Cook (2018) when using the canonical form (CF) feature (indicated by "+CF" in Table 2). This finding demonstrates that contextualized embeddings are able to better capture knowledge of the idiomaticity of PIEs than previous approaches. mBERT performs relatively poorly compared to BERT and RoBERTa, although it still outperforms the baselines, with the exception of King and Cook when using the CF feature.

We now examine the impact of the CF feature in the all expressions experimental setup.[6] For each model based on contextualized embeddings, incorporating the CF feature gives an improvement, but these improvements are small relative to the standard deviation across runs. This is in contrast to the substantial improvements obtained by King and Cook (2018) when using the CF feature. These findings suggest that contextualized embeddings are able to better capture the linguistic knowledge encoded in this feature than conventional word embeddings, which King and Cook use to represent VNC instances.

We now consider results for the unseen expressions experimental setup. Results are shown in the bottom panel of Table 2 (labelled "Unseen"). On EN-DEV, the best results are again obtained using BERT, however, the accuracy drops substantially on EN-TEST. RoBERTa performs more consistently across EN-DEV and EN-TEST, and performs best on EN-TEST. mBERT again performs relatively poorly compared to BERT and RoBERTa, but nevertheless substantially outperforms the most-frequent class baseline.

Focusing on the contribution of the CF feature, results for both BERT and RoBERTa on EN-DEV

[6] We do not consider the CF feature, which was developed for and evaluated on English VNCs (Fazly et al., 2009), for experiments with mBERT. We are primarily interested in mBERT as a point of comparison for cross-lingual experiments, and so do not incorporate this English-specific knowledge here. We also do not consider the CF feature in experiments on RUSSIAN or in cross-lingual experiments.

Setup	Model	EN-DEV		EN-TEST	
		−CF	+CF	−CF	+CF
All	MFC	63.4	63.4	62.9	62.9
	CForm	75.0	75.0	71.1	71.1
	King and Cook (2018)	82.5	85.6	81.5	84.7
	BERT	**90.7** ±0.53	**90.8** ±0.51	**89.3** ±1.11	**89.8** ±0.71
	RoBERTa	88.3 ±0.96	89.9 ±0.66	88.6 ±0.87	89.0 ±0.48
	mBERT	84.1 ±0.8	-	83.8 ±1.1	-
Unseen	MFC	60.9	60.9	63.3	63.3
	CForm	73.6	73.6	70.0	70.0
	King and Cook (2018)	72.3	76.4	74.6	77.8
	BERT	**83.5** ±0.97	**83.4** ±0.65	78.6 ±1.78	79.8 ±1.55
	RoBERTa	81.8 ±1.60	82.4 ±1.20	**82.3** ±1.76	**80.6** ±2.35
	mBERT	75.4 ±1.5	-	74.3 ±2.2	-

Table 2: % accuracy and standard deviation for the all and unseen expressions experimental setups on EN-DEV and EN-TEST, for BERT, RoBERTa, and mBERT, with and without the CF feature. % accuracy for the baselines is also shown. The best accuracy for each experimental setup, on each dataset, with and without the CF feature, is shown in boldface.

do not show a clear improvement when incorporating this feature when considering the standard deviation across runs. The impact of this feature in experiments on EN-TEST is similar. This finding again suggests that contextualized embeddings capture much of the linguistic knowledge encoded in this feature. We therefore focus on results for BERT and RoBERTa that do not incorporate the CF feature.

Focusing on results for EN-TEST (for which no hyper-parameter tuning was carried out), given the substantial improvements over the most-frequent class baseline, and over the CForm baseline, with the exception of mBERT when accounting for variation across runs, these findings suggest that the classifiers (including the approach of King and Cook) have learned information about the idiomaticity of PIEs, that is not restricted to specific expressions, as in the case of the all expressions experimental setup. Furthermore BERT and RoBERTa (without the CF feature) outperform the approach of King and Cook (2018), although given the standard deviation across runs, this difference does not appear to be significant for BERT when comparing against the approach of King and Cook when they use the CF feature.

In experiments until now we have used representations from the final layer of contextualized embedding models (BERT, RoBERTa, and mBERT). We now consider the effect of using different hidden layers, focusing on the unseen expressions experimental setup for BERT and RoBERTa, in an effort to explain the relatively poor performance of BERT here. Results are shown in Table 3.[7] In all cases, except for BERT on EN-TEST, the final layer performs best. This is inline with the findings of Jawahar et al. (2019) that the upper layers of BERT encode semantic information. For BERT, where accuracy was low on EN-TEST relative to EN-DEV in Table 2, on EN-TEST the second last layer performs best.

