Idioms: Humans or machines, it's all about context
PREPRINT VERSION

Manali Pradhan, Jing Peng, Anna Feldman, Bianca Wright

Montclair State University
Department of Linguistics
Department of Computer Science
Montclair, NJ, 07043
USA

Abstract. Expressions can be ambiguous between idiomatic and literal interpretation depending on the context they occur in ("sales hit the roof" vs "hit the roof of the car"). Previous studies suggest that idiomaticity is not a binary property, but rather a continuum or the so-called "scalar phenomenon" ranging from completely literal to highly idiomatic. This paper reports the results of an experiment in which human annotators rank idiomatic expressions in context on a scale from 1 (literal) to 4 (highly idiomatic). Our experiment supports the hypothesis that idioms fall on a continuum and that one might differentiate between highly idiomatic, mildly idiomatic and weakly idiomatic expressions. In addition, we measure the relative idiomaticity of 11 idiomatic types and compute the correlation between the relative idiomaticity of an expression and the performance of various automatic models for idiom detection. We show that our model, based on the distributional semantics ideas, not only outperforms the previous models, but also positively correlates with the human judgements, which suggests that we are moving in the right direction toward automatic idiom detection.

1 Introduction

Philip Johnson-Laird once said: "If natural language had been designed by a logician, idioms would not exist" [1]. According to [2], there are as many fixed expressions as there are words in American English, roughly 80,000. This means that people have at least 160,000 items memorized and available for use. What sets idioms from most other fixed expressions is the absence of any observable relation between their linguistic meaning and their idiomatic interpretation [1]. Researchers have not come up with a single agreed-upon definition of idioms that covers all members of this class [3–8]. The common property ascribed to the idiom is its relative non-compositionality. Additional properties include lexical and syntactic flexibility, i.e., *kick the bucket* is not the same thing as *kick the pail* and *the bucket was kicked* does not preserve the idiomatic meaning.

According to [9], the study of the identification and comprehension of ambiguous idiomatic expressions, like *sales* ***hit the roof*** vs. ***hit the roof*** *of the car*, shares many of the issues that are involved in the study of lexical ambiguity. One of the most important components involved in the comprehension of idioms is context ([9]). In particular, when ambiguous idioms are involved, local context seems to contribute to the selection of the particular sense of an idiom.

In this paper, we describe an experiment in which we use Amazon.com's Mechanical Turk (MTurk) to gather human subject rankings on 165 idiomatic expressions, from sixty raters. The purpose of the experiment was to determine whether subjects could rank idioms on a scale and whether the human rankings correlate with the performance of our automatic idiom classifier.

2 MTurk Experiment

2.1 Data

In both of the automatic classification and human judgement experiments, we use the VNC-Tokens dataset developed by [10], a resource of almost 3,000 English verb-noun combination (VNC) usages annotated as to whether they are literal or idiomatic. We selected expressions that were both idiomatic and ambiguous between idiomatic and literal interpretations. [10] report that in their analysis of 60 VNCs, approximately half of these expressions frequently appear in their literal sense in the British National Corpus (BNC) [11]. The original VNC-Tokens list was created by two annotators, both native English-speakers. According to [10], the annotators were presented with the single sentence containing the VNC usage. Sentences in the surrounding context were not included. If the annotator was unable to determine the class of a token based on the sentence in which it occurs, he or she could choose the unknown label.

An important observation that [10] make which is subsequently supported by [12] is that the idiomaticity of an expression is not binary. Expressions may be more or less idiomatic, falling on a continuum ranging from completely literal expressions to semantically opaque. While we do not agree that expressions which are completely literal can be still called idioms (perhaps, the authors meant "collocational continuum"), we do think that idiomaticity is a scalar property, and this observation is used in the experiments described below. [10] also notice that at the adjudication step, when the annotators were supposed to discuss the tokens on which the judges originally disagreed to achieve a consensus annotation, among the issues that arose were the expressions that fall in the middle of the literal-idiomatic continuum. For example, [10] mention that the idiomatic expression *have a word* is related to the literal meaning, as in *At the moment they only had the word of Nicola's husband for what had happened.*

[10] divide their data into three sets: development, test, and skewed. Skewed contains expressions for which one of the literal or idiomatic meanings is infrequent, while the expressions in the development and test sets are more balanced across the senses. [10] notice that while the observed agreement for all the sets

is quite high (in the 80s), the kappa scores are low on the test and the skewed sets. [10] mention that the judges consistently disagreed on the label for *have words*, *hold fire*, and *make hit*. Eliminating these three expressions improves the unweighted Kappa score significantly. We address this issue in this paper as well.

