Heuristically Informed Unsupervised Idiom Usage Recognition

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

Abstract

Depending on the surrounding context, an idiomatic expression may be interpreted figuratively or literally. This paper proposes an unsupervised learning method for recognizing the intended usages of idioms. We treat the possible usages as a latent variable in probabilistic models and train them in a linguistically motivated feature space. Crucially, we show that distributional semantics serves as a helpful heuristic for formulating a *literal usage metric* to estimate the likelihood that the idiom is intended literally. This information can then guide the unsupervised training process for the probabilistic models. Experiments show that our overall model performs competitively against supervised methods.

1 Introduction

Many idiomatic expressions may be interpreted both figuratively or literally. Their intended usages depend on how they fit with their contexts. For example, the idiom ”spill the beans” is used figuratively in the first instance below, and literally in the second:

> *(1)* [fig.] *The **beans have been spilled**. From what I've read on Twitter I could probably fill out the forms and submit it to the FISA court. I don't know what the big secret is.*[1]

> *(2)* [lit.] ***Spill the beans**, flip the fruit, bust open a box of hot pockets. Make a general mess of the kitchen.*[2]

This type of ambiguity is commonplace – prior work suggests that about half out of a sample of 60 idioms have a clear literal meaning as well as a figurative one (Fazly et al., 2009). Being able to distinguish the intended usage of an idiom in context has been shown to benefit many natural language processing (NLP) applications, e.g., machine translation and sentiment analysis (Salton et al., 2014; Williams et al., 2015).

While supervised models for idiom usage recognition have had some successes, they require appropriately annotated training examples (Peng et al., 2014; Byrne et al., 2013; Liu and Hwa, 2017). A more challenging problem is to recognize idiom usages without a dictionary or some annotated examples (Korkontzelos et al., 2013). Some previous unsupervised models tried to exploit linguistic differences in usages. For example, Fazly et al.(2009) observed that an idiom appearing in its canonical form is usually used figuratively; Sporleder and Li(2009) relied on the break in lexical coherence between the idioms and the context to signal a figurative usage. These heuristics, however, are not always applicable because the distinctions they depend upon may not be present or obvious. To improve generalization across different idioms and usage contexts, we need a more reliable heuristic, and appropriately incorporate it into an unsupervised learning framework.

We propose a heuristic that differentiates usages based on distributional semantics (Harris, 1954; Turney and Pantel, 2010). Our key insight is that when an idiom is used literally, its relationship with its context is more predictable than when it is used figuratively. This is because the literal meaning of an idiom is compositional (Katz and Giesbrecht, 2006), and the constituent words that make up the idiom are also meant literally. For example, in instance (2), *spill* is meant literally and can take on objects other than *beans*; moreover, one of the context words, *mess*, can often be seen to co-

[1] https://twitter.com/BTeboe/status/958792419302100993

[2] https://twitter.com/DukeRaccoon/status/477530732173471744

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1723–1731
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

occur with *spill* in other text, even without *beans*. Our strategy is to represent an idiom's literal usage in terms of the word embeddings of the idiom's constituent words and other words they frequently co-occur with. Then, for any instance in which the idiom's usage is not known, we only need to determine the semantic similarity between that instance and the idiom's literal representation. We define a *literal usage metric* that estimates the likelihood that an instance would be labeled ”literal”.

While the literal usage metric captures the distributional semantic information of the context, we find that some other linguistic cues are also significant for usage detection (such as whether the subject of the sentence is a person); therefore, we allow our model to further refine through unsupervised methods. Specifically, we treat the usage (*figurative* or *literal*) as a hidden variable in probabilistic latent variable models, and we define a set of features that are linguistically relevant for idiom usage detection as observables. We integrate our literal usage metric with the latent variable models by treating the metric outputs as *soft labels* to guide the latent variable models toward grouping by usages.

We hypothesize that unsupervised learning in a more linguistically motivated feature space, informed by soft labels from a semantically driven metric, will produce more robust classifiers. We conduct experiments comparing our approach against other supervised and unsupervised baselines. Results suggest that our approach achieves performances that are competitive to supervised models.

