UAlberta at SemEval 2022 Task 2: Leveraging Glosses and Translations for Multilingual Idiomaticity Detection

Bradley Hauer, Seeratpal Jaura, Talgat Omarov, Grzegorz Kondrak
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,seeratpa,omarov,gkondrak}@ualberta.ca

Abstract

We describe the University of Alberta systems for the SemEval-2022 Task 2 on multilingual idiomaticity detection. Working under the assumption that idiomatic expressions are non-compositional, our first method integrates information on the meanings of the individual words of an expression into a binary classifier. Further hypothesizing that literal and idiomatic expressions translate differently, our second method translates an expression in context, and uses a lexical knowledge base to determine if the translation is literal. Our approaches are grounded in linguistic phenomena, and leverage existing sources of lexical knowledge. Our results offer support for both approaches, particularly the former.

1 Introduction

In this paper, we describe the University of Alberta systems for the task of classifying multi-word expressions (MWEs) in context as either *idiomatic* or *literal* (Tayyar Madabushi et al., 2022). Each instance in the data includes a MWE (e.g., *closed book*), its language, and its context, composed of the three surrounding sentences. We participate in both the zero-shot and one-shot settings.

While the exact definitions of the two key terms are not stated explicitly in the task description[1], it is suggested that *idiomatic* is synonymous with *non-compositional*. The Pocket Oxford Dictionary defines *idiomatic* as "not immediately comprehensible from the words used," and *literal* as "taking words in their basic sense." Therefore, we adopt the following MWE *compositionality criterion*

$$\text{literal} \equiv \text{compositional} \equiv \neg\,\text{idiomatic}$$

where the three terms are considered to be Boolean variables. In addition, the shared task considers all proper noun MWEs (e.g., *Eager Beaver*) as literal.

[1] https://sites.google.com/view/semeval2022task2-idiomaticity

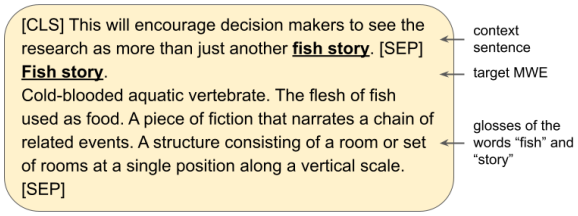

Figure 1: An example of defBERT input.

Our goal is to explore the idea that glosses and translations of word senses can help decide whether the meaning of a given MWE occurrence is compositional. Based on the above-stated compositionality criterion, this in turn could facilitate idiomaticity detection. In particular, we hypothesize that at least one of the words in any idiomatic expression is used in a non-standard sense. Following the intuition that a traditional word sense disambiguation (WSD) system can only identify senses that are included in a given sense inventory, we propose two methods that indirectly detect non-standard senses by leveraging either glosses or translations of senses from such an inventory.

Our gloss-based method follows from the intuition that the meaning of a given MWE occurrence is related to any of the existing sense glosses of its component words *only if the expression is compositional*. Therefore, the addition of the glosses to the context of the expression should help the classifier in deciding whether the MWE is used in a literal or idiomatic sense. We implement this method by adding the glosses of each sense of each individual word, retrieved from a lexical knowledge base, to the input to a neural classifier which fine-tunes multilingual BERT (mBERT; Devlin et al., 2019) for the idiomaticity detection task. We refer to this method as defBERT (Figure 1).

Our translation-based method follows from the observation that compositional expressions are typically translated word-for-word ("literally"), which implies that each content word and its translation

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 145 - 150

should have the same meaning. Therefore, each such multilingual word pair should share a multi-synset in a multi-wordnet (Hauer and Kondrak, 2020b). The procedure is as follows: (1) translate the MWE in context; (2) word-align the source and target sentences; (3) lemmatize and POS-tag the source MWE; and (4) for each lemma in the MWE, search for a multi-synset that contains both the lemma and its translation. This method is unsupervised, and we refer to it as MT.

