Cutting Idioms Both Ways: Detecting Literal vs. Idiomatic Instances in PIEs

Andy Polizzotto
adp527@nyu.edu

Argy Sakti
das9669@nyu.edu

Ethan Zarov
etz211@nyu.edu

Abstract

In this paper, we describe our research into the efficacy of various procedures for identifying idioms within a corpus of natural language. Idiomatic expressions present complex challenges seen little elsewhere in language, offering an array of challenges and research opportunities surrounding their identification and classification. Starting from a naive system of pattern-matching text to one of more than 8,000 idiomatic expressions in our corpus, we explore and compare the magnitudes of improvement of several different modifications to our baseline model. Our experiments demonstrate that the two main modifications made–categorial variation and word vectorization– both provide significant increases to one metric from our baseline. When combined, the two systems complement each other naturally, each addressing a core issue surrounding idiom identification. Our final model achieved an F-score of 59.62, far exceeding our baseline.

1 Introduction

Idiomatic phrases are an abundant source of special cases within natural language. On a semantic level, many idioms require high-level contextual understanding of the surrounding text. While *missing the boat* idiomatically implies losing out on some great opportunity, with ample context it may also be applied as a literal phrase (or even a literal and idiomatic phrase at once). Furthermore, idiomatic phrases at a morphosyntactic level present widely different degrees of mobility, even provided that two idioms possess identical formation and parts of speech. For example, *the stage was set* is an appropriate variation of *[someone] set the stage*, whereas *the dust was bitten* cannot be used in lieu of *[someone] bites the dust*. This mobility combined with idiomatic ambiguity makes detecting idioms within natural language a significant challenge.

This paper describes our experiments into various methods to improve idiom identification, and our work in tuning the variables involved in these methods. Our experimentation and tuning processes not only provided significantly more accurate results than our baseline, but also may help others understand the key problems to tackle when detecting idiomaticity. Furthermore, our final results can provide some insight into the general concepts that comprise idiomatic language.

Primarily, we evaluate the efficacy of *categorial variation* and *word vectorization*, both on their own and paired together. These two systems can rather neatly be organized into each solving one key problem in idiom identification. Categorial variation expands our base list of known idiomatic expressions with variants containing different conjugations or pronouns, addressing the mobility of many multi-word, transitive expressions. On its own, categorial variation only serves to expand a model's dictionary of idioms, and cannot distinguish when an individual phrase carries a literal or idiomatic meaning. Word vectorization, on the other hand, is our solution to the semantic ambiguity of idioms, presenting one possible approach to providing context to an idiom identification model. We set out to see how these two opposing systems may complement each other, and document our logic and methodology during the experimentation process to assist future research in the field.

1.1 MAGPIE

MAGPIE is a corpus comprised of more than 50,000 instances of *Potentially Idiomatic Expressions* (PIEs), gathered through careful and strict crowd-sourcing (Haagsma et al., 2020). Each entry provides ample context surrounding each PIE, as well as useful information regarding document genre and categorization. Most importantly, MAGPIE provides a confidence score with every idiom.

A confidence score of 1.0 indicates unanimous agreement by their trusted crowd-sourcing in regards to a phrase's idiomaticity, while more ambiguous examples have lower scores.

MAGPIE also provides a filtered version, containing only PIEs (along with their surrounding context) that earn high confidence scores (> 0.75). One instance of a low confidence score within MAGPIE was context surrounding *on the hop*, a phrase that while typically implies unpreparedness, given its surrounding context was referring to a 'world [brewing] expert on hops.' Removing many false positives and ambiguous results, using portions of the filtered corpus proved useful in developing our system, allowing for experiments within the model to be developed in a more controlled environment before conducting tests on the unfiltered data set.

1.2 Idiomaticity and PIEs

Idioms are an additionally fascinating subject because while widely-known cases (*kick the bucket* or *raining cats and dogs*) are universally declared idiomatic, there is no agreed upon definition for classifying idioms (Zeng and Bhat, 2021; Fazly et al., 2009). One useful distinction provided by MAGPIE is the difference between PIEs and *idiomatic expressions*. Rather than a phrase like *see the light* unconditionally declared as an idiomatic expression, MAGPIE classifies phrases like it without context as PIEs. Only with context added, for instance in a sentence such as "After speaking with his local pastor for hours, he *saw the light*" does a phrase qualify as an *idiomatic expression*. While this ambiguity doesn't allow for a concrete, binary distinction between phrases being literal or idiomatic in all cases, classifying phrases as PIEs is more accurate to the fuzziness of everyday idiom usage and is important to understanding the way we come to learn and create new idiomatic expressions.

