Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Representations of Context in Recognizing the Figurative and Literal Usages of Idioms

Changsheng Liu, Rebecca Hwa
Computer Science Department
University of Pittsburgh
Pittsburgh, PA 15260, USA
{changsheng,hwa}@cs.pitt.edu

Abstract

Many idiomatic expressions can be interpreted literally or figuratively, depending on the context in which they occur. Developing an appropriate computational model of the context is crucial for automatic idiom usage recognition. While many existing methods incorporate some elements of context, they have not sufficiently captured the interactions between the linguistic properties of idiomatic expressions and the representations of the context. In this paper we perform an in-depth exploration of the role of representations of the context for idiom usage recognition; we highlight the advantages and limitations of different representation choices in existing methods in terms of known linguistic properties of idioms; we then propose a supervised ensemble method that selects representations adaptively for different idioms. Experimental result suggests that the proposed method performs better for a wider range of idioms than previous methods.

Introduction

A sophisticated natural language processing application should be able to interpret figurative language. There has been extensive work in both metaphor detection (Shutova 2010; Mason 2004) and idiom processing. While metaphors are meant figuratively, an indiomatic expression may be interpreted literally or figuratively, depending on the context in which they occur. Our work focuses on discriminating these two interpretations. This problem is challenging because idioms do not conform to a set of linguistic patterns that can be easily characterized. Idioms have varying degrees of **context diversity**; some are only appropriate under specific situations (e.g., *break the ice*) while others might be more widely applicable (e.g., *rub it in*). Idioms also vary in terms of their **semantic analyzability** (Gibbs, Nayak, and Cutting 1989; Cacciari and Levorato 1998; Nunberg, Sag, and Wasow 1994); some contain words that suggest their figurative interpretations (e.g.,*in the fast lane*) while others are more dissimilar (*break a leg*).

The local context of an idiom holds clues for discriminating between its literal and figurative usages (Katz and Giesbrecht 2006). We argue that in order to fully exploit the information offered by the local context, an idiom usage recognizer ought to take the linguistic properties of the idioms into considerations. Although some previous works do make use of local context, they have not sufficiently taken into account the impact of context diversity and semantic analyzability.

We characterize representation choices into three categories: Lexical Representation (Rajani, Salinas, and Mooney 2014; Birke and Sarkar 2006; Byrne, Fenlon, and Dunnion 2013), Topical Representation (Li, Roth, and Sporleder 2010; Peng, Feldman, and Vylomova 2014) and Distributional Semantic Representation (Sporleder and Li 2009). Each has its own advantages and limitations. Consequently, previous systems tend to perform better for some idioms than others.

We hypothesize that a more flexible and adaptable representation of the context is necessary to account for both context diversity and semantic analyzability. We propose a supervised ensemble approach for learning to adapt multiple contextual representations for different idioms. Our studies compare leading methods against a diverse set of idioms and analyze the effects of contextual representations. We find that by drawing knowledge from multiple representations and adapting to different idioms, an automatic recognizer can achieve better stability without loss of accuracy.

Literal vs. Figurative Interpretations of Idioms

Despite the common perception that an idiomatic expression is mainly used in its figurative sense, an analysis of 60 idioms has revealed that about half of them also have a clear literal meaning (Fazly, Cook, and Stevenson 2009). In some cases, the literal usage of an idiom may even be dominant (Li and Sporleder 2009).The distinction between literal and figurative usages is important for NLP applications. For example, a machine translation system should translate *break the ice* differently in the following two sentences:

(1) When they finally **punched** through the **Arctic** ice cap just shy of the North Pole, it took them five hours to break the ice off their submarine's key hatches so they could reach the fresh air.

(2) US **President** Barack Obama and Cuba's Raul Castro will have a historic face-to-face **encounter** at the **Summit** of the Americas this week, breaking the ice after decades of glacial **relations**.

People can easily identify the correct usage by the surrounding context (e.g., *Arctic*, *North Pole* indicates a literal usage), but this is hard to automate because there are many type of linguistic cues. For example, while some figurative usages of idioms co-occur with lexical cues (e.g., certain prepositions appearing after *break the ice*), others may involve selectional preferences (e.g., having an abstract entity as the subject of *play with fire*) (Li and Sporleder 2010). Moreover, additional factors such as an idiom's context diversity and semantic analyzability also pose challenges for automatic usage recognition.