Our results provide evidence that leveraging lexical resources is beneficial for idiomaticity detection. In particular, our gloss-based method, when combined with a type-based UNATT heuristic, is among the top-scoring submissions in the one-shot setting. The heuristic is based on the observation that some MWEs are inherently idiomatic or literal, regardless of their context, which is confirmed by our analysis of the development set annotations.

2 Related Work

Early attempts to represent idiomatic MWEs involve treating idiomatic phrases as individual tokens and learning corresponding static embeddings (Mikolov et al., 2013). However, Cordeiro et al. (2016) show that the effectiveness of this method is limited by data sparsity for longer idiomatic expressions. Furthermore, Shwartz and Dagan (2019) and Garcia et al. (2021) conclude that idiomaticity is not yet accurately represented even by contextual embedding models. Tayyar Madabushi et al. (2021) create a new manually labeled dataset containing idiomatic and literal MWEs, and propose a method based on a pre-trained neural language model.

Regarding using lexical translations for idiomaticity detection, Moirón and Tiedemann (2006) measure semantic entropy in bitext alignment statistics, while Salehi et al. (2014) predict compositionality by presenting an unsupervised method that uses Wiktionary translation, synonyms, and definition information. We extend these ideas by applying machine translation, and consulting a multilingual lexical knowledge base.

Our prior work has already demonstrated the utility of lexical translations for various semantic tasks, including prior SemEval tasks on predicting cross-lingual entailment (Hauer et al., 2020) and contextual synonymy detection (Hauer et al., 2021), as well as word sense disambiguation (Luan et al., 2020), and homonymy detection (Hauer and Kondrak, 2020a; Habibi et al., 2021).

3 Methods

In this section, we describe our methods for idiomaticity detection.

3.1 Baseline mBERT

We re-implemented the mBERT classifier baseline (Devlin et al., 2019) following the methodology of Tayyar Madabushi et al. (2021). The model takes the context sentence and the relevant MWE as an input, and outputs a binary label indicating the idiomaticity of the target MWE. The input sequence is constructed by concatenating the MWE to the end of the context sentence after the special [SEP] token.

It is important to note the differences between our re-implementation and the official baseline provided by the task organizers. In the official baseline, the organizers add the target MWE as an additional feature in the one-shot setting but not in the zero-shot setting. Furthermore, the organizers include the sentences preceding and succeeding the target sentence only in the zero-shot setting. In our re-implementation, we add the target MWE and exclude the preceding and succeeding sentences in both zero-shot and one-shot settings.

3.2 Gloss-based Method

Our first method, defBERT, extends the baseline model by adding the glosses of all possible senses of each individual word in the target MWE to the classifier's input. The intuition is that the addition of the glosses to the input should help the classifier decide if the meaning of the target MWE can be deduced from the definitions of the individual words, i.e., if it is compositional. In the example in Figure 1, the disparity between the context in which *fish story* appears, and the glosses of the various senses of the words *fish* and *story* indicates that the MWE is idiomatic in this context.

The intuition for this method is that non-native speakers can identify idiomatic expressions, provided they understand the standard meanings of the words which comprise them. Suppose that the vocabulary of a non-native speaker covers most of the essential words necessary to understand a language, but not idiomatic expressions. Even if the speaker cannot deduce the meaning of an idiomatic expression in context, they can guess that the expression was used in an idiomatic sense because individual words of this expression do not make sense in the given context.

3.3 Translation-based Method

Our MT method is based on translating the target MWE in context, and leverages multilingual semantic resources. The intuition behind this method is that idioms are generally specific to a particular language, and, being non-compositional, their meanings cannot be conveyed simply by translating the individual words.

