Identification of Idiomatic Sequences

Aaron Shaw
College of Science and Engineering
Western Washington University
shawa8@students.wwu.edu

Abstract

I implement and explore a possible method for detecting idiomatic phrases. My method determines how strongly a phrase is linked to its surrounding context. If it determines the phrase has strong cohesive ties to the rest of the text, the phrase is deemed literal. If not, the phrase is labeled idiomatic. Through exploring this method I determine that this approach does merit a level of success, and can prove a useful tool in text classification.

1 Introduction

Idioms are a frequent occurrence in all languages. Defying traditional lexical composition, an idiomatic phrase's meaning is not derived from the individual meanings of its parts, but rather has a culturally-assigned arbitrary definition. For example, the phrase, ”to kick the bucket” is an English idiom for ”to die”. A simple solution for addressing these corner cases in natural language processing would be to construct a rules-based system that tokenizes idiomatic phrases and assigns them their arbitrary idiomatic phrase.

Problems arise, however, when an idiomatic phrase is meant literally. A natural language processor should be able to distinguish between the idiomatic and the literal, to avoid inaccurate classification. In this paper, I demonstrate a possible approach based on contextual lexical cohesion.

2 Related Work

This particular approach is primarily based on the study conducted by Sporleder and Li(2009) which explores lexical chains and cohesion networks as a tool for identification of non-literal phrases. However, there are several proposed approaches to idiom classification. Most methods of identifying idioms take advantage of certain properties that idioms share. One such property is that most idioms fall under the classification of formulaic language. Formulaic language is a term for expressions that have rigidity in their form. The phrase, ”to kick the bucket” will rarely, if ever, show up in the passive voice, ”the bucket was kicked”. This trait of steadfast structure and lack of deviation can be used for identification. The other major trait of idioms is their lack of compositional meaning, which this project is exploiting. An example of such an approach is the study by Katz and Giesbrecht(2006), which explores the connection of contextual similarity to non-compositionality of multi-word expressions.

3 Using Contextual Lexical Cohesion to Classify Idiomatic Sequences

3.1 Lexical Cohesion

This project uses lexical cohesion to detect idiomatic phrases. Lexical cohesion is the intuition that sentences in a text will usually contain words that are conceptually related to one another, forming what is called a lexical chain. Typically, a series of sentences from one English-Language text will have a higher degree of similarity to each other than they would have to a sentence from another text. This is because, typically, sequences of sentences in a text will share some level of similar context. This inter-sentence contextual similarity is what I use to identify phrases as idiomatic or literal. Since the figurative meaning of an idiom is arbitrary in respect to the individual literal meanings of the words comprising it, only one of these two meanings will apply to a given context. If a phrase is meant literally, then the component pieces of the phrase should show a high degree of cohesion to the context it is found in, and if the phrase is meant figuratively, then the figurative meaning will have a high degree of cohesion within the context of the piece, but the component pieces will not.

3.2 Modeling Text Cohesiveness

The intuition behind this approach is simple, but in application much is required. In order to apply this method, tools must be available which determine semantic relatedness between words. To this end, I used the WordNet database through the NLTK library. WordNet is a free lexical database of words and their semantic and conceptual links. for instance ”fire” is conceptually linked with ”burn”, and ”burn” is linked to ”pain”, and ”pain” to ”health”, and so on. With over 100,000 entries, this database covers most words that I expect to come across. With access to this, finding the relatedness between two words can be derived by computing the shortist conceptual path between two words. WordNet has this as a built-in function: WordNet.path_similarity(). WordNet.path_similarity() takes as input two WordNet nodes and returns a value between 0 and 1, derived from the shortest path that connects the two.

3.3 The Cohesion Graph

In order to determine how cohesive a series of word tokens in a text are, I designed a cohesion graph. Each vertex represents a word in the text and each vertex is connected by an edge. The edges connecting vertices on this graph correspond to the path similarity between the two words. By dividing the sum of the edge values by the number of edges, I have the value for the average path similarity between any two words in the text. The higher the value, the more cohesive (on average) the text. For each instance of a phrase I constructed two cohesion graphs; one graph containing the target phrase and one graph omitting it. The graph with the higher average edge weight I assume has a higher connectivity. If the graph containing the target phrase has a higher connectivity, then the classifier labeled the sample as ”literal”. If not, then it gave the label ”idiom”. The only library required in implementing this was nltk.

4 Evaluating the Cohesion Graph Approach

Section 4.2 gives details of the experiments and results. First, however, I will describe the data used in my experiments.

Expression	Literal	Non-Literal	All
break the ice	10	82	92
Toot one’s own Horn	0	16	16
Drop the Ball	5	42	47
Eat an Apple	33	0	33
Play with Fire	8	46	54
Bite more than one can Chew	1	14	15
Spill the beans	0	33	33
Get ones feet wet	10	15	25
All	67	248	315

Table 1: Idiom Statistics.

