A Distributional Semantics Model for Idiom Detection: *The Case of English and Russian*

Jing Peng[1], Katsiaryna Aharodnik[2] and Anna Feldman[1]
[1]*Montclair State University, Montclair, NJ, USA*
[2] *CUNY Graduate Center, NY, NY, USA*
{*pengj, feldmana*}*@montclair.edu, kaharodnik@gradcenter.cuny.edu*

Keywords: idiom recognition, corpus annotation, distributional semantics, English, Russian

Abstract: This paper describes experiments in English and Russian automatic idiom detection. Our algorithm is based on the idea that literal and idiomatic expressions appear in different contexts. This difference is captured by our distributional semantics model. We evaluate our model on both languages and compare its results. We show that our model is language-independent. We also describe a new annotated resource we created for our experiments.

1 INTRODUCTION

Idioms add color to language. Without idioms language would be dull and unexciting. Idioms reflect on our cultural values. Cross-linguistically, speakers use different types of idioms to express similar concepts. Thus for example, in American English, one *bites the bullet* while in Russian, one *squeezes the teeth*; in American English one puts *a fly in the ointment* whereas in Russian one adds *a spoon of tar to the barrel of honey*. Both Russians and Americans *shed crocodile tears*. Many Natural Language Processing applications, such as machine translation (MT), natural language understanding (NLU), sentiment and emotion analysis could improve their performance if idioms could be detected automatically with good accuracy. It turns out that a large number of expressions are ambiguous between their idiomatic and literal interpretation and their status (idiomatic vs. literal) can only be determined in context (e.g., *sales hit the roof* vs. *hit the roof of the car*).

Several approaches have been explored in finding a better solution to this problem (e.g.,(Katz and Giesbrecht, 2006; Cook et al., 2007; Fazly et al., 2009; Sporleder and Li, 2009; Li and Sporleder, 2010; Peng et al., 2014a; Peng et al., 2015a; Peng and Feldman, 2016a; Peng and Feldman, 2016c; Pradhan et al., 2017) among others). However, a number of questions about automatic processing of semantic relationships specifically those that are not trivial to define and disambiguate still remain unanswered.

The current paper addresses 1) the problem of determining automatically whether an expression is literal or idiomatic in a specific context, and 2) whether the same methodology can be generalized to other languages besides English. In this paper, we only consider those expressions that are ambiguous in nature and can be interpreted either literally or figuratively depending on the context they occur in. Below we describe our approach.

2 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 (Firth, 1957; Katz and Giesbrecht, 2006)).

2.1 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 (Mikolov et al., 2013a; Mikolov et al., 2013b).

A word can be represented by a vector of fixed dimensionality q that best predicts its surrounding words in a sentence or a document (Mikolov et al., 2013a; Mikolov et al., 2013b). 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 (Mikolov et al., 2013b), 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}

$$P = V^t \sigma_{vn} \quad (1)$$

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

$$p \cdot idf. \quad (2)$$

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 (Mikolov et al., 2013a; Mikolov et al., 2013b). 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,

$$p \cdot tf \cdot idf. \quad (3)$$

2.2 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 (Mikolov et al., 2013a; Mikolov et al., 2013b).

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 (Mikolov et al., 2013a; Mikolov et al., 2013b). Assuming w_is have been centered, we compute the scatter matrix

$$\Sigma = d^t d, \quad (4)$$

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 (Fukunaga, 1990). 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.

3 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 2.2, 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.

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**.

4 DATASETS

4.1 English

We use BNC and a list of VNCs (Cook et al., 2008) (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 1.

Table 1: 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

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.