2 Related Work

Despite the common perception that idioms are mainly used figuratively, many can also be meant literally. A number of models have been proposed in the literature to recognize an idiom's usages under different context. Many rely on specific linguistic property to draw a clear-cut decision boundary between literal and figurative usages. For example, Fazly et al. (2009) proposed a method that relies on the concept of *canonical form*. Based on the observation that while literal usages are less syntactically restricted, figurative usages tend to occur in a small number of canonical form(s). As shown in the examples above, however, this rule of thumb does not always hold. Sporleder and Li (2009) proposed a method by building a cohesion graph to include all content words in the context; if removing the idiom improves cohesion, they assume the instance is figurative. Later, Li and Sporleder (2009) used their cohesion graph method to label a subset of the test data with high confidence. This subset is then passed on as training data to the supervised classifier, which then labels the remainder of the dataset.

When manually annotated examples are available, supervised classifiers are effective. Rajani et al. (2014) extracted all non-stop-words in the context and used them as ”bag of words” features to train a L2 regularized Logistic Regression (L2LR) classifier (Fan et al., 2008). As local context of an idiom holds clues for discriminating between its literal and figurative usages, Liu and Hwa (2017) find that context representation also plays a significant role in idiom usage recognition. They took an adaptive approach, applying supervised ensemble learning over three classifiers based on different context representations (Peng et al., 2014; Birke and Sarkar, 2006; Rajani et al., 2014).

3 Our Approach

Given a target idiomatic expression and a collection of instances in which the idiom occurs, our proposed system (Figure 1) determines whether the idiom in each instance is meant figuratively or literally. We first build a **Literal Usage Representation** for each idiom by leveraging the distributional semantics of its constituents (Sec 3.1). Given an instance of idiom, we can determine its usage by the semantic similarity between the context of the instance and the **Literal Usage Representation**. We define a **Literal Usage Metric** to transform the semantic similarity score into soft label, i.e., an initial rough estimation of the instance's usage (Sec 3.2). Finally, we treat the soft labels as distant supervision for downstream probabilistic latent variable models, in which the usages are considered as the hidden variables and are represented over a set of features.

3.1 Literal Usage Representation

An idiom co-occurs with different sets of words depending on whether it is meant literally or figuratively. For example, when used literally, *get wind* is more likely to co-occur with words such as *rain*, *storm* or *weather*; in contrast, when used figuratively, it frequently co-occurs with *rumor* or