2.2 Procedure

In our experiment, we wanted to see whether human annotators are capable of ranking the idioms on a scale and later correlate their judgement with the performance of our algorithm. Using Turktools [13], we randomized and formatted the target material into an html template compatible with Mechanical Turk. The 165 target items were split into three separate Mechanical Turk Human Intelligence Tasks (HITs), each of which contained five target idiomatic expressions presented in context, from all eleven idiom types. Each target item was to be assigned a ranking ranging from 1 to 4, or "not idiomatic" to "highly idiomatic". The purpose of rankings 2 and 3, was to allow for the possibility of an idiomatic expression to be perceived as neither strictly literal nor idiomatic. Prior to the beginning of the experiment, participants were presented with four example questions to aid in understanding the four possible rankings. In order to ensure the turkers were paying attention, each example question had instructions to select a specific ranking. Participant responses were primarily rejected if they consisted of numerous missing entries or an abnormally large number of low rankings. To increase the likelihood of participants being the native speakers of English, we required that all turkers had a high school diploma obtained in the US [1]
Here is an HIT excerpt:

> Instructions
> You will be presented with 55 text excerpts which contain various focus phrases (highlighted in bold). Your task is to rate how idiomatic each phrase is in its respective text excerpt. The contexts in which the phrases appear, will determine the degree of idiomaticity. There is no "correct" response, simply follow your native speaker intuitions. Below are some typical properties of idiomatic and literal phrases:
>
> Idiomatic phrases tend to be:
> - Abstract/complex
> - Vague
> - Commonly used by native speakers in casual speech and difficult for English learners
>
> Literal phrases tend to be:
> - Straightforward in meaning

[1] Naturally, this requirement does not guarantee a native speaker, but we had not a better option to control for it.

- Basic

Here is a scale you can follow. Not idiomatic (1)[2]: The meaning sounds fairly straightforward. Little idiomatic (2): The meaning seems like it could be taken literally, but not completely. It almost seems literal (or not idiomatic) but there's a hint of figurativeness. Somewhat idiomatic (3): The meaning seems to be figurative, but not completely. It almost seems idiomatic, but there is a hint of literalness. Idiomatic (4): It is figurative and cannot be taken literally. So, subjects were supposed to rank idiomatic expressions in context using the scale above. The context is a paragraph from BNC in which an idiomatic expression occurs. It's exactly the same context our automatic classifier uses to tell apart idioms from literal expressions. Here's an example:
We decided to go out to dinner the other day, but I was a little worried because I wasn't sure if she was still mad at me or not. So whatever, we still went and we got into the same argument we had last week. She ended up ***making a huge scene*** *right there, in the middle of the restaurant!*

2.3 Results

The results show that a ranking of "4", or "highly idiomatic", was the most frequent among all sixty raters, while the average ranking was 3.2. Although all of the paragraphs presented to participants consisted of strictly idiomatic expressions, lower rankings were assigned consistently by all raters across the 165 target items. The ratings for each idiom type show that some expressions received lower rankings than others. One expression (*have word*) in particular, received a very high assignment of low rankings in comparison to the others, resulting in an average ranking of 2.29. This result is consistent with what was reported in [10]. Apart from this expression, the other ten were assigned a ranking of "4" most frequently. The agreement among the raters was low in terms of both the unweighted B-statistic [14] (0.31) and Cohen's Kappa [15](0.06). The weighted measures were 0.70 for the B-statistic and 0.11 for the Cohen's Kappa.

Table 2.3 summarizes the experiment. Table 2.3 reports the average ranking per idiom type.

As has been mentioned in section 2.1, Cook et al. [10] report high observed agreement, but low kappa values on the data. They eliminate three expressions that the annotators consistently disagree on to improve the unweighted kappa score. Shankar and Baugdiwala [16] address the paradox earlier noticed by [17], namely, (1) low kappa values despite high observed agreement under highly symmetrically imbalanced marginals, and (2) higher kappa values for asymmetrical

[2] We should clarify here that even though all items that we used were already marked as idiomatic in the [10]'s data, we decided to keep the option of ranking them as literal, just in case of a mistake or a different interpretation. Remember that [10]'s dataset is annotated by only two annotators.