Context Diversity: This measures how diversified the context of an expression can be. For some, the figurative or literal usages might be closely related to a small range of topics. For example, the figurative use of *break the ice* is not very diverse; it is often associated with political topics, so its contexts are likely to contain words such as *country*, *nation*, *relation*, and *war*. Other idioms, such as *under the microscope*, might be used figuratively with a wider range of topics. If an expression has a low degree of context diversity, even a small set of training examples may be sufficient for developing automatic usage recognizer. For expressions with a very high degree of context diversity, however, supervised learning may be impractical due to training data sparsity.

Semantic Analyzability: This measures the extent to which the meanings of the words forming an idiom contribute to its figurative interpretation (Cacciari and Levorato 1998). For idiom with a high degree of semantic analyzability, its figurative meaning is semantically close to its constituent words, thus the overall figurative context would also be close to its literal context. This could make the usage recognition difficult for methods using distributional semantics such as that of Sporleder and Li (2009).

Although idioms are well studied in the linguistics literature, their observed properties do not always translate to ideas modeled by current computational methods. To the best of our knowledge, this work is the first to quantitatively analyze the impact of context diversity and semantic analyzability from a computational perspective.

Representation of the Usage Context

The current methods that model context can be characterized into three categories: Lexical Representation, Topical Representation and Distributional Semantic Representation.

Lexical Representation

A straightforward representation is to extract surface words from the context. The assumption is that the contexts of an expression used in the same way should have many words in common. The exact range of the context varies from methods to methods. For example, Byrne, Fenlon, and Dunnion (2013) extracted only the left and right boundary words of a target phrase to train Naive Bayesian classifiers. On the other hand, Rajani, Salinas, and Mooney (2014) extracted all non-stop-words and used them as "bag of words" features to train a L2 regularized Logistic Regression (L2LR) classifier (Fan et al. 2008)

One potential drawback for methods using Lexical Representation is that shared context words are not very strong indicators. Expressions with different usages may nonetheless share some words in common in their contexts; and conversely, even when two contexts do not share any common words, an expression may still have the same usage. Another drawback is that if an idiom has a high degree of context diversity, its contexts would contain too many surface words for them to serve as reliable features.

Topical Representation

Instead of directly setting surface words as the feature space, Topical Representation models a context as a point in an idiomatic expression's topic space. The assumption is that even if an idiom is used in different contexts, if the contexts have similar topics, their usage should be similar. One example of a method in this category is the work of Li, Roth, and Sporleder (2010), in which the context is represented as a mixture over latent topics. Another example is the work of Peng, Feldman, and Vylomova (2014), in which the context is represented as a set of topic words extracted by Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003).

An advantage of Topical Representation over Lexical Representation is that it could filter out words that are unrelated to the main topics of the context. The discriminative power of words in the context are different; Lexical Representations generally treat all the words equally. Topical Representation extracts the most critical words for the relevant topics. It can be seen as a refined version of Lexical Representation.

A possible drawback of Topical Representation is that it might overlook some syntactic information which could be used in the usage recognition for some idioms. For example, a figurative usage for *break the ice* may be indicated by the occurrence of the prepositions *over* or *between* after it (Li and Sporleder 2010). These words are generally ignored by methods using Topical Representations, whereas methods using Lexical Representation may include them. Also, similar to Lexical Representation, the context diversity will also influence the effectiveness of Topical Representation.

Distributional Semantic Representation

Methods using the previous two representations essentially rely on the calculation of common words between contexts, which is problematic for idioms with a high degree of context diversity. Distributional Semantic Representation can overcome this problem by using external resource or knowledge base to calculate words similarity. For instance, the following sentence has no word overlap with example (1) and (2). However, the word *monarch* is semantically close to the word *president* in example (2), which suggests they might have the same usage.

(3) Edwards usually manages to break the ice with the taciturn **monarch**.

One method that used Distributional Semantic Representation is the work of Sporleder and Li (2009). They used distributional semantic similarity to calculate the lexical cohesion (Halliday and Hasan 2014) between constituent words

of a idiom and its contextual words. The hypothesis of this method is that if the constituents of a potentially idiomatic expression do not 'fit' in any lexical chains, it is highly likely that the expression is used figuratively.