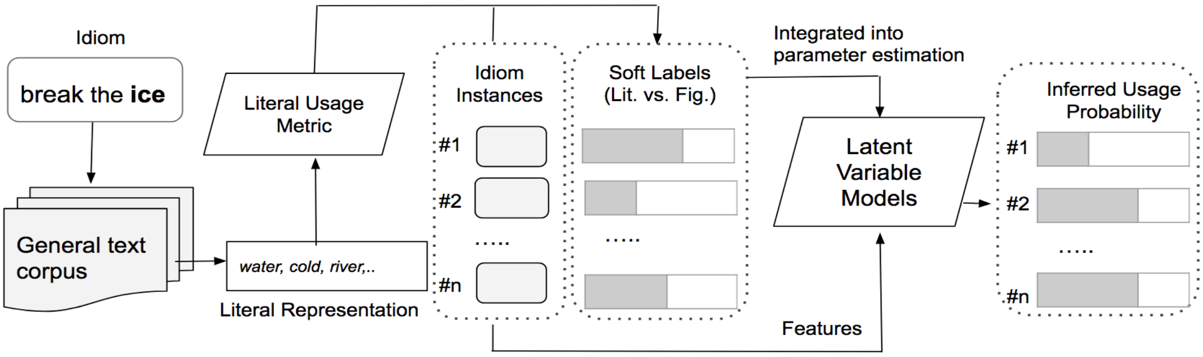

Figure 1: An overview of our unsupervised idiom usage recognition model

story, etc. Comparing the two sets of words associated with the idiom, we see that the literal set of words also tend to co-occur with just *wind*, a constituent word within the idiom. Therefore, even without annotated data or dictionary, we may still approximate a representation for the literal meaning of an idiom by the idiom's constituent words and their semantic relationship to other words. To do so, we begin by initializing a *literal meaning set* to just the idiom's main constituent words[3]; we then grow the set by adding two types of semantically related words. First, we look for co-occuring words in a large textual corpus (e.g., (David et al., 2005)): for each constituent word w, we randomly sample s sentences that contain w from the corpus; we extract the top n most frequent words (excluding stop words) and add them to the literal meaning set. Second, we look for words that are semantically close in a word embedding space: we train a continuous bag-of-words (CBOW) embedding model (Mikolov et al., 2013) and add additional t words that are the most related to w using cosine similarity.

All together, the literal usage representation is a collection of vectors, i.e., the embeddings of the words in the final extended literal meaning set. The size of the set depends on parameters s, n, and t; if the chosen values are too small, we do not end up with a word collection that is representative enough; if the numbers are too large, we would only be wasting computing resources chasing Zipfian tails. Parameter setting choices are discussed further in the experiment section.

3.2 Literal Usage Metrics

Among all the instances to be classified, we expect the context words of the literal cases to be more semantically close to the literal usage representation we just formed. Let L denote the set of words in the literal usage representation for the target idiom. For each instance, let C be the set of non-stop context words in the instance. We calculate s, the semantic similarity score between the context of the instance and the literal usage representation as follows:

$$s = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|L|} \sum_{l \in L} sim(c, l) \qquad (1)$$

where c denotes a word in C, l denotes a word in L and $sim(c, l)$ refers to the cosine similarity between the word embeddings of c and l.

Let $S = \{s_1, s_2, ...s_n\}$ be the set of semantic similarity scores for all the instances we wish to classify. Instances with higher scores are more likely to use the idiom literally. A naive literal usage metrics is to choose a predefined threshold for all idioms and label all the instances with score above the threshold as literal usages. This approach is unlikely to work well in practice. As noted by previous work, idioms have different levels of **semantic analyzability** (Gibbs et al., 1989; Cacciari and Levorato, 1998). When an idiom has a high degree of semantic analyzability, its contextual words will be more semantically close to the literal usage representation, thus a higher threshold is needed.

In this work, we select a different decision threshold for each idiom adaptively based on the similarity scores distribution. And most importantly, rather than generate a hard label, we transform these scores into a probabilistic metric,

[3] We observe that the nouns tend to be the most indicative of the idiom's literal meaning, but if the idiom does not contain any noun, we back off to any constituent word that is not a stop word.

where 0 means the usage in the instance is almost certainly figurative while 1.0 means it is literal.

We propose a metric based on the principle of **Minimum Variance (MinV)**. That is, we first sort the scores in S and choose the threshold (from these scores) that minimizes the sum of variances of the two resulting clusters. For each instance i, we then apply the following metric to estimate the probability that the idiom in instance i is meant **literally** based on its semantic similarity score s_i :

$$Pr_i = \frac{1}{1 + e^{-k*(s_i - t)}} \tag{2}$$

where k is a constant weighting factor and t indicates the learned threshold. The intuition is that the larger the difference between s_i and the threshold is, the more likely the instance i is literal; the probability of literal usage is not linearly correlated to the difference, we use the sigmoid function to account for this non-linearity. We incorporate *k* to scale the value of the difference since it is generally very small (close to 0). Without *k*, all the Pr values gravitate toward 0.5, rendering the soft label being equivalent to random guess. We set k to 5 for all the idioms based on a development set.

3.3 Heuristically Informed Usage Recognition

The soft label, generated by MinV (the literal usage metric), captures the distributional semantic information of the context. In practice, there are a variety of other linguistic features which are also informative of the intended usage of idiom. We explore probabilistic latent variable models over a collection of features that are linguistically relevant for idiom usage detection. The soft label is integrated into the unsupervised learning of hidden usages as a distant supervision. In this section, we will describe the proposed features in the latent variable models and how we integrate the soft label into the learning process.