Table 1. Human Ranking Experiment

Number of subjects:	20 per experiment (60 total)
Each experiment:	50 tokens, across 11 idiom types
Total number of tokens tested:	165
Average ranking:	3.2
Most frequently used ranking:	4
Agreement:	
B-stat unweighted:	0.31
B-stat weighted:	0.70
Cohen's K unweighted:	0.06
Cohen's K weighted:	0.11

imbalanced marginal distributions. [16] examine the behavior of alpha, kappa and B-statistic [14] under different scenarios of marginal distributions, balanced or not, symmetrical or not. They show that while all statistics are affected by lack of symmetry and imbalances in the marginal totals, the B-statistic comes closest to resolving the paradoxes identified by [17]. Therefore, based on the B-statistic scores, we assume that the results of our human ranking experiment are reliable.

Table 2. Average human rankings of 11 idiom types

hold fire	3.28	hold horse	3.37	lose head	3.35
blow whistle	3.16	have word	**2.29**	make scene	3.02
give sack	3.33	take hear	3.30	blow top	**3.44**
hit wall	3.19	hit roof	3.34		

2.4 Related work

A similar experiment was conducted by [18]. They use a dataset with human judgements of compositionality [19] and ask the subjects to judge the compositionality of verb-noun combinations. The focus of their experiment is the detection of the more non-compositional verb-noun combinations, but they do not pay attention to the ambiguity of the expressions. Their list is largely idiomatic, whereas our experiment only deals with ambiguous expressions which can only be disambiguated in context. We are also aware of [20]'s dataset of 1048 noun-noun compounds annotated as non-compositional, compositional, conventionalized and not-conventionalized. The reason why we chose to work with verb-noun constructions is that we wanted to compare our algorithms with the state-of-the-art.

3 Automatic Approach

Our approach is based on two hypotheses: (1) words in a given text segment that are representatives of the local context are likely to associate strongly with a literal expression in the segment, in terms of projection of word vectors onto the vector representing the literal expression; (2) the context word distribution for a literal expression in word vector space will be different from the distribution for an idiomatic one (similarly to [21, 22]).

3.1 Projection Based On Local Context Representation

To address the first hypothesis, we propose to exploit recent advances in vector space representation to capture the difference between local contexts [23, 24].

A word can be represented by a vector of fixed dimensionality q that best predicts its surrounding words in a sentence or a document [23, 24]. Given such a vector representation, our first proposal is the following. Let v and n be the vectors corresponding to the verb and noun in a target verb-noun construction, as in *blow whistle*, where $v \in \Re^q$ represents *blow* and $n \in \Re^q$ represents *whistle*. Let $\sigma_{vn} = v + n \in \Re^q$. Thus, σ_{vn} is the word vector that represents the composition of verb v and noun n, and in our example, the composition of *blow* and *whistle*. As indicated in [24], word vectors obtained from deep learning neural net models exhibit linguistic regularities, such as additive compositionality. Therefore, σ_{vn} is justified to predict surrounding words of the composition of, say, *blow* and *whistle* in a literal context. Our hypothesis is that on average, the projection of v onto $\sigma_{blowwhistle}$, (i.e., $v \cdot \sigma_{blowwhistle}$, assuming that $\sigma_{blowwhistle}$ has unit length), where vs are context words in a literal usage, should be greater than $v \cdot \sigma_{blowwhistle}$, where vs are context words in an idiomatic usage.

For a given vocabulary of m words, represented by matrix

$$V = [v_1, v_2, \cdots, v_m] \in \Re^{q \times m},$$

we calculate the projection of each word v_i in the vocabulary onto σ_{vn}

$$P = V^t \sigma_{vn} \tag{1}$$

where $P \in \Re^m$, and t represents transpose. Here we assume that σ_{vn} is normalized to have unit length. Thus, $P_i = v_i^t \sigma_{vn}$ indicates how strongly word vector v_i is associated with σ_{vn}. This projection forms the basis for our proposed technique.