Despite its advantage, Distributional Semantic Representation still has its limitations. First, for some idioms, it is more effective to just use the surrounding words to detect its usage, such as the preposition *over* or *between* after *break the ice*. Second, since the approach assumes that the overall literal context and figurative context is semantically distant, it is poor at handling idioms with a high degree of semantic analyzability.

Our Model

We treat literal and figurative usage recognition as a special word sense disambiguation problem in the same spirit as Birke and Sarkar (2006). Specifically, we use similarity-based models because they have been shown to be effective in the general problem of word sense disambiguation (Abdalgader and Skabar 2012; Karov and Edelman 1998). In this section, we describe two variants of our model for integrating different contextual representations within our similarity-based framework.

Representation fusion strategies To fuse different context representations, one straightforward strategy is to concatenate all the features using the three representations and build a single similarity based classifier that applies to the concatenated feature (*early fusion*) (Bruni, Tran, and Baroni 2014). Another option is a per-representation strategy; different classifiers are trained independently on the three representations, and afterwards, the results are combined to generate a final output (*late fusion*). We have experimented with both strategies.

The Late Fusion Model

In this model, three classifiers are developed based on Lexical similarity, Topical similarity and Distributional semantic similarity; and a variant of averaged perceptron learning is applied to learn the weights for each classifier according to its discriminative power over different idioms.

Lexical similarity: Given two contexts T_i and T_j of a target expression, we use cosine similarity to calculate their similarity as shown in the Equation 1, where T^i_{bow} and T^j_{bow} denote the bag of word vector of the two contexts. We remove all the stop words in the context except the preceding and following words of the target expression, which tend to be useful for some idioms (Byrne, Fenlon, and Dunnion 2013).

$$Sim_1(T_i, T_j) = \frac{T^i_{bow} \cdot T^j_{bow}}{|T^i_{bow}| \cdot |T^j_{bow}|} \quad (1)$$

Topical similarity: For an idiom, we first run LDA to all the instances and get a set of *m* topics.

$$Topics = \{t_1, t_2, ..., t_m\} \quad (2)$$

For each instance, we represent the context using its probabilities over these topic set.

$$T_{topic} = \{P(t_1), P(t_2), ..., P(t_m)\} \quad (3)$$

Given two contexts T_i and T_j, we use T^i_{topic} and T^j_{topic} to denote their Topical Representations. Their topic similarity is calculated also using cosine similarity.

$$Sim_2(T_i, T_j) = \frac{T^i_{topic} \cdot T^j_{topic}}{|T^i_{topic}| \cdot |T^j_{topic}|} \quad (4)$$

Distributional semantic similarity: Given two contexts T_i and T_j, we calculate their semantic similarity $Sim_3(T_i, T_j)$ using doc2vec (Le and Mikolov 2014). In detail, we use gensim toolkit (Řehůřek and Sojka 2010) and train our model on Wikipedia articles[1]. We empirically set the dimensionality of vector to 200.

$$Sim_3(T_i, T_j) = doc2vec\_sim(T_i, T_j) \quad (5)$$

We distinguish the usage of the target expression by calculating its average similarity (using one of the similarity metrics) to both the literal and figurative example set and assign the label of the set which has higher similarity. Since we have three types of similarity metrics, we now have three "voters". We use v_i to denote the voting vector with each entry representing the voting results for the ith instance of a idiom.

Because idioms vary in properties that may impact each representation differently, we propose to learn the weight for each voter by applying a variant of averaged perceptron learning method (Collins 2002). In addition, we augment the weight learning algorithm by incorporating a novel confidence measure (Schapire and Singer 1999). In our case, the confidence is related to the similarity difference. Let Sim_f be the similarity between the context of the target expression and figurative example set, Sim_l be the similarity between the context of the target expression and literal example set (using any of the three similarity metrics). The ratio between the two similarities is a reasonable confidence measure at first glance. The intuition is that the bigger the difference between the two similarities Sim_f and Sim_l, the more confident the voter is. However, both our empirical evidence and observation from Schapire and Singer (1999) suggest such confidence measure could lead to large and overly confident predictions and ultimately increases the possibility of overfitting. To overcome such issue, we use a smoothed ratio between the two similarities as the confidence value shown in Equation 6.