3.3.1 Latent Variable Models

To predict an idiom's usage in instances, we consider two representative probabilistic latent variable models: Latent Dirichlet Allocation (LDA) (Blei et al., 2003)[4] and unsupervised Naive Bayes (NB). For both models, the latent variable is the idiom usage (figurative vs. literal); the observables are linguistic features that can be extracted from the instances, described below:

Subordinate Clause We encode a binary feature indicating whether the target expression is followed by a subordinate clause (the Stanford Parser (Chen and Manning, 2014) is used). This feature is useful for some idioms such as *in the dark*. It usually suggests a figurative usage as in *You've kept us totally in the dark about what happened that night*.

Selectional Preference Violation of selectional preference is normally a signal of figurative usage (e.g., having an abstract entity as the subject of *play with fire*). We encode this feature if the head word of the idiom is a verb and focus on the subject of the verb. We apply Stanford Name Entity tagger (Finkel et al., 2005) with 3 classes ("Location", "Person", "Organization") on the sentence containing the idiom. If the subject is labeled as an Entity, its class will be encoded in the feature vector. Pronouns such as "I" and "he" also indicate the subject is a "Person". However, they are normally not tagged by Stanford Name Entity tagger. To overcome this issue, we add Part-of-Speech of the subject into the feature vector.

Abstractness Abstract words refer to things which are hard to perceive directly with our senses. Abstractness has been shown to be useful in the detection of metaphor, another type of figurative language (Turney et al., 2011). A figurative usage of an idiomatic phrase may have relatively more abstract contextual words. For example, in the sentence *She has lived life in the fast lane*, the word *life* is considered as an abstract word. This is a useful indicator that *in the fast lane* is used figuratively. We use the MRC Psycholinguistic Database Machine Usable Dictionary (Coltheart, 1981) which contains a list of 4295 words with their abstractness measure between 100 and 700. We calculate the average abstractness score for all the contextual words (with stop words being removed) in the sentence containing the idiom. The score is then transformed into categorical feature to overcome sparsity problem based on the following criteria: concrete (450 - 700), medium (350 - 450), abstract (100 - 350).

Neighboring Words Words preceding and following the idiomatic expression can be very informative in terms of usage recognition. For example, words such as *relax* or *shower* before the idiom *in hot water* often signal a literal usage.

[4] Although originally conceived for modeling document content, LDA can be applied to any kind of discrete input

Part-of-Speech of the Neighboring Words Class of neighboring words might be useful as well. For example, a pronoun preceding *dog's age* generally indicates a literal usage, as in *I think my dog's age is starting to catch up. She sometimes needs help to jump on to my bed*, while a determiner usually marks a figurative usage, as in *It's been a dog's age since I've used Twitter.*

3.3.2 Incorporating Soft Label into Usage Recognition

Given a collection of instances and their features, either LDA or NB can separate the instances into two groups (hopefully, by usages), but it does not associate the right label (i.e., "figurative" or "literal") to the groups. We do not want to rely on any manual annotations for this step. Therefore, we integrate the automatically generated soft labels (based on MinV, our literal usage metric) into the unsupervised learning procedure as a weak form of supervision. Formally, we want to estimate each instance's posterior distribution over (literal/figurative) usages θ_{du} and usage-feature distribution ϕ_{uf}. For LDA, we derive a Gibbs sampling algorithm which incorporates the soft label into the learning procedure. We refer it as informed Gibbs sampling (infGibbs). For unsupervised naive Bayes model, we adapt the classical Expectation-Maximization algorithm to integrate the soft label. We refer it as informed Expectation-Maximization (infEM).

Informed Gibbs Sampling The Gibbs sampling algorithm (Griffiths and Steyvers, 2004) used in traditional LDA initializes each word token a random hidden topic. The system needs to interpret the learned topics post-hoc, e.g., by human annotation. In our case, for each feature f in each instance, an initial random usage **biased** by the instance's soft label is assigned to f (i.e., a Bernoulli trial). Since the soft label explicitly encodes an instance's literal and figurative usage distribution, we do not need to interpret the learned usages at the end of the algorithm. Based on these assignments, we build a feature-usage counting matrix C^{FU} and instance-usage counting matrix C^{DU} with dimensions $|F| \times 2$ and $|D| \times 2$ respectively ($|F|$ is the feature size and $|D|$ is the number of instances): $C^{FU}_{i,j}$ is the count of feature i assigned to usage j; $C^{DU}_{d,j}$ is the count of features assigned to usage j in instance d. Then for each feature f in each instance, we resample a new usage for f and matrices C^{FU} and C^{DU} will be updated accordingly. This step will be repeated for T times. The resampling equation is:

$$p(u_i = j | u_{-i}, f) \propto p_j \cdot \frac{C^{f_i}_{-i,j} + \beta}{C^{(*)}_{-i,j} + |F|\beta} \cdot \frac{C^{d_i}_{-i,j} + \alpha}{C^{d_i}_{-i,*} + |U|\alpha} \tag{3}$$

where i indexes features in the instance d, j is an index into literal and figurative usages, $*$ indicates a summation over that dimension and $-$ means excluding the corresponding instance. The first factor p_j is the soft label encoding prior usage distribution. The second factor represents the probability of feature f under usage j ($C^{f_i}_{-i,j}$ is the count of the feature f assigned to usage j, excluding the current usage assignment u_i). The third factor represents the probability of usage j in the current instance ($C^{d_i}_{-i,j}$ is the count of linguistic features which are assigned to usage j in the current instance, excluding the current feature f). The value of $|U|$ is 2, representing the number of usages (i.e., figurative and literal). α and β are the hyper-parameters from the Dirichlet priors (we set both of them to 1). The core idea of Equation 3 is to integrate both distribution semantic information (soft label, the first factor) and linguistically motivated features (the second and third factors) into the inference procedure.

The matrices of C^{FU} and C^{DU} from the last $10\% * T$ iterations are averaged and then normalized to approximate the true usage-feature distribution ϕ_{uf} and instance-usage distribution θ_{du} respectively. The final result is determined by θ_{du}, i.e., assigning each instance with the usage of probability higher than 0.5. We do average to have a more stable result because an accidental bad sampling would affect our model negatively if we only use the C^{FU} and C^{DU} from the last iteration. This procedure is important for some idioms if their feature space is sparse. The iteration number T is set to 500 based on a development set.

Informed Expectation Maximization Combining a Naive Bayes classifier with the EM algorithm has been widely used in text classification and word sense disambiguation (Hristea, 2013; Nigam et al., 2000). In our case, we want to construct a model to recover the missing literal and figurative labels of the instances of the target idiom. This section describes two extensions to the basic EM algorithm for idiom usage recognition. The extensions help improve parameter estimation by taking the automatically learned soft labels into consideration.

Our informed EM method extends a basic version for NB (Hristea, 2013), where the initial parameter values θ_{du} and ϕ_{uf} are chosen randomly. At each iteration, the E-step of the algorithm estimates the expectations of the missing values (i.e. the literal and figurative usage) given the latest iteration of the model parameters; the M-step maximizes the likelihood of the model parameters using the previously-computed expectations of the missing values. As we've done with extending Gibbs sampling for LDA, we also perform two similar adaptations on conventional EM for NB to incorporate soft labels. First, we assign each instance an initial usage distribution θ_{du} directly using the soft label, and then initialize the usage-feature distribution ϕ_{uf} using these assignments. We refer it as informed initialization. Second, in the E-step, we multiply the expectation result of the basic EM with the soft label as the new expected usage for each instance (i.e., updating θ_{du}). The M-step is the same as basic EM to update the usage-feature distribution ϕ_{uf}.

4 Evaluation

We conduct experiments to address three questions:

1. How effective is our overall approach? How does it compare against previous work?

2. How effective is our literal usage metric (i.e., MinV) compared to other heuristics?

3. How effective is our literal usage metric at informing downstream learning processes?

4.1 Experimental Setup

Models Our main experiments will evaluate the two variants of the proposed fully unsupervised model as described in section 3: MinV+infGibbs and MinV+infEM. We report the average performance of our models over 5 runs. Performing multiple runs is necessary because we have a sampling process. They are compared with three baseline unsupervised models: Sporleder and Li (2009), Li and Sporleder (2009)[5] and Fazly et al. (2009); and two baseline supervised models: Rajani et al. (2014) and Liu and Hwa (2017) (using 5-fold cross validation).