Table 2: Russian idioms: Examples of different syntactic constructions

Syntactic Construction	Example	Count
Adj(Poss. Pron) + Noun	černyj voron	20
Prep+Noun	bez golovy	17
Prep+Adj+Noun	na moju golovu	3
Verb+(Prep)+Noun	vtsepit'sja v glotku	50
Adv + Verb	žirno budet	2
Noun + Short Adj	kontsert okončen	4
Prep+Noun+Verb	kuda veter duet	4

Table 3: Russian examples: Is = idioms; Ls = literals

Target	Gloss	Interpretation	I	L
s bleskom	with flying colors	brilliantly	222	38
na svoju golovu	on your own head	pain in the neck	119	39
na vysote	at the height	rise to the occasion	147	223
smotret' v glaza	look into the eyes	face (challenges)	45	72
čerez golovu	over the head	go over someone's head	58	224
na nožax	with the knives	to be at daggers drawn	40	39
po barabanu	on the drums	couldn't care less	64	19
vtoroj dom	second home	second home	13	33
vyše sebja	above oneself	beyond the possible	36	9
dlinnyj jazyk	long tongue	chatterbox	26	22

For our experiments reported here, we obtained English word vectors using the word2vec tool (Mikolov et al., 2013a; Mikolov et al., 2013b) 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.

4.1.1 English Datasets

Table 1 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.

4.2 Russian

4.2.1 Corpus Collection

For the list of idioms, a Russian-English dictionary of idioms was used as a primary source (Lubensky, 2013). Initially, 150 idioms (target expressions) were included in the list. The rationale for choosing a

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
						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
						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

certain target expression was that each expression could be interpreted as either idiomatic or literal depending on the context. For example, an expression *postavit' točku* ('put a stop') can appear in a sentence like *Učitelnitsa napomnila Maše čto nužno* ***postavit' točku*** *v kontse predoženija* ('The teacher reminded Masha to put a period at the end of a sentence') with the literal interpretation and also in a sentence like *Ona rešila effektno* ***postavit' točku*** *v svoej kar'ere.* ('She decided to effectively put an end to her career').

The list of idioms includes only multiword expressions (MWE). Each target expression consists of more than one word token, with their length ranging from two, e.g., *dlinnyj jazyk* (long tongue), to four word tokens as in *s penoj u rta* (with frothing at the mouth). Unlike for English, syntactically, target expressions were not limited to a single structure. We collected prepositional phrases, such as (*bez golovy*) ('without head'), nouns with adjectival or possessive modifiers, e.g., *vtoroj dom* ('second home'), verb phrases, e.g., *plyt' po tečeniju* (to go with the flow, Verb +PP), and *postavit' točku* (to put an end, Verb+NP). Table 2 provides a list of syntactic constructions with their counts. The list included idioms in their dictionary form, but each idiomatic expression was extracted from the compiled corpora in any form it appeared in files (conjugated forms for verbs or declined forms for adjectives and nouns).

For the Russian experiments, we used pretrained word vectors, trained on Wikipedia using fastText. These vectors in dimension 300 were obtained using the skip-gram model described in (Bojanowski et al., 2016) with default parameters.

4.2.2 Extracting Target Expressions

A target token is defined as a multiword expression that can be identified as either idiomatic or literal within the text. Each target expression was extracted with one preceding it and one following it paragraph from a source text file. Thus, one entry is defined as a three paragraph text in one file. Each target expression was extracted following the steps below:

1. Convert the online text file to html format. This was done to preserve the html tags and use the tags for paragraph extraction.
2. Save each file as a plain text document with preserved html tags.
3. Extract each target expression (token) from each html document in a three paragraph format, with the second paragraph containing a target expression.
4. Save each three paragraph entry in a separate text file.

Overall, 100 tokens/target expressions were used to create the idiom-annotated corpus.

4.2.3 Annotation

Once the expressions were extracted, each file was annotated manually by two Russian native speakers. The overall inter-annotator agreement was high (Kappa 0.81). Each target expression was assigned a tag, Idiomatic (I) or Literal (L).

A list of 10 target expressions extracted for the corpus is provided in Table 3. It also includes the

Table 5: Average performance of competing methods on Russian idioms.