Let $D = \{d_1, d_2, \cdots, d_l\}$ be a set of l text segments (local contexts), each containing a target VNC (i.e., σ_{vn}). Instead of generating a term by document matrix, where each term is *tf-idf* (product of term frequency and inverse document frequency), we compute a term by document matrix $M_D \in \Re^{m \times l}$, where each term in the matrix is

$$p \cdot idf. \tag{2}$$

That is, the product of the projection of a word onto a target VNC and inverse document frequency. That is, the term frequency (tf) of a word is replaced by the projection of the word onto σ_{vn} (1). Note that if segment d_j does not contain word v_i, $M_D(i,j) = 0$, which is similar to *tf-idf* estimation. The motivation is that topical words are more likely to be well predicted by a literal VNC than by an idiomatic one. The assumption is that a word vector is learned in such a way that it best predicts its surrounding words in a sentence or a document [23, 24]. As a result, the words associated with a literal target will have larger projection onto a target σ_{vn}. On the other hand, the projections of words associated with an idiomatic target VNC onto σ_{vn} should have a smaller value.

We also propose a variant of $p \cdot idf$ representation. In this representation, each term is a product of p and typical *tf-idf*. That is,

$$p \cdot tf \cdot idf. \tag{3}$$

3.2 Local Context Distributions

Our second hypothesis states that words in a local context of a literal expression will have a different distribution from those in the context of an idiomatic one. We propose to capture local context distributions in terms of scatter matrices in a space spanned by word vectors [23, 24].

Let $d = (w_1, w_2 \cdots, w_k) \in \Re^{q \times k}$ be a segment (document) of k words, where $w_i \in \Re^q$ are represented by a vectors [23, 24]. Assuming w_is have been centered, we compute the scatter matrix

$$\Sigma = d^t d, \tag{4}$$

where Σ represents the local context distribution for a given target VNC.

Given two distributions represented by two scatter matrices Σ_1 and Σ_2, a number of measures can be used to compute the distance between Σ_1 and Σ_2, such as Choernoff and Bhattacharyya distances [25]. Both measures require the knowledge of matrix determinant. We propose to measure the difference between Σ_1 and Σ_2 using matrix norms. We have experimented with the Frobenius norm and the spectral norm. The Frobenius norm evaluates the difference between Σ_1 and Σ_2 when they act on a standard basis. The spectral norm, on the other hand, evaluates the difference when they act on the direction of maximal variance over the whole space.

4 Experiments

4.1 Methods

We carried out an empirical study evaluating the performance of the proposed techniques. The following methods are evaluated:

1. $p \cdot idf$: compute term by document matrix from training data with proposed $p \cdot idf$ weighting (2).
2. $p \cdot tf \cdot idf$: compute term by document matrix from training data with proposed p*tf-idf weighting (3).

3. $CoVAR_{Fro}$: proposed technique (4) described in Section 3.2, the distance between two matrices is computed using Frobenius norm.
4. $CoVAR_{Sp}$: proposed technique similar to $CoVAR_{Fro}$. However, the distance between two matrices is determined using the spectral norm.
5. *Context+ (CTX+)*: supervised version of the CONTEXT technique described in [26].
6. *GMM*: Gaussian Mixture Model as described in [27].

For methods **3** and **4**, we compute the literal and idiomatic scatter matrices from training data (4). For a test example, compute a scatter matrix according to (4), and calculate the distance between the test scatter matrix and training scatter matrices using the Frobenius norm for method **3**, and the spectral norm for method **4**. Method **5** corresponds to a supervised version of CONTEXT described in [26]. CONTEXT is unsupervised because it does not rely on the "gold-standard". Rather it uses knowledge about automatically acquired canonical forms (C-forms). Thus, the gold-standard is "noisy" in CONTEXT. Here we provide manually annotated training data. Therefore, CONTEXT+ is a supervised version of CONTEXT. For Method **6**, [27]'s work uses Normalized Google Distance to model semantic relatedness in computing features [28, 29]. We use inner product between word vectors. The main reason is that Google's custom search engine API is no longer free.

4.2 Data Preprocessing

We use BNC and a list of VNCs [30] (described above) and labeled as L (Literal), I (Idioms), or Q (Unknown). For our experiments we only use VNCs that are annotated as I or L. We only experimented with idioms that can have both literal and idiomatic interpretations. Each document contains three paragraphs: a paragraph with a target VNC, the preceding paragraph and following one. Our data is summarized in Table 3.