1.3 Contributions

This paper describes our experiments and research into the efficacy of various idiom and PIE identification methods. Our models use various implementations of *categorial variation* and *word vectorization*, studying the affects of gradually applied subsystems to our results (e.g., categorial variation on verbs versus nouns). Many modifications made to these two systems and the methods used to arrive at them may be useful in conducting larger scale experiments in idiom-related studies. Our hope is that the research conducted can be used as a touchstone for more advanced models in the future, and perhaps provide some further insight into the broad themes of idiomatic language.

This paper is organized as follows. Section 2 presents former research done on idiom identification and classification. In Section 3, we describe some of the problems that MAGPIE's data poses to our research and our solutions. In Section 4, we describe the methodology behind each attempted improvement to our baseline model. Section 5 presents the experiments and comparisons, including some issues identified in our models. Section 6 discusses potential future work to fix these issues and improve our results and Section 7 concludes our findings.

2 Related Works

Because of the rich research depth that idiomatic phrases provide both in cultural and linguistic value, there is no shortage of related research conducted in this field. But, given the diversity in definition regarding what constitutes idiomaticity, methodology and classification can vary widely. One important distinction between various models and their evaluation is *phrase classification* versus *token classification* (Muzny and Zettlemoyer, 2013). Many previous systems provide absolute phrase classification, where a given phrase is determined always idiomatic or literal. Token classification is far more accurate to the ambiguous nature of idioms in natural language, accounting for PIEs with more-often literal interpretations. Given the confidence scores and organization provided by MAGPIE, token classification– assigning literal or idiomatic to a phrase dependent on context– is the binary distinction we will use when evaluating our models.

Several previous models rely on dictionaries, idiom-definition pairs or lexical cohesion for token classification (Ehren, 2017; Verma and Vuppuluri, 2015). In recent work conducted by Zeng and Bhat, attention flow networks were used to determine what they coin as *semantic compatibility* between each PIE and its surrounding context, identifying how associated the words and meanings are within each data entry. PIEs with low semantic compatibility are likely to be figurative and can generally be declared idiomatic. Our use of word vectorization in our models aims to address this issue, however

notably done without the same high-level semantic understanding, definitions, and implementation required by these models.

3 Difficulties with the MAGPIE Corpus

Despite the MAGPIE corpus' best efforts to provide a complete, perfect data set of labelled PIEs, by nature of its structure and data collection process there are some misidentified or unlabelled PIEs that exist within the corpus. During our testing and analyzing, we found several categories of inaccurate PIEs, each with their own ways of muddying our final evaluations.

One such problem were some entries within the corpus sometimes containing more than one PIE. Regardless of whether these extra expressions would've had confidence scores high enough to be regarded idiomatic, many of our false positives during evaluation in the testing phase were attributed to entries with multiple PIEs, causing a deflated precision score. There were also many entries with PIEs deemed idiomatic that we would dispute as either literal interpretations– or in some cases, not even PIEs– within the confines of MAGPIE's definition for idiomatic expression. For example, entries like "A white *spot on* a yellow model..." and "Dissolve more [of the sugar] *in the hot water*" both had confidence scores exceeding the threshold despite being used literally or worded as invalid transitives of an actual PIE (not only is *in the hot water* literal, but its phrasing could never be used idiomatically).

Despite these shortcomings, the MAGPIE corpus still remains one of the best and most dependable sources for idiom and PIE-based research. To create a more accurate test set, we randomly selected PIEs until we arrived on 250 "clean" PIEs, that is, entries with correct labels and only a single idiom. Further discussed in Section 5, when comparing the same model's results on the unfiltered data and our manually filtered version applied to the same model, we found that precision increased from 18.30 to 19.56 and recall increased from 73.73 to 80.00.

4 Methodology

Our baseline for the experiments conducted naively pattern-matches each exact occurrence of an idiom using the Aho-Corasick (AC) algorithm. The AC algorithm searches through each PIE in the corpus, looking for an exact match from a list of over 8,000 unique idioms collected and organized from Wiktionary. Indeed, this initial model does little to address idioms with mobility or idioms that rely on context to be interpreted as non-literal.

4.1 Categorial Variation

The first improvement appended to our baseline was implementing *categorial variation* (Habash and Dorr, 2003) to add additional permutations of mobile and transitive PIEs to our idiom list. We use categorial variation (CatVar, or CV) to take commonly interchangeable words within each idiomatic phrase– generally appearing as verbs or pronouns, expanding our list of 8,000 phrases to more than 90,000. For example, *ace up one's sleeve* expands to also include *ace up [his/her] sleeve*, and *see red* conjugates to *seeing red* and *had seen red*. Some additional data processing was needed for these generated idioms to match the punctuation and style formatting of MAGPIE.