Under this hypothesis, to classify an MWE as literal or idiomatic, we need only determine whether the words in the MWE are translated literally. We do this by first identifying the translation of each word via alignment. We then consult a multilingual wordnet, or *multi-wordnet*, a lexical knowledge-base which organizes words in two or more languages into multilingual synonym sets, or *multi-synsets*. Each multi-synset corresponds to a unique concept, and contains the words which express that concept. Given a word in context, and a translation of that word in that context, we consider the word to be literally translated if it shares at least one multi-synset with its translation.

For example, consider an instance in which the MWE *wedding anniversary* is translated into Italian as *anniversario di matrimonio*. Our method checks if either of the translation pairs (*wedding*, *matrimonio*) and (*anniversary*, *anniversario*) share a multi-synset in a multi-wordnet. We test two versions of this method: in MT(all), this condition must be satisfied for all content words in the MWE; in MT(one), detecting a literal translation for one word is sufficient to classify the MWE as literal. In addition, multiple languages of translation may be considered.

3.4 Additional Heuristics

The annotation methodology for this shared task includes proper nouns in the literal class. We therefore use a part-of-speech tagger to detect proper nouns; if any word in the MWE is tagged as a proper noun, MT automatically classifies it as literal without further consideration.

In the one-shot setting, we also use a type-based heuristic which we refer to as UNATT. The intuition behind this heuristic is that certain MWEs are inherently idiomatic or literal, regardless of the context that they appear in. If the training data has no example of an MWE in a particular class, the heuristic exploits this fact as evidence that the MWE should always be classified as the opposite, attested class. For example, this heuristic always classifies *life vest* as idiomatic and *economic aid* as literal, as these are the only classes in which these MWEs appear in the training data. In practice, since UNATT returns no classification if the training set contains instances that belong to either class, this heuristic must be used in combination with another method.

3.5 Combination

Our defBERT and MT methods take different views of the data, with the former using a neural language model and gloss information, and the latter using translation and a lexical knowledge base. We therefore consider combining the two methods. In this approach, we independently apply defBERT and MT to a given instance. If the two methods agree, we return the agreed-upon classification; if they disagree, we return a default class, which is a tunable parameter. As with the other methods, we can combine this method with the UNATT heuristic in the one-shot setting.

4 Experiments

We now describe our experiments, including the tools and resources, the experimental setup, the results, and a discussion of our findings.

4.1 Lexical Resources

As lexical resources for sense translations and glosses, we use two different multi-wordnets: BabelNet (BN; Navigli and Ponzetto, 2010, 2012), and Open Multilingual WordNet (OMW; Bond and Foster, 2013). The defBERT method and the alignment tool access BN 4.0 via the provided Java API[2]. For the MT method, we access the BN 5.0 via the HTTP API. We access OMW via the NLTK interface (Bird et al., 2009). For the MT method, we consider the translation of a word to be literal if it shares a multi-synset with the word in either BN or OMW. For lemmatization and POS tagging, we use TreeTagger[3] (Schmid, 2013).

Both BN and OMW contain English glosses for most concepts, but the availability of glosses in other languages varies. In particular, OMW contains no Portuguese or Galician glosses. With BabelNet, we experimented with two techniques: using English glosses for all languages, and using glosses from the language of the instance, i.e. the

[2] https://babelnet.org/guide

[3] We use the pre-trained models for English, Portuguese, and Galician from https://cis.uni-muenchen.de/~schmid/tools/TreeTagger.

source language, when available. We refer to these variants as defBERT-BN-en and defBERT-BN-src, respectively. Since defBERT uses a multilingual pre-trained language model, it can seamlessly handle input from multiple languages. Furthermore, because of the relatively poor coverage of Galician in the lexical resources (only 54% of glosses are available in this language), we attempt to leverage its close relationship to Portuguese by processing Galician as if it was Portuguese.

4.2 Translation and Word Alignment

We translate the context sentence of each MWE with Google Translate API[4]. We translated English instances into Italian, and Portuguese/Galician instances into English, because of the good coverage of these languages in our resources. We also conducted development experiments with translation into less related languages, as well as with combining translation information from multiple languages, but we observed no consistent improvements.