Since BNC did not contain enough examples, we extracted additional ones from COCA, COHA and GloWbE (http://corpus.byu.edu/). Two human annotators labeled this new dataset for idioms and literals. The inter-annotator agreement was relatively low (Cohen's kappa = .58); therefore, we merged the results keeping only those entries on which the two annotators agreed. For our experiments reported here, we obtained word vectors using the word2vec tool [23, 24] and the text8 corpus. The text8 corpus has more than 17 million words, which can be obtained from `mattmahoney.net/dc/text8.zip`. The resulting vocabulary has 71,290 words, each of which is represented by a $q = 200$ dimension vector. Thus, this 200 dimensional vector space provides a basis for our experiments.

4.3 Datasets

Table 3 describes the datasets we used to evaluate the performance of the proposed technique. All these verb-noun constructions are ambiguous between literal and idiomatic interpretations.

Table 3. Datasets: Is = idioms; Ls = literals

Expression	Train	Test
BlowWhistle	20 Is, 20 Ls	7 Is, 31 Ls
LoseHead	15 Is, 15 Ls	6 Is, 4 Ls
MakeScene	15 Is, 15 Ls	15 Is, 5 Ls
TakeHeart	15 Is, 15 Ls	46 Is, 5 Ls
BlowTop	20 Is, 20 Ls	8 Is, 13 Ls
GiveSack	20 Is, 20 Ls	26 Is, 36 Ls
HaveWord	30 Is, 30 Ls	37 Is, 40 Ls
HitRoof	50 Is, 50 Ls	42 is, 68 Ls
HitWall	90 Is, 90 Ls	87 is, 154 Ls
HoldFire	20 Is, 20 Ls	98 Is, 6 Ls
HoldHorse	80 Is, 80 Ls	162 Is, 79 Ls

5 Results

Table 4 shows the average precision, recall and accuracy of the competing methods on 11 datasets over 20 runs. (The average best performance is in bold face. We calculate accuracy by adding true positives and true negatives and normalizing the sum by the number of examples. The results show that the *CoVAR* model outperforms the rest of the models overall.

Interestingly, the Frobenius norm outperforms the spectral norm. One possible explanation is that the spectral norm evaluates the difference when two matrices act on the maximal variance direction, while the Frobenius norm evaluates on a standard basis. That is, Frobenius measures the difference along all basis vectors. On the other hand, the spectral norm evaluates changes in a particular direction. When the difference is a result of all basis directions, the Frobenius norm potentially provides a better measurement. The projection methods ($p \cdot idf$ and $p \cdot tf \cdot idf$) outperform $tf \cdot idf$ overall but not as pronounced as *CoVAR*.

Finally, we have noticed that even the best model ($CoVAR_{Fro}$) does not perform as well on certain idiomatic expressions. We hypothesize that the model works the best on highly idiomatic expressions.

6 Is there a correlation between the human judgements and the automatic approach?

We measure the correlation between the human judgements and the competing algorithms in terms of Pearson's correlation coefficient. Figure 1 shows the plots of the correlation matrices between the average human judgements per idiom type shown in Table 2.3 and the judgements by the algorithms. The resulting correlation matrices show that the performance of the proposed algorithm $CoVar_{Fro}$ is highly correlated with the human judgements, followed by $CoVar_{Sp}$. This once again demonstrates that $CoVar_{Fro}$ is capable of exploiting context

Table 4. Average accuracy of competing methods on 11 datasets: BlWh (BlowWhistle), LoHe (LoseHead), MaSe (MakeScene), TaHe (TakeHeart), BlTo (BlowTop), GiSa (GiveSack), HaWo (HaveWord), HiRo (HitRoof), HiWa (HitWall), HoFi (HoldFire), and HoHo (HoldHorse).