Parameter setting Recall that in order to build the literal usage representation of an idiom, we need to sample s sentences that contain each constituent word w from an external corpus; extract from them the top n most frequently co-occurring words with w; then separately find t words that are semantically similar to w using word embeddings. To set parameters with values in reasonable ranges, we evaluated MinV on a small development set. We picked 10 idioms that are different from the evaluation set, scraped 50 instances from the web for each idiom, and labeled them ourselves. We find that $s >= 100$, n=10, and t=5 yield good results.

We use the gensim toolkit (Řehůřek and Sojka, 2010) and train our word embedding model using the continuous bag of word model on Text8 Corpus[6]. Negative sampling is applied as the training method; the min_count is set to 2. For the other parameters, we use the default settings in gensim.

Evaluative Data Our goal is to compare all the methods under two public available corpora: SemEval 2013 Task 5B corpus (Korkontzelos et al., 2013), which is used by prior supervised methods (Liu and Hwa, 2017; Rajani et al., 2014) and verb–noun combination (VNC) dataset (Cook et al., 2008), which is used by a prior unsupervised method (Fazly et al., 2009). However, there are some methods-datasets conflicts that have to be resolved. Because the idioms in the SemEval dataset are all in their canonical forms, and because the idioms are not restricted to the verb-noun combination, we cannot evaluate the method by Fazly et al. on this dataset (as their method is tailored to verb-noun combination). Some idioms from the VNC dataset are almost always used figuratively (or literally), which presents a problem for supervised methods. To facilitate full comparisons, we select the subset of idioms from the VNC corpus whose number of literal and figurative instances are both higher than 10. A summary of the two corpora is shown in Table 1. Note that each instance in SemEval corpus has about 3∼5 sentences; for consistency, we use 3 sentences as the context: the sentence with the target idiom and two neighboring sentences.

Evaluation metric Following the convention in prior works, we report the F-score for the recognition of figurative usages and the overall accuracy.

[5] We replace Normalized Google Distance (NGD) with word embeddings to measure the semantic relatedness between words due to the query frequency restriction on the API of NGD.

[6] From http://mattmahoney.net/dc/text8.zip

	SemEval	VNC
# of Idiom	10	11
# of Literal	1185	239
# of Figurative	1186	470
Idiom Type	Mixture	Verb-Noun
Syntactic Form	Canonical	Mixture
Context Size	3 ∼ 5 Sentences	1 Sentence

Table 1: Statistics of the two corpora

4.2 The Performance of Our Full Models

Table 2 shows the result of our models and the other comparative methods. Our proposed models show consistent performance across the two corpora, outperforming the unsupervised baselines from Sporleder and Li (2009), Li and Sporleder (2009) and the supervised model from Rajani et al. (2014). Moreover, there is no statistical significance in the F-score difference between the supervised ensemble model from Liu and Hwa (2017) and our models.

On the VNC corpus, our models have comparable average scores as that of Fazly et al. (2009); our scores are more stable across different idioms. While the method of Fazly et al. is nearly perfect for some idioms (0.98 on ”take heart”), it performs poorly for others (e.g., 0.33 on ”pull leg”). Their algorithm has trouble with idioms whose canonical and non-canonical forms can appear frequently both in literal and figurative usages.

4.3 Effectiveness of MinV

The core of our approach is MinV, the literal usage metric we developed to generate soft labels to guide the unsupervised learning. This experiment examines its effectiveness by creating usage classifications directly from it (i.e., if MinV predicts a probability of >0.5, predict ”literal”). We compare MinV against two alternative heuristics.

MinV is based on two core ideas. First, if an idiom is used figuratively, we expect to see a big difference (low similarity scores) between its context and the semantic representation of idiom's literal usage. The idea is similar to that of Sporleder and Li (2009), but they relied on lexical chain instead of distributional semantics. Second, instead of choosing a predefined threshold to separate the raw semantic similarity scores, we select a different decision threshold for each idiom adaptively based on the distribution of the scores. So as an alternative, we compare MinV against a Fixed-Threshold heuristic that labels an instance as ”literal” if its raw score is higher than some global threshold (set to 0.346 based on development data).