4.1 Data

I chose 9 phrases to test on. The idioms were selected more or less randomly, although I made a point of picking a few idioms that could feasibly show up both figuratively and literally. For each phrase, I extracted at least 15 samples from the Corpus of Contemporary American English and manually formatted and labeled them as idiomatic instances or figurative. Because COCA is only free to access through their online interface and there is a hard cap on the number of queries allowed for free per day, the actual amount of data I was able to process was quite small. Entries that were extracted were instances of the phrase which appeared in canonical form. For instance, ”broke the ice” and ”break the ice” would be extracted, but not ”break a ice”. The only variation in the phrasing was found in the head verb of each phrase. For each example, I included for context the sentence containing the phrase, the 4 sentences following the phrase, and the 4 sentences preceding. I hand-annotated each instance as idiomatic or literal. One of the phrases ”eat an apple” was used to determine the accuracy of the cohesion=graph method in detecting literal phrases, which should never be reported as idiomatic.

4.2 Experimental Results

The success of this method could not be solidly determined. With accuracies ranging from 33% to 95%,the method’s performance appeared to wildly change based on which phrase was being evaluated. A few things can be determined with this data set, however. The cohesion graph model has a precision of approximately 85% and a recall of 59%, giving it an F_1-score of .70.

The most notable outlier in these results would be the phrase ”play with fire”. The reason for this

	Guessed	"Idiom"	Guessed	"Literal"	
Sample Set	**Correct**	**Incorrect**	**Correct**	**Incorrect**	%**Correct**
breakicetxt	60	3	7	22	73%
toothorn.txt	10	0	0	6	62%
dropball.txt	23	3	2	19	53%
eatapple.txt	0	9	24	0	73%
playfire.txt	10	0	8	36	33%
bitechew.txt	13	1	0	1	87%
spillbeans.txt	16	0	0	17	48%
feetwet.txt	15	10	0	0	60%
All	147	26	41	101	60%

Table 2: table

is likely due to the usage of the idiom. In many instances, the phrase was "completed" with "If you play with fire, you're going to get burned" the instance of "burned has a strong lexical connectivity to fire, and thus would support the incorrect conclusion that the instance was to be taken literally. Indeed this was observed for several samples. Many times, writers appear to use an idiom as a basis for analogy, and decorate the surrounding context with semantically related words. The cohesion graph approach used on the phrase, "eat an apple" reported "literal" a majority of the time. In the instances where "idiom" was reported, the phrase was isolated from context, serving as minor decoration to the text. By removing minor details, the cohesiveness of a piece is also improved, and the graph approach deems the phrase idiomatic.

5 Conclusion

In this paper, I demonstrated a possible approach to identifying abnormal phrases that show up in language. My approach was based on the assumption that non-literal phrases will typically not show strong ties to their surrounding contexts. My results were encouraging and indicate I may be on to something, but my method was far from perfect. In future work, I would like to do a number of things.

Firstly, I need to procure a larger dataset. With only 315 samples to work with, my test data was tiny. I had originally sought to test 20 commonly-occurring idioms, but soon realized this was out of my capabilities. With more data, I could achieve more reliable results, and identify if the trends I see are correct.

Secondly, I am currently not taking into account homonyms. The noun "horn", for instance, has two distinct meanings. Currently my program defaults to the most commonly-occurring instance of that word for that part of speech. If I refine my program to correctly identify the intended meaning, I could greatly improve my results, as overall cohesiveness, I believe, would increase.

Another way in which I could improve this method is change the way word similarity is computed. Although working well enough, the path_similarity() function is only a rough estimate of an intangible thing, If my lexical connectedness quantifier is increased in accuracy, so are my results.

The amount of text before and after a phrase which I used for context was chosen because that was the maximum amount of context the Corpus of Contemporary American English would give me. I believe I could play around with different lengths of context to determine the optimal distance to get the most context without losing cohesion

Another possible way of improving my data would be to explore the application tf lexical chains. Like my graph approach, lexical chains are series of words that are contextually connected. I generate lexical chains in a text and determine if the target phrase is contained in any of them. If so, I can use this information to weigh the likelihood of its literal or non-literal usage.

Ultimately, the cohesion graph can be used as a reasonable determiner of idioms. However, my results are still far, far lower than other approaches towards idiom identification. Even with the improvements I've listed above, this model simply will never identify an idiomatic phrase which happens to make sense both figuratively and literally. A possible solution to this may be to test a graph with the target phrase replaced with the figurative meaning. For example, test the data with ”kick” and ”bucket” removed and ”die” inserted. If the figurative meaning makes for a stronger cohesiveness, then that is a strong indicator towards idiomatic usage.

Perhaps the most intriguing aspect of this project, however, is the non-idiom applications of this method. Although working plenty fine in identifying idioms, what this method truly measures is how relevant a collection of words are to a given context. With minimal modifications, I could change this method to search a collection of text and select the ”most relevant” or ”least relevant” sentence. This has implications for. among other things, text summarization.

References

Sporleder, Caroline and Linlin Li. 2009. *”Unsupervised Recognition of Literal and Non-literal Use of Idiomatic Expressions.” Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on* - EACL ’09 (2009)

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics.*

Graham Katz and Eugenie Giesbrecht 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 1219.