$$c = 1 + \ln \frac{max(Sim_f, Sim_l)}{min(Sim_f, Sim_l)} \quad (6)$$

Similar to voting vector v_i, we construct the confidence vector c_i for the ith instance; the confidence rated voting vector x_i is the point-wise product of v_i and c_i. Then we apply the voting weight learning algorithm to get the weight w for each voter and classify the target expression usage using Equation 7.

$$y^* = \text{sign}(wx_i) \quad (7)$$

[1]https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

The Early Fusion Model

In this case, we perform L-2 normalization and simply concatenate the vectors of the three representations and then apply cosine similarity metric. The classification process is identical to the single classifier in late fusion strategy.

Experiment

To verify our hypothesis that the automatic recognition of idiomatic usages depends on addressing the interactions between properties of the idioms (i.e., context diversity and semantic analyzability) and the contextual representations of the idiom, we conduct a comparative study across four representative state-of-the-art methods: two for Lexical Representation (Rajani, Salinas, and Mooney 2014; Birke and Sarkar 2006)[2]; one for Topical Representation (Peng, Feldman, and Vylomova 2014); and one for Distributional Semantic Representation (Sporleder and Li 2009). We then compare our proposed methods against these four. The experiments address the following questions:

- To what extent can usage recognizers reliably predict figurative versus literal usages for a wide variety of idioms?
- What is the relative contribution from contextual information compared to other features?
- Does our proposed model of adapting multiple contextual representations succeed in capturing the interactions between representational choices and context diversity and semantic analyzability?

Data

The data is from SemEval 2013 task 5B (Korkontzelos et al. 2013). This corpus consists of 10 target idioms of different types that can be used both literally and figuratively. For each idiom, several instances are provided corresponding to its literal or figurative usages. There are 4 instances labeled as *both* which could lead to ambiguity are removed and we get 2371 instances in total, among which 1185 instances are literal usages and 1186 instances are non-literal usages.

Evaluation metric

For comparison and analysis, we primarily rely on the standard F1 score for the recognition of the figurative usage. The overall accuracy of both figurative and literal usage is not ideal for analysis because it can be misleading for idioms with a heavily skewed usage distribution. Nonetheless, we do report it because it is ultimately what downstream applications care about.

Implementation

We reimplmented the four methods, but with two minor changes. First, Sporleder and Li used Normalized Google Distance (NGD) to measure the semantic relatedness (Cilibrasi and Vitanyi 2007), but the API that NGD has a restriction on the number of queries it can make; therefore, we use word embeddings for calculating the distributional semantic similarity (Mikolov et al. 2013). Second, we did not encode Birke and Sarkar's SuperTags feature because they reported that the overall gain was only 0.5%. We do not expect these two changes to have significant impact on the findings.

We run ten fold cross validation for the two supervised methods (Rajani et al. and Peng et al.). In each round of the cross validation, we randomly select half of the training sample as the example set; the remaining half of the training sample is used to learn the weight for the three representations.

Results and Observations

Table 1 reports the performances of the four comparative state-of-the-art methods. As expected, the supervised classifier by Rajani et al.'s achieves the best performance while the unsupervised method by Sporleder and Li has the lowest scores for most idioms.

1) Reliability across idioms Comparing across different idioms for each method, we observe large performance variances. For Rajani et al., the F_{fig} is as low as 0.54 for *break a leg* and as high as 0.83 for *through the roof*. Similarly, Peng et al., the lowest F_{fig} is 0.46 for *under the microscope* and the highest is 0.75 for *at the end of the day*.

2) Significance of contextual features Table 2 shows the performances of the two supervised methods limited to just the contextual features. Compared to their full model counterparts in Table 1, we see that the contribution from the additional features is limited, and its impact varies from idiom to idiom. For some, the additional features might have negative effect on the performance (cf. *in the bag*). These results suggest that contextual features are essential to the idiom usage recognition task.

3) Results of the proposed model Table 3 reports the performances of our proposed models (both early fusion and late fusion), each of the three component representations in the late fusion model, and the best of the comparative methods for each idiom. The performance of our full late fusion model is competitive; most of our F_{fig} are higher than the best results from the other methods. The late fusion model is more stable than the other methods, with a narrow range of F_{fig} scores, from 0.68 (*under the microscope*) to 0.85 (*at the end of the day*).