Of course, not all custom derivations generated by CatVar will remain idiomatic or even grammatical. *See red*, for example, is almost exclusively used in simple and continuous tenses, with past perfect and conditional conjugations such as *they had seen red* not seen in use idiomatically. Fortunately, since most newly-generated phrases that are agrammatical do not appear in typical speech or text, casting a net as wide as possible during this step will have negligible consequence to either our model's precision and recall. Still, generated phrases that remain grammatically correct but are no longer idiomatic as exemplified above can create a number of false positives. Given that most idioms involve some degree of non-compositionality, however, in constructing our model we believe the potential sacrifice to precision is trivial in the face of the vast improvements CatVar can offer to recall.

4.2 Word Vectorization

Another improvement we used in creating our models was applying *word vectorization* with the intent to address PIEs with more commonly ambiguous and literal uses. As discussed in the previous paragraph, detecting every phrase that matches a PIE as idiomatic unreservedly can cause a number of false positives. Word vectorization, implemented in our models using Word2Vec, was used to provide some of the same contextual understanding to our models that is often needed to interpret idioms in day-to-day conversation.

Our model uses cosine similarity scores to compare each meaningful word within a PIE to the neighboring words surrounding the phrase. For each word in the idiom, we find the two lowest cosine distances to the surrounding context and average them. This value, which we call *word-to-context distance* (WTC), is only calculated on word pairs with meaningful and relevant information, excluding calculations that would include a preset list of determiners, prepositions, and conjunctions that provide irrelevant context clues and wildly variable cosine distances. Below is the final equation used to determine a PIE's idiomaticity. Let *WTC[i]* be the word-to-context distance of word pair at index *i*, with *i=0* containing the word pair with the lowest WTC score.

$$(\sum_{i=0}^{n-1} WTC[i] * r^i)/(\sum_{i=0}^{n-1} r^i) > a$$

We apply geometric decay to our sorted list of WTC scores by falloff rate *r*, which in our final models had a value of 0.85. While it was important to include a falloff to our scores to favor stronger word pairs, low *r* values with steep falloffs could cause a key context clue to be missed. For example, in the PIE *missed the boat*, word pairs including *boat* are far more likely to inform a phrase's idiomaticity than pairs including *missed*, as *missed* plays a similar role in both the literal or idiomatic interpretation of the PIE. The denominator in our equation is used to normalize our final scores to between 0 and 1. Finally, we compare scores to alpha *a*, a threshold at which a PIE run through our model can be declared as idiomatic. Values *r* and *a* were arrived on through experimental evidence, running the model on the test data set and finding the values with the strongest F-score. These constants provided the best complements to CatVar's shortcomings, supplementing it with just enough contextual information to correctly identify literal instances of PIEs without overshooting and misidentifying idiomatic expressions.

5 Evaluation

We measure the efficacy of our models on a portion of the MAGPIE corpus, calculating precision, recall, and F-Score, given that each PIE tested is only deemed idiomatic if it has a confidence score of 0.75 or higher.

CatVar Model	Pre	Rec	F1
No Expansion	20.14	40.92	27.00
Verbs	19.51	53.07	28.53
Verbs+Nouns	20.03	53.73	29.18
Verbs+Custom Nouns	14.46	55.53	22.95

Table 1: Results of CatVar expansion applied at various degrees.

5.1 CatVar Performance

Table 2 provides the results of CatVar applied to the baseline (labelled here as "No Expansion"). Each type of CatVar expansion is tested incrementally to see its effects on our model. The main improvement of verb expansion–or essentially, conjugation–gave a sharp increase in recall with little loss in precision. The small loss is likely the cause of an increase in false positives due to the expanded word list. Upon investigation, most newly incorrect answers given by the model were primarily caused by PIEs with ambiguous confidence scores falling close to or below 0.75. Incorporating nouns gave a slight boost to all metrics, though the rarity of circumstances in which categorial variation is needed for these cases might make it irrelevant for future attempts at idiom identification. Indeed, when going further by adding our own custom list of CatVar expansions (for example, *someone* becoming *his/her*), we actually noticed a decrease in precision.

5.2 Word2Vec and Combined Performance

We first analyze the effect of various alpha thresholds (*a*) on our Word2Vec model, tuned to maximize each evaluation metric. Table 2 illustrates our findings, with each maximized metric underlined. Naturally, our model achieved the highest recall with a threshold at $a = .000$, as a PIE can never be misidentified as literal. This performance essentially serves as our baseline. Higher alpha values provided a dramatic increase in precision, though an even more dramatic decrease to our recall. The optimal threshold for balancing these two extremes and providing the highest F-score is $a = .599$, offering the greatest boost to precision with only a minimal reduction to the recall score.

Table 3 highlights the results of Word2Vec (W2V) on its own and paired with CatVar (CV). Both systems provide consistent improvements to the F-score when compared to our baseline. While CatVar mainly seeks to address the transitivity of

a	Pre	Rec	F1
.000	25.95	80.10	39.20
.897	59.70	7.41	13.19
.599	39.00	60.53	47.44

Table 2: Word2Vec alpha values (a) on the test data set tuned to provide the best performance in each metric.