In Table 3, we observe that Minv outperforms both Sporleder and Li's model as well as Fixed-Threshold, but using MinV by itself is not sufficient. It has great fluctuations, e.g., the F-Score for individual idioms varies from 0.43 to 0.88. Recall that MinV +infGibbs has a smaller fluctuation across different idioms in Table 2. These results suggest that the subsequent learning process is effective.

Through error analysis, we find two major factors contributing to the performance fluctuation. First, the context itself could be misleading. An error case of **play ball** by MinV is:

All 10-year-old Minnie Cruttwell wants to do is play with the boys , but the ***Football Association*** *are not playing ball. She is a* ***member*** *of a mixed* ***team*** *called Balham Blazers , but the FA say she must* ***join*** *a girls'* ***team*** *when she is 12.*

The context words in bold (which are related to the word ”ball”) mislead MinV to predict a ”literal” usage when it is actually a ”figurative” usage (since an organization such as the Football Association cannot literally *play ball*). Second, not all content words in the context are relevant for distinguishing the idiom's usage. A future direction is to prune contextual words more intelligently.

4.4 Integration of MinV into Learning

We have argued that an advantage of using a metric with a probabilistic interpretation instead of a binary class heuristic is that its scores can be incorporated into subsequent learning models as soft labels. In this set of experiments, we evaluate the impact of the metric on the learning methods. First, we consider unsupervised learning without input from the literal usage metric. We cluster the instances with the original Gibbs sampling and EM algorithms and then label the two clusters with the majority usage within the clusters. Second, we explore using the information from the literal usage metric as ”noisy gold standard” to perform supervised training on a nearest neighbors (NN) classifier. Specifically, the literal and figurative instances labeled by MinV with high confidence (top 30%) are used as example set. Then for each test instance, we calculate its cosine similarity (in feature space) to the literal and figurative example sets and assign the label of the closest set. We refer this model as MinV +NN.

		SemEval		VNC	
Type	Model	Avg. F_{fig}	Avg.Acc	Avg. F_{fig}	Avg.Acc
Unsupervised	Sporleder & Li	0.58* (0.42 $\sim$ 0.72)	0.52*(0.32 $\sim$ 0.7)	0.61* (0.46 $\sim$ 0.73)	0.57*(0.41 $\sim$ 0.75)
	Li & Sporleder	0.64* (0.41 $\sim$ 0.76)	0.62*(0.43 $\sim$ 0.71)	0.67* (0.48 $\sim$ 0.77)	0.66*(0.52 $\sim$ 0.77)
	Fazly et al.	-	-	0.73 (0.33 $\sim$ 0.98)	0.74 (0.35 $\sim$ 0.98)
Supervised	Rajani et al.	0.71* (0.54 $\sim$ 0.83)	0.75(0.67 $\sim$ 0.81)	0.69* (0.49 $\sim$ 0.8)	0.7*(0.6 $\sim$ 0.79)
	Liu & Hwa	0.77 (0.68 $\sim$ 0.85)	0.77(0.71 $\sim$ 0.85)	0.75 (0.65 $\sim$ 0.88)	0.75(0.67 $\sim$ 0.89)
Our Model	**MinV + infGibbs**	0.75 (0.64 $\sim$ 0.91)	0.74(0.63 $\sim$ 0.87)	0.73 (0.64 $\sim$ 0.86)	0.75(0.66 $\sim$ 0.83)
	MinV + infEM	0.73 (0.58 $\sim$ 0.88)	0.73(0.61 $\sim$ 0.85)	0.72 (0.62 $\sim$ 0.87)	0.72(0.6 $\sim$ 0.84)

Table 2: The performances of different models. Avg. F_{fig} denotes average figurative F-score, Avg.Acc denotes average accuracy. We report the range in the parenthesis. * indicates the difference is significant with our MinV+ infGibbs model at the 95% confidence level. Since the method from Fazly et al. (2009) restricted their experiment to VNC type, we only report their performance on the VNC corpus.