Model	Dataset	Layer			
		9	10	11	12
BERT	EN-DEV	82.0	82.2	82.6	**83.5**
BERT	EN-TEST	79.2	79.8	**80.2**	78.6
RoBERTa	EN-DEV	75.6	78.2	79.8	**81.8**
RoBERTa	EN-TEST	71.8	77.7	79.5	**82.3**

Table 3: % accuracy and standard deviation for the unseen expressions experimental setup on EN-DEV and EN-TEST using BERT and RoBERTa with representations from the indicated layers. The best results for each model and dataset are shown in boldface.

5.2 Russian

For monolingual experiments on Russian, we again consider the all and unseen expressions experimental setups. Here we compare against a most-frequent class baseline. Although Aharodnik et al.

[7] Results are only shown for layers 9–12. The overall trend for other layers is that lower layers achieve lower accuracy.

Setup	Model	% Accuracy
All	MFC	54.1
	RuBERT	87.4 ±4.7
	mBERT	**88.2** ±2.8
Unseen	MFC	54.3
	RuBERT	**74.6** ±2.2
	mBERT	73.6 ±3.8

Table 4: % accuracy and standard deviation for the all and unseen expressions experimental setups on RUSSIAN for RuBERT, mBERT, and the most-frequent class baseline (MFC). The best accuracy for each experimental setup is shown in boldface.

(2018) report preliminary results on this dataset, they are not for the same experimental setups that we consider, and so we do not compare against their results. Here we consider RuBERT, a monolingual Russian model, and mBERT, which includes Russian text in its pre-training. For the all and unseen expressions experimental setups we use the best hyper-parameter settings for EN-DEV using BERT for the unseen and all expressions experimental setups, respectively; i.e., we do not do any hyper-parameter tuning on RUSSIAN.

Results are shown in Table 4. We see that in both the all and unseen expressions experimental setups, both RuBERT and mBERT substantially outperform the most-frequent class baseline. We also see that, accounting for variation across runs, the performance of RuBERT and mBERT is similar within each experimental setup.

These findings add to those of Section 5.1, and again indicate that contextualized embeddings encode knowledge of PIE idiomaticity, although in this case the experiments consider a range of PIE syntactic constructions, as opposed to only VNCs. These findings also again indicate that the classifier for the unseen expressions experimental setup has learned information about the idiomaticity of PIEs that is not restricted to expressions that were observed during training. In the following section we consider whether contextualized embeddings encode knowledge of idiomaticity that can be generalized across languages.

6 Cross-lingual Results

In this section we consider cross-lingual experiments in which we train on instances of PIEs in a source language, and evaluate on instances of PIEs in a (different) target language. We consider the case of both English-to-Russian and Russian-to-English. For English we consider both EN-DEV and EN-TEST. In these experiments we train on the entire source language dataset (i.e., when Russian is the source language we train on RUSSIAN, and when English is the source language we train on either EN-DEV or EN-TEST), and evaluate on the entire target language dataset. We use the best hyper-parameter settings for EN-DEV using BERT for the unseen expressions experimental setup from Section 5.1; i.e., we do not attempt any hyper-parameter tuning for this cross-lingual experimental setup. We again compare results against a most-frequent class baseline, and when English is the target language, also against the unsupervised CForm baseline (Fazly et al., 2009).

Results are shown in Table 5. For English-to-Russian, and Russian-to-English, mBERT outperforms the most-frequent class baseline in each case. In experiments with English as the target language, mBERT also outperforms the CForm baseline, although in the case of EN-DEV the difference does not appear to be significant given the standard deviation across runs. Furthermore, the results are, remarkably, roughly on par with monolingual results for the unseen expressions experimental setup. Focusing on experiments involving EN-TEST and RUSSIAN, where for both datasets no hyper-parameter tuning was considered in previous experiments, for English-to-Russian (i.e., EN-TEST source, RUSSIAN target) mBERT achieves 72.4% accuracy, whereas in the monolingual Russian unseen expressions experimental setup, RuBERT and mBERT achieve accuracies of 74.6% and 73.6%, respectively (Table 4). These differences are relatively small considering the standard deviations across runs. For Russian-to-English (i.e., RUSSIAN source, EN-TEST target) mBERT achieves an accuracy of 80.1%, while the accuracies for contextualized embedding models for EN-TEST in the unseen expressions experimental setup range from 74.3% for mBERT to 82.3% for RoBERTa (Table 2).