	BlWh	LoHe	MaSe	TaHe	BlTo	GiSa	HaWo	HiRo	HiWa	HoFi	HoHo	Ave
Precision												
$p \cdot idf$	0.29	0.49	0.82	0.9	0.59	0.55	0.52	0.54	0.55	0.97	0.86	0.64
$p \cdot tf \cdot idf$	0.23	0.31	0.4	0.78	0.54	0.54	0.53	0.41	0.39	0.95	0.84	0.54
$CoVAR_{Fro}$	0.65	0.6	0.84	0.95	0.81	0.63	0.58	0.61	0.59	0.97	0.86	**0.74**
$CoVAR_{sp}$	0.44	0.62	0.8	0.94	0.71	0.66	0.56	0.54	0.5	0.96	0.77	0.68
$CTX+$	0.17	0.55	0.78	0.92	0.66	0.67	0.53	0.55	0.92	0.97	0.93	0.70
GMM	0.18	0.46	0.67	0.79	0.41	0.45	0.42	0.4	0.41	0.94	0.73	0.53
Recall												
$p \cdot idf$	0.82	0.27	0.48	0.43	0.58	0.47	0.53	0.84	0.92	0.83	0.81	0.63
$p \cdot tf \cdot idf$	0.99	0.3	0.11	0.11	0.53	0.64	0.53	0.98	0.97	0.89	0.97	0.64
$CoVAR_{Fro}$	0.71	0.78	0.83	0.61	0.87	0.88	0.49	0.88	0.94	0.86	0.97	**0.80**
$CoVAR_{sp}$	0.77	0.81	0.82	0.55	0.79	0.75	0.53	0.85	0.95	0.87	0.85	0.78
$CTX+$	0.56	0.52	0.37	0.66	0.7	0.83	0.85	0.82	0.57	0.64	0.89	0.67
GMM	0.55	0.48	0.54	0.36	0.49	0.47	0.41	0.55	0.73	0.72	0.57	0.53
Accuracy												
$p \cdot idf$	0.6	0.48	0.53	0.44	0.68	0.62	0.54	0.66	0.7	0.81	0.78	0.62
$p \cdot tf \cdot idf$	0.37	0.49	0.33	0.18	0.65	0.55	0.53	0.45	0.43	0.85	0.86	0.52
$CoVAR_{Fro}$	0.87	0.58	0.75	0.62	0.86	0.72	0.58	0.74	0.74	0.84	0.87	**0.74**
$CoVAR_{sp}$	0.77	0.61	0.72	0.56	0.79	0.73	0.58	0.66	0.64	0.84	0.73	0.69
$CTX+$	0.4	0.46	0.45	0.64	0.75	0.76	0.57	0.67	0.71	0.64	0.88	0.63
GMM	0.46	0.5	0.52	0.39	0.49	0.53	0.49	0.51	0.53	0.7	0.57	0.52

information. Interestingly, the supervised version of the CONTEXT technique described in [26] negatively correlates with the human rankings, suggesting that this model does not use contextual information in the most optimal way.

6.1 Related Work

Previous approaches to idiom detection can be classified into two groups: 1) type-based extraction, i.e., detecting idioms at the type level, e.g., [6, 26, 31, 32]; 2) token-based detection, i.e., detecting idioms in context. Type-based extraction is based on the idea that idiomatic expressions exhibit certain linguistic properties such as non-compositionality that can distinguish them from literal expressions [6, 26]. While many idioms do have these properties, many idioms fall on the continuum from being compositional to being partly unanalyzable to completely non-compositional [33]. [22, 34, 26, 35, 27, 36–38], among others, notice that type-based approaches do not work on expressions that can be interpreted idiomatically or literally depending on the context and thus, an approach that considers tokens in context is more appropriate for idiom recognition. To address these problems, [39] investigate the bag of words *topic* representation