Model	Avg. F_{fig}	Avg.Acc
Fixed-Threshold	0.6 (0.23 $\sim$ 0.82)	0.62 (0.47 $\sim$ 0.83)
MinV	0.66 (0.43 $\sim$ 0.88)	0.65 (0.51 $\sim$ 0.89)
Sporleder & Li	0.59 (0.42 $\sim$ 0.73)	0.54(0.32 $\sim$ 0.75)

Table 3: A comparison of classifying by different heuristics. Results are averaged across all the idioms in the two corpora.

Model	Avg. F_{fig}	Avg.Acc
Gibbs	0.58 (0.31 $\sim$ 0.78)	0.57 (0.4 $\sim$ 0.78)
EM	0.56 (0.31 $\sim$ 0.71)	0.6 (0.42 $\sim$ 0.77)
MinV+NN	0.68 (0.41 $\sim$ 0.83)	0.67 (0.55 $\sim$ 0.86)

Table 4: The performance of MinV+NN and models without soft label on all the idioms in the two corpora

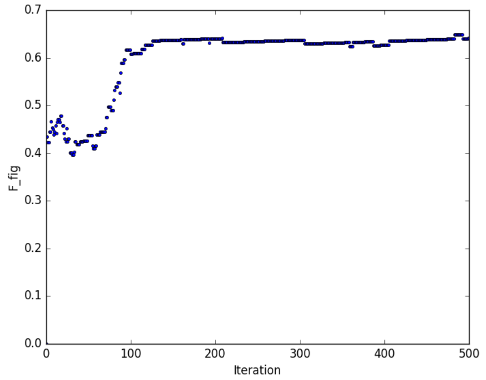

Figure 2: The performance of MinV+infGibbs on the idiom ”break a leg”

Table 4 shows the performances of the new models, which are all worse than our full models MinV +infGibbs and MinV +infEM. This highlights the advantage of integrating distributional semantic information and local features into one single learning procedure. Without the informed prior (encoded by the soft labels), the Gibbs sampling and EM algorithms only seek to maximize the probability of the observed data, and may fail to learn the underlying usage structure.

The model MinV +NN is not as competitive as our full models. It is too sensitive to the selected instances. Even though the training examples are instances that MinV is the most confident about, there are still mislabelled instances. These ”noisy training examples” would lead the NN classifier to make unreliable predictions. In contrast, our unsupervised learning is less sensitive to the performance of MinV; it can achieve a decent performance for an idiom even when the quality of the soft labels is poor. For example, when using MinV as a stand-alone model for *break a leg*, its figurative F-score is only 0.43, but through further training, the full model MinV+infGibbs achieves 0.64. Fig. 2 shows the training curve. A possible reason for this phenomenon is that the soft label is integrated into the learning process by biasing the sampling procedure (see Equation 3). We only encourage our model to follow the distributional semantic evidence captured by soft label and do not force it. So if there are strong evidences encoded by the linguistically motivated features in the instances to overcome the soft label it still has the freedom to do so.

5 Conclusion

We have presented an unsupervised method for idiom usage recognition built upon the heuristic that instances that use the idiom literally are semantically closer to constituent words of the idiom. Experimental results on two different corpora suggest that our models are competitive against supervised methods and prior unsupervised methods.

References

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Lorna Byrne, Caroline Fenlon, and John Dunnion. 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).

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

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset.

Graff David, Kong Junbo, Chen Ke, and Maeda Kazuaki. 2005. English gigaword second edition ldc2005t12. *Linguistic Data Consortium*.

Rong En Fan, Kai Wei Chang, Cho Jui Hsieh, Xiang Rui Wang, and Chih Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

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

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370.

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

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Florentina T. Hristea. 2013. *The Naïve Bayes Model in the Context of Word Sense Disambiguation*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *EMNLP*, pages 315–323.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. In *AAAI*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying idiomatic and literal expressions using topic models and intensity of emotions. *EMNLP*, pages 2019–2027.

Nazneen Fatema Rajani, Edaena Salinas, and Raymond Mooney. 2014. Using abstract context to detect figurative language.

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

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *EACL*, pages 754–762.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *EMNLP*, pages 680–690.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.