We align each input sentence with its translation using BabAlign (Luan et al., 2020), which consults BabelNet to refine the alignments generated by a base aligner, FastAlign (Dyer et al., 2013). To further improve the alignment quality, we augment the set of sentence-translation pairs with additional parallel data from the OpenSubtitles parallel corpus (Lison and Tiedemann, 2016). We note that the English-Galician bitext is less than 1% of the size of the other two bitexts.

4.3 mBERT and defBERT

We fine-tune the mBERT-based models using the binary classification objective on the labeled training dataset. In the zero-shot setting, the MWEs in the training data are disjoint from those in the development and test splits, while in the one-shot setting, all MWEs in the development and test splits have at least one example in the training data. In the zero-shot setting, we trained the models only on the zero-shot training set, while in the one-shot setting, we trained the models on both training sets. In particular, we fine-tuned the models for 20 epochs with a maximum sequence length of 256, a learning rate of 2e-5, and a per device batch size of 16, using the HuggingFace Transformers library.[5]

[4] https://cloud.google.com/translate

[5] https://huggingface.co

4.4 Development experiments

Table 1 contains the results of the following models: the official mBERT-based baseline (row 0) as reported by the shared task organizers, our re-implementation of the official baseline (row 1), three variants of defBERT method which is based on mBERT (rows 2-4), defBERT combined with the UNATT heuristic (row 5), and the MT method combined with defBERT (rows 6-7)[6]. For rows 1-5 we average the macro F1 score obtained over five runs with random initializations.

Our experiments with defBERT explored the impact of adding glosses to the mBERT model, including the source and language of the glosses. With English glosses retrieved from BabelNet, defBERT improves the total score over the mBERT model in the zero-shot setting, especially on Portuguese. The results also suggest that the English glosses may be preferable to glosses in the source language, a finding which could simplify work on lower-resourced languages, where glosses may not be available.

Combining the predictions of the mBERT-based models with the UNATT heuristic improves the one-shot F1 scores in all cases (row 5 vs. row 4).

The MT methods achieve the best results when combined with defBERT on the development set in the zero-shot setting: MT(one) for English (row 6), and MT(all) for Portuguese (row 7). This demonstrates the utility of using lexical translation information for idiomaticity detection when annotated training data is not available.

4.5 Error Analysis

We found that the defBERT method performs slightly better, by about 1% F1, on literal instances as compared to idiomatic instances in the one-shot setting. In other words, the method is less likely to make an error when given a literal instance. We speculate that this is explained by the model's consistent classification of proper nouns as literal expressions. Indeed, a proper noun is identified incorrectly in only one instance. The fraction of idiomatic vs. literal instances is 39% in English and 56% in Portuguese.

For the MT method, a large number of of errors were caused by a literal translation of an idiomatic expression by Google Translate, even though the

[6] After the test output submission deadline, we discovered errors in our implementation of the MT methods. We report our original results for consistency with the official results.

		Development results				Test results							
		Zero-Shot		One-Shot		Zero-Shot				One-Shot			
		EN	PT	EN	PT	EN	PT	GL	ALL	EN	PT	GL	ALL
0	Baseline	66.2	63.9	87.0	86.7	70.7	68.0	50.7	65.4	88.6	86.4	81.6	86.5
1	mBERT	74.6	62.5	85.7	85.9	**75.1**	63.3	**61.1**	68.2	90.0	83.6	86.6	87.7
2	defBERT-BN-src	75.5	64.8	85.4	86.7	72.0	66.4	57.8	67.2	**95.7**	88.5	88.9	92.2
3	defBERT-BN-en	75.3	66.4	87.6	86.6	73.4	**68.4**	59.7	**69.5**	95.0	**89.3**	87.9	91.8
4	defBERT-OMW-en	74.8	64.5	87.1	84.5	71.0	65.6	56.5	66.5	92.4	86.7	88.5	90.1
5	UNATT + defBERT	-	-	**92.0**	**87.7**	-	-	-	-	94.5	89.2	**91.2**	**92.4**
6	MT(one) + defBERT	**77.3**	64.9	84.5	78.0	68.2	54.6	56.3	62.7	85.9	70.6	78.2	80.6
7	MT(all) + defBERT	66.4	**69.2**	73.7	78.0	65.4	62.5	54.3	62.1	80.3	73.8	73.9	77.3