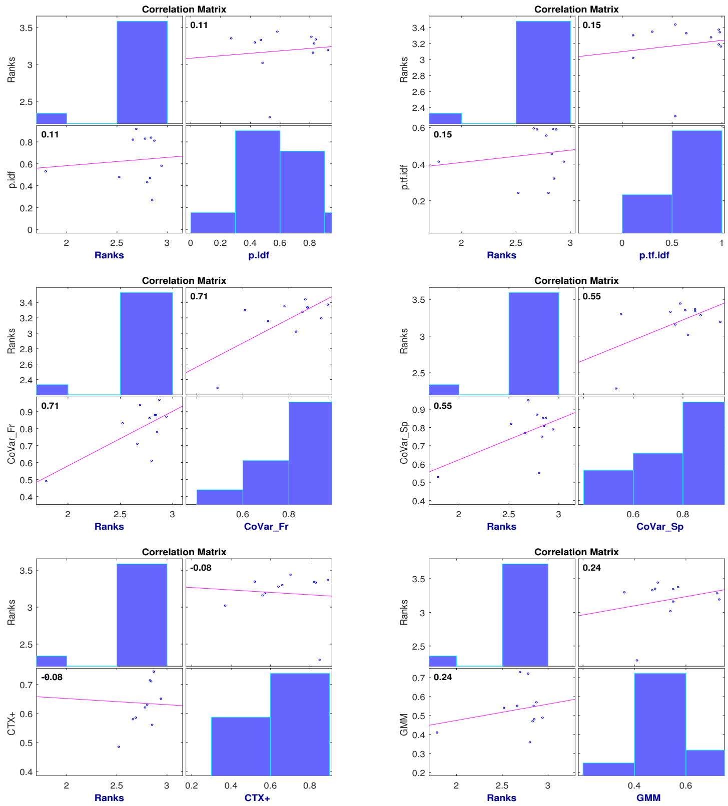

Fig. 1. Pairwise Pearson's correlation matrix between the human judgements and the competing algorithms. Top row: $p{\cdot}idf$ and $p{\cdot}tf{\cdot}idf$. Middle row: $CoVar_{Fro}$ and $CoVar_{Sp}$. Botton row: $CTX+$ and GMM.

and incorporate an additional hypothesis–contexts in which idioms occur are more affective. Still, they treat idioms as semantic outliers. [40–45] explore a range of distributional vector-space models for semantic composition.

7 Conclusions

In this paper we reported the results of an experiment in which human annotators ranked idiomatic expressions in context on a scale from 1 (literal) to 4 (highly idiomatic). Our experiment supports the hypothesis that idioms fall on a continuum and that one might differentiate between highly idiomatic, mildly idiomatic and weakly idiomatic expressions. In addition, we measured the relative idiomaticity of 11 idiomatic types and computed the correlation between the relative idiomaticity of an expression and the performance of various automatic models for idiom detection. We have shown that our model, based on the distributional semantics ideas, positively correlates with the human judgements. This suggests that we are moving in the right direction toward automatic idiom detection.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-16-1-0261 and a National Science Foundation grant IIS-1319846.

References

1. Glucksberg, S.: From Metaphors to Idioms. Number 36 in Oxford Psychology Series. Oxford University Press (2001)
2. Jackendoff, R.: The boundaries of the lexicon. Idioms: Structural and psychological perspectives (1995) 133–165
3. Glucksberg, S.: Idiom Meanings and Allusional Content. In Cacciari, C., Tabossi, P., eds.: Idioms: Processing, Structure, and Interpretation. Lawrence Erlbaum Associates (1993) 3–26
4. Cacciari, C.: The Place of Idioms in a Literal and Metaphorical World. In Cacciari, C., Tabossi, P., eds.: Idioms: Processing, Structure, and Interpretation. Lawrence Erlbaum Associates (1993) 27–53
5. Nunberg, G., Sag, I.A., Wasow, T.: Idioms. Language **70** (1994) 491–538
6. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A Pain in the Neck for NLP. In: Proceedings of the 3rd International Conference on Intelligence Text Processing and Computational Linguistics (CICLing 2002), Mexico City, Mexico (2002) 1–15
7. Villavicencio, A., Copestake, A., Waldron, B., Lambeau, F.: Lexical Encoding of MWEs. In: Proceedings of the Second ACL Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain (2004) 80–87
8. Fellbaum, C., Geyken, A., Herold, A., Koerner, F., Neumann, G.: Corpus-based Studies of German Idioms and Light Verbs. International Journal of Lexicography **19** (2006) 349–360
9. Colombo, L.: The comprehension of ambiguous idioms in context. Idioms: Processing, structure, and interpretation (1993) 163–200
10. Cook, P., Fazly, A., Stevenson, S.: The vnc-tokens dataset. In: Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008). (2008) 19–22
11. Burnard, L.: The British National Corpus Users Reference Guide. Oxford University Computing Services. (2000)
12. Wulff, S.: Rethinking idiomaticity: A usage-based approach. A&C Black (2010)
13. Erlewine, M.Y., Kotek, H.: A streamlined approach to online linguistic surveys. Natural Language & Linguistic Theory **34** (2016) 481–495
14. Bangdiwala, S.: A graphical test for observer agreement. In: 45th International Statistical Institute Meeting. (1985) 307–308
15. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Education and Psychological Measurement (1960) 37–46