Discussion: Performance Variance

We have hypothesized that the variance in performance is partially due to context diversity. To assess the diversity of contextual words for a target idiom, we measure the diversity of topics in which the idiom can be used. To do so, we run LDA on the examples for a given idiom and vary the parameter of topic number. For each topic number, a log-likelihood value is calculated, indicating how well the generated topic model fits the example set. We select the number of topics with the highest log-likelihood value to approximate the measurement of diversity of topics for the idiom

[2]We include Rajani et al.'s method because it achieves the best performance on the SemEval 2013 task 5B corpus.

Idiom	Rajani et al.		Peng et al.		Sporleder and Li.		Birke and Sarkar	
	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A
at the end of the day	**0.81**	**0.73**	0.75	0.63	0.72	0.59	0.69	0.63
bread and butter	**0.81**	**0.8**	0.75	0.70	0.66	0.58	0.67	0.70
break a leg	0.54	**0.8**	0.49	0.63	**0.67**	0.7	0.61	0.65
drop the ball	**0.61**	**0.79**	0.58	0.67	0.45	0.32	0.52	0.76
in the bag	**0.72**	**0.71**	0.68	0.66	0.65	0.50	0.64	0.71
in the fast lane	**0.78**	0.67	0.72	**0.69**	0.52	0.61	0.68	0.65
play ball	**0.75**	0.72	0.68	0.67	0.51	0.40	0.73	**0.75**
rub it in	**0.67**	**0.69**	0.5	0.47	0.55	0.46	0.44	0.49
through the roof	**0.83**	**0.81**	0.68	0.69	0.61	0.51	0.69	0.74
under the microscope	**0.55**	0.74	0.46	0.64	0.42	0.41	**0.55**	**0.79**

Table 1: Result of different methods. Rajani et al., Birke and Sarkar use Lexical Representation; Peng et al. use Topical Representation; Sporleder and Li. use Distributional Semantic Representation. F_{fig} denotes F1 score of figurative usage recognition and A denotes the overall accuracy. For each idiom, the boldfaced number shows the best performance among the four methods while underlined shows the worst.

Idiom	Rajani et al.		Peng et al.	
	F_{fig}	A	F_{fig}	A
at the end of the day	0.8	0.71	0.73	0.61
bread and butter	0.85	0.84	0.74	0.69
break a leg	0.57	0.77	0.46	0.60
drop the ball	0.59	0.77	0.59	0.68
in the bag	0.75	0.75	0.66	0.62
in the fast lane	0.78	0.68	0.68	0.64
play ball	0.84	0.82	0.64	0.61
rub it in	0.66	0.67	0.51	0.49
through the roof	0.78	0.77	0.67	0.62
under the microscope	0.5	0.74	0.51	0.66

Table 2: Result of two supervised methods using only contextual features. F_{fig} denotes F1 score of figurative usage recognition and A denotes the overall accuracy.

(see Formula 8, D denotes the example set, M_n denotes the generated model with n as the topic number).

$$\underset{n}{\operatorname{argmax}}\ logP(D|M_n) \tag{8}$$

We randomly select 32 literal instances and 29 figurative instances (the minimum number of instances among all the target idioms) for each idiom from the corpus and ran the process mentioned above. The results are shown in Table 4.

We observe that *under the microscope* has the highest topic number, suggesting that it has a high context diversity; it is an idiom that is difficult for all four methods. In contrast, the optimal topic numbers for *bread and butter* is the lowest, suggesting that it has a low context diversity; accordingly, methods using Lexical Representation and Topical Representation performed well on it. We also calculate the Pearson correlation between F_{fig} and the total topic number.[3] The r value is -0.86 for Rajani et al., which suggests strong negative correlation; while the r values for Peng et al. and Birke and Sarkar are -0.72 and -0.62 respectively, a more moderate negative correlation. Although the r value for Sporleder and Li is -0.72, which also suggests a moderately negative correlation, its trend is less reliable. For example, *through the roof* has the lowest topic number (12), but the F_{fig} score (0.61) is well below the best result (0.72); *break a leg* has a relatively high topic number (18), but the F_{fig} score (0.67) is better than the other three methods. These observations suggest that context diversity does influence performances, especially for methods using Lexical or Topical Representation.