	na svoju golovu	na vysote	smotret' v glaza	
	get into trouble	to be at one's best	to face (a challenge)	Ave
		Precision		
$p \cdot idf$	0.75	0.49	0.40	0.55
$p \cdot tf \cdot idf$	0.80	0.50	0.50	0.60
$CoVAR_{Fro}$	0.80	0.71	0.49	**0.67**
$CoVAR_{sp}$	0.78	0.64	0.54	0.65
		Recall		
$p \cdot idf$	0.73	0.83	0.40	0.65
$p \cdot tf \cdot idf$	0.76	0.81	0.42	0.66
$CoVAR_{Fro}$	0.88	0.81	0.50	**0.73**
$CoVAR_{sp}$	0.76	0.76	0.50	0.67
		Accuracy		
$p \cdot idf$	0.63	0.64	0.57	0.61
$p \cdot tf \cdot idf$	0.68	0.66	0.67	0.67
$CoVAR_{Fro}$	0.76	0.82	0.65	**0.74**
$CoVAR_{sp}$	0.68	0.77	0.68	0.71

counts of idiomatic and literal interpretations for each idiom. This paper is just a pilot study of the Russian idioms, therefore, we only report the performance of our system on three constructions, but in our future work we will use the entire corpus to evaluate the system.

5 RESULTS

5.1 English

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.

5.2 Russian

The results of the experiments using the new Russian corpus are reported in Table 5. We evaluate our models on three expressions. Right now, our preliminary numbers indicate that the Russian model performs similarly to English, even though Russian is a more morphologically complex language and has free word order.

6 Human judgements correlate with 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 6 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 information.

6.1 Related Work

Previous approaches to idiom detection can be classified into two groups: 1) type-based extraction, i.e., de-

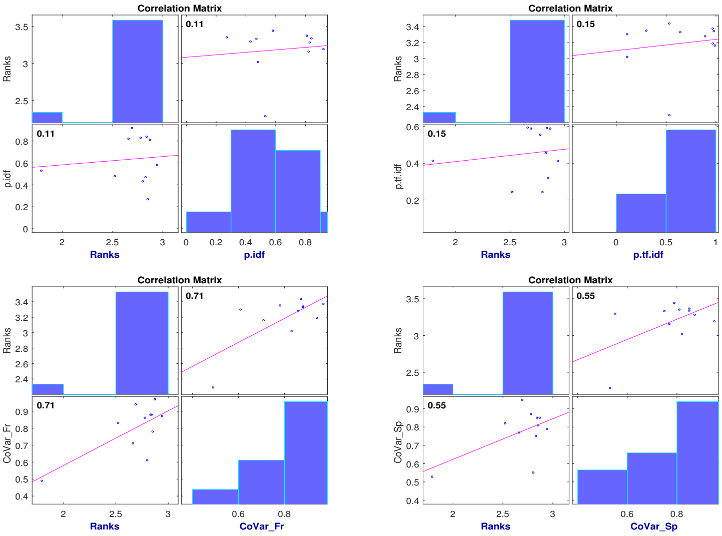

Figure 1: Pairwise Pearson's correlation matrix between the human judgements and the competing algorithms. Top row: $p \cdot idf$ and $p \cdot tf \cdot idf$. Middle row: $CoVar_{Fro}$ and $CoVar_{Sp}$.

tecting idioms at the type level, e.g., (Sag et al., 2002; Fazly et al., 2009; Widdows and Dorow, 2005; Hearst, 1992); 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 (Sag et al., 2002; Fazly et al., 2009). While many idioms do have these properties, many idioms fall on the continuum from being compositional to being partly unanalyzable to completely non-compositional (Cook et al., 2007). (Katz and Giesbrecht, 2006; Birke and Sarkar, 2006; Fazly et al., 2009; Sporleder and Li, 2009; Li and Sporleder, 2010; Bu et al., 2010; Boukobza and Rappoport, 2009; Reddy et al., 2011), 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, (Peng et al., 2014b) investigate the bag of words *topic* representation and incorporate an additional hypothesis–contexts in which idioms occur are more affective. Still, they treat idioms as semantic outliers. (Yazdani et al., 2015; Salehi et al., 2015; Peng et al., 2015b; Salton et al., 2016; Peng and Feldman, 2016b; Cordeiro et al., 2016) explore a range of distributional vector-space models for semantic composition.

7 CONCLUSIONS

In this paper we described a distributional approach to idiom detection and tested it on English and Russian data. Our results suggest that the proposed approach is applicable to languages other than English, with more complex morphology and more flexible word order compared to English.

We also reported the results of an experiment in which human annotators ranked English 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.