16. Shankar, V., Bangdiwala, S.I.: Observer agreement paradoxes in 2x2 tables: comparison of agreement measures. BMC medical research methodology **14** (2014) 100
17. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of clinical epidemiology **43** (1990) 543–549
18. McCarthy, D., Keller, B., Carroll, J.: Detecting a Continuum of Compositionality in Phrasal Verbs. In: Proceedings of the ACL-SIGLEX Workshop in Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan (2003) 73–80
19. Venkatapathy, S., Joshi, A.K.: Measuring the relative compositionality of verb-noun (vn) collocations by integrating features. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2005) 899–906
20. Farahmand, M., Smith, A., Nivre, J.: A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In: Proceedings of NAACL-HLT. (2015) 29–33
21. Firth, J.R.: {A synopsis of linguistic theory, 1930-1955}. (1957)
22. Katz, G., Giesbrecht, E.: 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. (2006) 12–19
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR. (2013)
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS. (2013)
25. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press (1990)
26. Fazly, A., Cook, P., Stevenson, S.: Unsupervised Type and Token Identification of Idiomatic Expressions. Computational Linguistics **35** (2009) 61–103
27. Li, L., Sporleder, C.: Using gaussian mixture models to detect figurative language in context. In: Proceedings of NAACL/HLT 2010. (2010)
28. Cilibrasi, R., Vitányi, P.M.B.: The google similarity distance. IEEE Trans. Knowl. Data Eng. **19** (2007) 370–383
29. Cilibrasi, R., Vitányi, P.M.B.: Normalized web distance and word similarity. CoRR **abs/0905.4039** (2009)
30. Cook, P., Fazly, A., Stevenson, S.: The VNC-Tokens Dataset. In: Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco (2008)
31. Widdows, D., Dorow, B.: Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In: Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition. DeepLA '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 48–56
32. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics - Volume 2. COLING '92, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 539–545
33. Cook, P., Fazly, A., Stevenson, S.: Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In: Proceedings of the ACL 07 Workshop on A Broader Perspective on Multiword Expressions. (2007) 41–48

34. Birke, J., Sarkar, A.: A clustering approach to the nearly unsupervised recognition of nonliteral language. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), Trento, Italy (2006) 329–226
35. Sporleder, C., Li, L.: Unsupervised Recognition of Literal and Non-literal Use of Idiomatic Expressions. In: EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2009) 754–762
36. Bu, F., Zhu, X., Li, M.: Measuring the non-compositionality of multiword expressions. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics (2010) 116–124
37. Boukobza, R., Rappoport, A.: Multi-word expression identification using sentence surface features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, Association for Computational Linguistics (2009) 468–477
38. Reddy, S., McCarthy, D., Manandhar, S.: An empirical study on compositionality in compound nouns. In: IJCNLP. (2011) 210–218
39. Peng, J., Feldman, A., Vylomova, E.: Classifying idiomatic and literal expressions using topic models and intensity of emotions. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics (2014) 2019–2027
40. Yazdani, M., Farahmand, M., Henderson, J.: Learning semantic composition to detect non-compositionality of multiword expressions. In: EMNLP. (2015) 1733–1742
41. Salehi, B., Cook, P., Baldwin, T.: A word embedding approach to predicting the compositionality of multiword expressions. In: HLT-NAACL. (2015) 977–983
42. Peng, J., Feldman, A., Jazmati, H.: Classifying idiomatic and literal expressions using vector space representations. In: RANLP. (2015) 507–511
43. Salton, G.D., Ross, R.J., Kelleher, J.D.: Idiom token classification using sentential distributed semantics. In: Proceedings of the 54th Annual Meeting on Association for Computational Linguistics. (2016) 194–204
44. Peng, J., Feldman, A.: Experiments in idiom recognition. In: COLING. (2016) 2752–2762
45. Cordeiro, S., Ramisch, C., Idiart, M., Villavicencio, A.: 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), Association for Computational Linguistics (2016) 1986–1997