Table 1: The macro F1 scores on the development and test datasets. Our official submissions are in rows 4-7. Where not otherwise specified, defBERT is in the OMW-en configuration.

corresponding expression is not meaningful in the target language. For example, "she was different, like a closed book" is translated into Italian as "era diversa, come un libro chiuso" even though the Italian translation does not carry the meaning of a person being secretive. In a few cases, the translation would simply copy the source language expression, yielding output which is not fully translated. In addition, some correct lexical translations are not in our lexical resources. Finally, a number of incorrect idiomatic predictions could be traced to word alignment errors, especially in cases of many-to-one alignments (e.g., *bow tie* correctly translated as *papillon*).

Manual analysis performed on the development set corroborates our hypothesis that most multiword expressions are inherently idiomatic (e.g., *home run*) or literal (e.g., *insurance company*). Only about one-third of the expressions are ambiguous in the sense that they can be classified as either class depending on the context (e.g. *closed book*). Our judgements are generally corroborated by the gold labels, with the exception of proper nouns, which are consistently marked as literal. The UNATT heuristic (Section 3.4), which is based on this observation, obtains a remarkable 98.3% precision and 55.8% recall on the set of 739 instances in the development set.

4.6 Test set results

The results on the test set are shown in Table 1. Our best results are produced by defBERT-BN-en in the zero-shot setting, and the combination of defBERT with the UNATT heuristic in the one-shot setting. The latter also obtains the best result on Galician, which demonstrates its applicability to low-resource languages, as this method only requires English glosses.

The results of combining defBERT with MT are well below the baseline, which may be due to a different balance of classes in the test set, omissions in lexical resources, and/or errors in our initial implementation. Another possible reason is that modern idiomatic expressions are often translated word-for-word ("calqued"), especially from English into other European languages. Examples from the development set include *flower child, banana republic*, and *sex bomb*.

5 Conclusion

Our top result ranks third overall in the one-shot setting. The corresponding method is applicable to a wide variety of languages. It takes advantage of the ability of neural language models to seamlessly incorporate textual information such as glosses, even if it is expressed in a different language. These results strongly support our hypothesis that the gloss information of individual words can improve idiomaticity detection. Moreover, our development results support the hypothesis that non-compositional expressions can be identified through their translations. These findings conform with our prior work on leveraging translation for various semantic tasks (Section 2). We hope that this work will motivate further investigation into the role of multilinguality in semantics.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

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, Sofia, Bulgaria.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997.

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.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

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

Amir Ahmad Habibi, Bradley Hauer, and Grzegorz Kondrak. 2021. Homonymy and polysemy detection with multilingual information. In *Proceedings of the 11th Global Wordnet Conference*, pages 26–35.

Bradley Hauer, Hongchang Bao, Arnob Mallik, and Grzegorz Kondrak. 2021. UAlberta at SemEval-2021 task 2: Determining sense synonymy via translations. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 763–770, Online.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269, Barcelona (online).

Bradley Hauer and Grzegorz Kondrak. 2020a. One homonym per translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7895–7902.

Bradley Hauer and Grzegorz Kondrak. 2020b. Synonymy = translational equivalence. *arXiv preprint arXiv:2004.13886*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016*, pages 923–929. European Language Resources Association.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Begona Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

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.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, Doha, Qatar.

Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

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.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*.

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.