Whereas the findings for the monolingual unseen expressions experimental setup indicate that the classifier is able to generalize to expressions that are unseen during training, these findings for cross-lingual experiments indicate that the classifier is able to generalize across languages. This suggests that the classifier has learned information about idiomaticity that is not restricted to specific expressions, nor to a specific language. The cross-

Source Language	Target language	Source dataset	Target dataset	Model	% Accuracy
English	Russian	EN-DEV	RUSSIAN	MFC	54.3
				mBERT	75.7 ±3.0
		EN-TEST	RUSSIAN	MFC	54.3
				mBERT	72.4 ±5.7
Russian	English	RUSSIAN	EN-DEV	MFC	60.9
				CForm	73.6
				mBERT	75.2 ±2.0
		RUSSIAN	EN-TEST	MFC	63.3
				CForm	70.0
				mBERT	80.1 ±1.3

Table 5: % accuracy and standard deviation for cross-lingual experiments from English to Russian (top panel) and Russian to English (bottom panel) using mBERT, a most-frequent class (MFC) baseline, and for English, the unsupervised CForm baseline.

lingual findings furthermore seem to be inline with the findings of Pires et al. (2019) that cross-lingual transfer with mBERT works reasonably well even when languages do not share the same script (as for English and Russian), but works less well when the languages do not share the same word order (where English is an SVO language, and Russian has freer word-order, but SVO is considered dominant (Dryer, 2013)).

7 Conclusions

In this paper we proposed a supervised model based on contextualized embeddings to predict the idiomaticity of PIE instances. In contrast to most prior work on this topic, we considered the ability of the model to generalize to expressions that were not observed during training, and also to generalize across languages. Code to reproduce these experiments is available.[8]

We first considered monolingual experiments for English, focusing on verb–noun combinations, a common type of PIE. In experiments in which we train and test on instances of the same PIEs, we demonstrated that an approach based on contextualized embeddings improves over previous approaches based on conventional word embeddings. We then considered experiments in which we evaluate on PIEs that were not observed during training, and showed that the proposed approach improves over a strong, linguistically-informed unsupervised baseline. We further found that, in contrast to prior models based on conventional word embeddings, incorporating information about the lexico-syntactic fixedness of VNCs does not lead to clear improvements, suggesting that contextualized embeddings capture this rich linguistic knowledge.

In monolingual experiments on Russian we considered a wider range of types of PIEs. Here we showed that, as for English, the proposed approach improves over baselines when evaluating on expressions that were, and were not, observed during training. The experimental setup in which the model is tested on instances of PIE types that were not observed during training is particularly interesting because we do not expect to have annotated instances of all PIE types available for training supervised models. The findings in this experimental setup, for both English and Russian, indicate that the model is capturing knowledge of PIE idiomaticity that is not restricted to specific expressions.

Finally, we considered cross-lingual experiments in which we train on instances of either English or Russian PIEs, and evaluate on PIE instances in the other language. Here the proposed model again improves over baselines, and achieves performance that is roughly on par with that of monolingual experiments in which we evaluate on PIEs that were not observed during training. This finding indicates that contextualized embeddings encode knowledge of PIE idiomaticity that is not restricted to specific expressions, nor to a specific language.

In future work, we plan to further explore cross-lingual idiomaticity prediction. We would like to include more languages in the analysis to be able to measure the impact of training on multiple source languages. We further intend to consider including

[8] https://github.com/SaminFakharian/Contextualized-Embeddings-Encode-Monolingual-and-Cross-lingual-Knowledge-of-Idiomaticity

the target language amongst the source languages, to measure the impact of augmenting training data for the target language with data from other languages. Finally, we intend to consider cross-lingual approaches for other MWE prediction tasks, such as predicting noun compound compositionality.

Acknowledgments

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada, the New Brunswick Innovation Foundation (NBIF), and the University of New Brunswick.

References

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a Russian idiom-annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2nd edition. CRC Press, Boca Raton, USA.

Lou Burnard. 2000. *The British National Corpus Users Reference Guide*. Oxford University Computing Services.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop on Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22, Marrakech, Morocco.

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.

Matthew S. Dryer. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

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.

Hessel Haagsma, Malvina Nissim, and Johan Bos. 2018. The other side of the coin: Unsupervised disambiguation of potentially idiomatic expressions by contrasting senses. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 178–184, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia. Association for Computational Linguistics.

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*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3276–3284. Curran Associates, Inc.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint*

Workshop on Multiword Expressions and Electronic Lexicons, pages 85–94, online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.