Model	Pre	Rec	F1
Baseline	30.52	62.40	41.00
CV	27.92	88.80	42.49
W2V	54.37	54.80	54.58
CV+W2V	49.21	75.60	59.62

Table 3: Results of different system combinations.

idioms and increase recall scores, Word2Vec addresses non-compositionality and precision. When combined, these two systems complement each other naturally resulting in our best performing model for accurate and consistent idiom identification.

Other modern models tested on the MAGPIE corpus have achieved F-scores ranging between 82.73 and 87.78 (Zeng and Bhat, 2021). Notably, however, many of these models rely on semantic compatibility and require a deeper understanding of definitions in and around a PIE than our model uses.

5.3 Commonly Missed Idiom Types

When analyzing our results, several categories of commonly mislabelled PIEs became apparent. First, and most common, was a particular idiom within our MAGPIE test corpus not appearing in any form within our reference list compiled from Wiktionary. While certain idiomatic phrases like *on a silver platter* and *in the flesh* appear on Wiktionary, they are classified as English lemmas as opposed to idioms. Other phrases such as *like clockwork* are disputable as to whether they're idiomatic or similes regardless of context, accounting for unavoidable "error" in any system.

There were also some boundary issues stemming from variations of a particular idiom between MAGPIE and Wiktionary. *Point the finger at* was correctly identified by our model as idiomatic, but was declared invalid in evaluation as the MAGPIE corpus has the idiom's boundary defined as *point the finger*. However, there are other instances in which MAGPIE does include prepositions as part of an idiom. For example, *to the hilt* was identified by the model as idiomatic, but deemed incorrect in evaluation because MAGPIE bounds the idiom as *up to the hilt*. As such, a more refined test set that has a clear, set rule in regards to the beginning and end of an idiom would enable the model to be adjusted accordingly and likely result in greater precision during evaluation.

As discussed in Section 3, training on somewhat faulty data may have made our final models weaker than their potential. A perfectly annotated corpus could produce a much more reliable falloff rate and alpha threshold in our word vectorization research, resulting in better scores in our final evaluation. Regardless, these errors seem to occur across all idiom-related research, and will continue to occur for as long as idioms remain to be constantly evolving and nebulous phrases.

6 Future Work

More work could be done to refine our Word2Vec model in future papers, and expand its use cases to other realms of idiom-based research. For example, additional word-to-context distances could be calculated for the definitions of idioms. In addition to checking every relevant WTC score in a phrase like *seeing red*, WTC scores for relevant words in *getting angry* could also be factored into the final decision, adding to the overall semantic understanding of our model. Our Word2Vec model could also prove a valuable asset in identifying novel idioms. Using a high alpha threshold to prioritize precision could help in finding PIEs that do not exist within Wiktionary's database or our base list (Muzny and Zettlemoyer, 2013).

Our system and others would also benefit from further increasing the reliability of existing data sets, as a majority of unidentified idioms are a result of phrases not being labelled idioms within Wiktionary. We could also refine the scoring system to better reflect the true accuracy of our model at identifying idioms; this would mean modifying the precision and recall to account for cases where the boundaries do not exactly match. We could modify this system to measure the proportion of the model's output that correctly identifies a labelled idiom when calculating the precision, and the proportion of the labelled idiom that the model identifies when calculating the recall, so that the model is not punished as severely for missing or including an extra word.

Finally, we would like to investigate the usage of

transformers and an attention flow model to identify idioms as parts of speech that have a low degree of mutability when changed to semantically similar words (for example, *bites the dust* is grammatical while *bites the sand* is not), building off of previous work in using transformer-based models to fix grammatical errors (Alikaniotis and Raheja, 2019).

7 Conclusion

In this work, we studied how two main modifications to our baseline model– categorial variation and word vectorization– improve reliability of idiom identification. These solutions are lightweight and complementary, and put together far surpass baseline results, even staying near to state-of-the-art model findings despite our models' relative simplicity.

Results gathered from tuning our systems may also provide a broad-stroke understanding of idiomaticity. Our research with CatVar and various degrees of dictionary expansion can provide some insight into the types of idioms that have mobility and their pervasiveness across natural language. Additionally, our methodology for tuning our word vectorization can be built upon in later standalone models, or as a valuable extension to idiom identification or detection systems in the future.

References

Dimitris Alikaniotis and Vipul Raheja. 2019. The unreasonable effectiveness of transformer language models in grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–133, Florence, Italy. Association for Computational Linguistics.

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

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

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Nizar Habash and Bonnie Dorr. 2003. A Categorial Variation Database for English. In *Proceedings of the North American Association for Computational Linguistics (NAACL'03)*, pages 17–23, Edmonton, Canada.

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

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

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9(1):1546–1562.