Performance variance may also be due to semantic analyzability, especially for methods using Distributional Semantic Representation. We quantify semantic analyzability in the following way. For an idiom, we prepare two sets of instances; one consists of literal instances and the other consists of figurative instances. Then we approximate the overall figurative and literal context similarity of the idiom by measuring the averaged semantic similarity between the two sets. We use L and F to represent the literal and figurative set respectively. The averaged similarity of F and L is calculated using the following Formula:

$$S_{set}(F,L) = \frac{1}{|F|}\sum_{\forall T_f \in F} \max_{\forall T_l \in L} doc2vec_{sim}(T_f, T_l) \tag{9}$$

Table 5 shows our semantic analyzability measure on the 10 idioms. The idiom with the highest similarity score is *drop the ball*, indicating that literal and figurative usages are hard to separate. This corresponds to the poor performance of Sporleder and Li's method on it. In contrast, *break a leg* has the lowest similarity score, which corresponds to the high F_{fig} using Sporleder and Li's method. We also calculate the Pearson correlation coefficient between the F_{fig} and $S_{set}(F,L)$; the r value is -0.77 for Sporleder and Li's method, which suggests moderate negative correlation between the two variables; the r values for the other three methods are -0.03, 0.17, 0.06, respectively. These findings

[3]For methods from Rajani et al. and Peng et al, we use the F_{fig} from Table 2 (the implementation without additional features).

Idiom	Best other		Lexical		Topical		Distributional		Early fusion		Late fusion	
	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A
at the end of the day	0.81	0.73	0.82	0.75	0.81	0.74	0.72	0.69	0.79	0.73	**0.85***	**0.81***
bread and butter	0.81	0.8	0.83	0.79	**0.84**	0.80	0.57	0.61	0.82	0.71	**0.84**	**0.83**
break a leg	0.67	0.7	0.58	0.7	0.56	0.63	0.69	0.71	0.66	0.7	**0.73***	**0.71**
drop the ball	0.61	0.79	0.65	0.81	0.59	0.77	0.51	0.69	0.67	0.82	**0.72***	**0.85***
in the bag	0.72	0.71	0.67	0.66	0.67	0.69	0.74	0.71	0.73	0.65	**0.75***	**0.74**
in the fast lane	**0.78**	0.67	0.68	0.69	0.70	0.73	0.59	0.65	0.54	0.69	0.72*	**0.74***
play ball	0.75	0.72	0.76	0.77	0.71	0.76	0.61	0.71	0.78	0.74	**0.82***	**0.81***
rub it in	0.67	0.69	0.65	0.68	0.73	0.71	0.62	0.71	0.7	0.71	**0.78***	**0.76***
through the roof	**0.83**	0.81	0.81	0.8	0.71	0.69	0.65	0.72	0.81	0.66	0.81	**0.85**
under the microscope	0.55	**0.79**	0.64	0.73	0.47	0.66	0.52	0.69	0.58	0.75	**0.68***	0.75

Table 3: The comparison between our method and competing methods. The "Best other" column shows the best result from the other methods. $*$ indicates the difference between the "Late fusion" and "Best other" is statistically significant, χ^2 text, p = 0.05. The boldfaced number shows the best performance.

lend credence to our argument that semantic analyzability influences the effectiveness of Distributional Semantic Representation.

Idiom	T_{Fig}	T_{Lit}	Total
at the end of the day	9	4	13
bread and butter	7	5	12
break a leg	12	6	18
drop the ball	13	8	21
in the bag	11	6	17
in the fast lane	9	7	16
play ball	9	7	16
rub it in	12	5	17
through the roof	8	4	12
under the microscope	16	7	**23**

Table 4: Optimal topic numbers for different idiom instances. T_{Fig} means the topic number of figurative set, T_{Lit} means the topic number of literal set.

Discussion: Combining Different Representations

Throughout this paper, we have argued for the importance of combining different representations of the context. As shown in Table 3, the stability of the late fusion model did improve. But do the results of the individual components corroborate our arguments about the interactions between linguistic properties and specific representations?