Our best performing Russian idioms syntactically represent prepositional phrases (PPs): *na* (Prep) *svoyu* (Attribute) *golovu* (Noun); *na* (Prep) *vysote* (Noun), thus suggesting that our model is able to perform well not just on verb-noun constructions reported for English. Noticeably, the two best performing expressions, when idiomatic, are highly idiomatic (according to our annotators) and we think that the average idiomaticity correlates with the model's performance, similarly to the English case. Like in English, the best performing idioms are those that are highly idiomatic in certain contexts and unambiguously non-idiomatic in others. For instance, it can be seen from the corpus that *na svoju golovu* is associated with certain verbs that appear with literal but not idiomatic interpretations. In addition, those idioms that are harder to disambiguate by human annotators (e.g., *smotret' v glaza*) are also harder to disambiguate automatically.

In our current work we are running experiments on a larger Russian dataset, exploring a variety of syntactic constructions, but experiments described in this paper suggest that we are moving in the right direction toward automatic idiom detection.

Table 6: Average human rankings of 11 idiom types

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

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1319846

REFERENCES

Birke, J. and Sarkar, A. (2006). 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)*, pages 329–226, Trento, Italy.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Boukobza, R. and Rappoport, A. (2009). 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*, pages 468–477. Association for Computational Linguistics.

Bu, F., Zhu, X., and Li, M. (2010). Measuring the non-compositionality of multiword expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 116–124. Association for Computational Linguistics.

Cook, P., Fazly, A., and Stevenson, S. (2007). 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*, pages 41–48.

Cook, P., Fazly, A., and Stevenson, S. (2008). The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco.

Cordeiro, S., Ramisch, C., Idiart, M., and Villavicencio, A. (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. Association for Computational Linguistics.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103.

Firth, J. R. (1957). {A synopsis of linguistic theory, 1930-1955}.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katz, G. and Giesbrecht, E. (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 12–19.

Li, L. and Sporleder, C. (2010). Using gaussian mixture models to detect figurative language in context. In *Proceedings of NAACL/HLT 2010*.

Lubensky, S. (2013). *Russian-English Dictionary of Idioms*. Yale University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Peng, J. and Feldman, A. (2016a). Experiments in idiom recognition. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*.

Peng, J. and Feldman, A. (2016b). Experiments in idiom recognition. In *COLING*, pages 2752–2762.

Peng, J. and Feldman, A. (2016c). In god we trust. all others must bring data. — w. edwards deming — using word embeddings to recognize idioms. In *Proceedings of the 3rd Annual International Symposium on Information Management and Big Data — SIMBig, Cusco, Peru.*

Peng, J., Feldman, A., and Jazmati, H. (2014a). Classifying idiomatic and literal expressions using topic models and intensity of emotions. In *Proceedings of the 2014 Empirical Methods for Natural Language Processing Conference (EMNLP).*

Peng, J., Feldman, A., and Jazmati, H. (2015a). Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP) conference 2015, Hissar, Bulgaria.*

Peng, J., Feldman, A., and Jazmati, H. (2015b). Classifying idiomatic and literal expressions using vector space representations. In *RANLP*, pages 507–511.

Peng, J., Feldman, A., and Vylomova, E. (2014b). 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)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Pradhan, M., Peng, J., Feldman, A., and Wright, B. (2017). Idioms: Humans or machines, it's all about context. In *Proceedings of the 18th International Conference on Computational Linguistics and Inteeligent Text Processing (CICling).*

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *IJCNLP*, pages 210–218.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). 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)*, pages 1–15, Mexico City, Mexico.

Salehi, B., Cook, P., and Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *HLT-NAACL*, pages 977–983.

Salton, G. D., Ross, R. J., and Kelleher, J. D. (2016). Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting on Association for Computational Linguistics*, pages 194–204.

Sporleder, C. and Li, L. (2009). 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*, pages 754–762, Morristown, NJ, USA. Association for Computational Linguistics.

Widdows, D. and Dorow, B. (2005). Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, DeepLA '05, pages 48–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yazdani, M., Farahmand, M., and Henderson, J. (2015). Learning semantic composition to detect non-compositionality of multiword expressions. In *EMNLP*, pages 1733–1742.