Consider *break a leg*, which has a higher context diversity (18 topics) but lower semantic analyzability (0.27 similarity score). Our model's Lexical Representation and Topical Representation components are not as effective as the Distributional Semantic Representation component; they have an F_{fig} score of 0.58, 0.56, and 0.69 respectively. Similarly, for an idioms with a higher semantic analyzability but a lower context diversity like *bread and butter*, our model's Distributional Semantic Representation component performed worse individually than the Lexical Representation and Topical Representation components.

Idiom	Similarity
at the end of the day	0.28
bread and butter	0.32
break a leg	0.27
drop the ball	**0.37**
in the bag	0.29
in the fast lane	0.35
play ball	0.34
rub it in	0.28
through the roof	0.32
under the microscope	0.34

Table 5: A measure of Semantic Analyzability

In both cases, our method has effectively adapted to the particulars of the idioms and increased the contributions from the well performing components. For *break a leg*, the weights of the components are $[0.23, 0.19, 0.58]$, favoring the Distributional Representation to obtain an F_{fig} of 0.73. For *bread and butter*, the weights are appropriately shifted to the Lexical Representation and Topical Representation components ($[0.4, 0.43, 0.17]$) for an overall F_{fig} of 0.84.

Finally, we observe that Lexical and Topical Representation generally perform better than Distributional Semantic Representation. This may be due to the challenges of calculating semantic similarity between short texts.

Conclusion

We have argued for the importance of two linguistic properties in idioms (context diversity and semantic analyzability) for distinguishing their figurative and literal usages. Experimental results show that leading methods with fixed representations do not perform equally well on different types of idioms. We have proposed a supervised ensemble approach to adaptively combine multiple contextual semantic representations for different idioms. Evaluated on a diverse set of idioms, we find that our method can achieve better stability without loss of accuracy.

References

Abdalgader, K., and Skabar, A. 2012. Unsupervised similarity-based word sense disambiguation using context vectors and sentential word importance. *ACM Transactions on Speech and Language Processing (TSLP)* 9(1):2.

Birke, J., and Sarkar, A. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3:993–1022.

Bruni, E.; Tran, N. K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49(1):1–47.

Byrne, L.; Fenlon, C.; and Dunnion, J. 2013. IIRG: A naive approach to evaluating phrasal semantics. *In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation* 45(4).

Cacciari, C., and Levorato, M. C. 1998. The effect of semantic analyzability of idioms in metalinguistic tasks. *Metaphor and Symbol* 13(3):159–177.

Cilibrasi, R. L., and Vitanyi, P. 2007. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on* 19(3):370–383.

Collins, M. 2002. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 1–8. Association for Computational Linguistics.

Fan, R. E.; Chang, K. W.; Hsieh, C. J.; Wang, X. R.; and Lin, C. J. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research* 9:1871–1874.

Fazly, A.; Cook, P.; and Stevenson, S. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1):61–103.

Gibbs, R. W.; Nayak, N. P.; and Cutting, C. 1989. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of memory and language* 28(5):576–593.

Halliday, M. A. K., and Hasan, R. 2014. *Cohesion in English*. Routledge.

Karov, Y., and Edelman, S. 1998. Similarity-based word sense disambiguation. *Computational Linguistics* 24(1):41–59.

Katz, G., and Giesbrecht, E. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 12–19. Association for Computational Linguistics.

Korkontzelos, I.; Zesch, T.; Zanzotto, F. M.; and Biemann, C. 2013. Semeval-2013 task 5: Evaluating phrasal semantics.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Li, L., and Sporleder, C. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 315–323. Association for Computational Linguistics.

Li, L., and Sporleder, C. 2010. Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 683–691. Association for Computational Linguistics.

Li, L.; Roth, B.; and Sporleder, C. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1138–1147. Association for Computational Linguistics.

Mason, Z. J. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics* 30(1):23–44.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Nunberg, G.; Sag, I. A.; and Wasow, T. 1994. Idioms. *Language* 491–538.

Peng, J.; Feldman, A.; and Vylomova, E. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. *EMNLP* 2019–2027.

Rajani, N. F.; Salinas, E.; and Mooney, R. 2014. Using abstract context to detect figurative language.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Schapire, R. E., and Singer, Y. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning* 37(3):297–336.

Shutova, E. 2010. Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 688–697. Association for Computational Linguistics.

Sporleder, C., and Li, L. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 754–762. Association for Computational Linguistics.