# Unsupervised Classification of Verb Noun Multi-Word Expression Tokens

Mona T. Diab and Madhav Krishna

Columbia University, New York, NY 10115
mdiab@ccls.columbia.edu, madhkrish@gmail.com

**Abstract.** We address the problem of classifying multiword expression tokens in running text. We focus our study on Verb-Noun Constructions (VNC) that vary in their idiomaticity depending on context. VNC tokens are classified as either idiomatic or literal. Our approach hinges upon the assumption that a literal VNC will have more **in common** with its component words than an idiomatic one. Commonality is measured by contextual overlap. To this end, we set out to explore different contextual variations and different similarity measures. We also identify a new data set OPAQUE that comprises only non-decomposable VNC expressions. Our approach yields state of the art performance with an overall accuracy of 77.56% on a TEST data set and 81.66% on the newly characterized data set OPAQUE.

## 1 Introduction

A Multi-Word Expression (MWE), for our purposes, can be defined as a multiword unit that refers to a single concept, for example - *kick the bucket*, *spill the beans*, *make a decision*, etc. An MWE typically has an idiosyncratic meaning that is *more* or *different* than the meaning of its component words. An MWE meaning is transparent, i.e. predictable, in as much as the component words in the expression relay the meaning portended by the speaker compositionally. Accordingly, MWEs vary in their degree of meaning compositionality; compositionality is correlated with the level of idiomaticity. An MWE is compositional if the meaning of an MWE as a unit can be predicted from the meaning of its component words such as in *make a decision* meaning *to decide.* If we conceive of idiomaticity as being a continuum, the more idiomatic an expression, the less transparent and the more non-compositional it is. Some MWEs are more predictable than others, for instance, *kick the bucket*, when used idiomatically to mean *to die*, has nothing in common with the literal meaning of either *kick* or *bucket*, however, *make a decision* is very clearly related to *to decide.* Both of these expressions are considered MWEs but have varying degrees of compositionality and predictability.

MWEs are pervasive in natural language, especially in the ever more abundant web based texts and speech genres. Identifying MWEs and understanding their meaning is essential to language understanding, hence they are of crucial importance for any Natural Language Processing (NLP) applications that aim at

A. Gelbukh (Ed.): CICLing 2009, LNCS 5449, pp. 98–110, 2009.

handling language meaning and use. In fact, the seminal paper [1] refers to this problem as a *key* issue for the development of high-quality NLP applications.

[2] note that the majority of MWEs are verbal expressions, such as light verb constructions (LVC), verb particle constructions (VPC), and verb noun constructions (VNC). To date, most research has addressed the problem of MWE *type* classification for VNC expressions [3, 4, 5, 6, 7, 8], not *token* classification. For example: *he spilt the beans over the kitchen counter* is most likely a literal usage. This is given away by the use of the prepositional phrase *over the kitchen counter*, since it is plausable that beans could have *literally* been spilt on a location such as a kitchen counter. Most previous research would classify *spilt the beans* as idiomatic irrespective of usage. A recent study by [9] of 60 idiom MWE types concluded that almost half of them had clear literal meanings and over 40% of their usages in text were actually literal. Thus, it would be important for an NLP application such as machine translation, for example, when given a new token of an MWE, to be able to determine whether it is used idiomatically or not.

In this paper, we address the problem of MWE classification for verb-noun (VNC) token constructions in running text. We investigate the binary classification of an unseen VNC token expression as being either **Idiomatic** (IDM) or **Literal** (LIT). An IDM expression is certainly an MWE, however, the converse is not necessarily true. We attempt to handle the problem of *sparsity* for the purpose of MWE classification. We explore several vector similarity metrics. We exploit more linguistically oriented feature sets to model the VNC vector space. We evaluate our results against a standard data set from the study by [10]. We achieve state of the art performance in classifying VNC tokens as either literal (F-measure: $F_{\beta_1}$=0.69) or idiomatic ($F_{\beta_1}$=0.83), corresponding to an overall accuracy of 77.56%. Recognizing the gray zone in such a binary classification set up, another thrust of our work focuses on a new evaluation set we term OPAQUE. The OPAQUE set comprises MWEs that have a clear distinction between their idiomatic senses and their literal ones.

This paper is organized as follows: In Section 2 we describe our understanding of the various classes of MWEs in general. Section 3 is a summary of previous related research. Section 4 describes our approach. In Section 5 we present the details of our experiments. We discuss the results in Section 6. Finally, we conclude the paper with some future directions in Section 7.

## 2 Multi-Word Expressions

MWEs are typically not productive, though they allow for inflectional variation [1]. They have been conventionalized due to persistent use. MWEs can be classified based on their semantic types as follows:

*Idiomatic:* This category includes expressions that are semantically non-compositional. For example, these include fixed expressions such as *kingdom come, ad hoc,* and, *ins and outs.* Fixed expressions tend to be more or less frozen in form. Idiomatic expressions also include non-fixed expressions such as *break new ground, speak of the devil*, and *break the ice*. Non-fixed expressions may undergo inflectional variations and lexical insertions.

*Semi-idiomatic:* This class includes expressions that seem semantically non-compositional, yet their semantics are more or less transparent. This category consists of Light Verb Constructions (LVC) and Verb Particle Constructions (VPC). An example of an LVC is *make a living*. The verb *make* would *prefer* a physical entity as an argument [11]. Examples of VPCs are *write-up, call-up* and *phone-up*. The particle *up* is an aspectual modifier of the verb rather than a preposition.

*Non-Idiomatic:* This category includes expressions that are semantically compositional: Compound nominals such as *prime minister*, proper nouns such as *New York Yankees*, and collocations such as *machine translation*. These expressions are *statistically idiosyncratic*. For instance, *traffic light* is the most likely lexicalization of the concept and would occur more often in text than, say, *traffic regulator* or *vehicle light*.

## 3 Previous Related Work

Several researchers have addressed the problem of MWE classification [5, 12, 13], however the most similar work to ours is the research by [10] and [7].

Cook *et al.* [10] develop an unsupervised technique that classifies a token instance of a VNC expression as idiomatic if the similarity between its context vector and that of its idiomatic usages is higher than the similarity between its context vector and that of its literal usages. They define the vector dimensions in terms of the co-occurrence frequencies of 1000 most frequent content bearing words (nouns, verbs, adjectives, adverbs and determiners) in the corpus. A context vector for a VNC expression is defined in terms of the words in the sentence in which it occurs. They employ the cosine measure to estimate similarity between contextual vectors. They assume that every instance of an expression occurring in a certain *canonical* syntactic form is idiomatic, otherwise it is literal. This assumption holds for many cases of idiomatic usage since many of them are conventionalized, however in cases such as *spilt the beans on the counter top*, the expression would be misclassified as idiomatic since it does occur in a canonical form though the meaning in this case is literal. They estimate the context vectors of literal usages in two ways: by either using those for the 'non-canonical' forms of the expression, or by adding the co-occurrence vectors of the component words. Their method achieves an accuracy of 52.7% on a data set containing expression tokens used mostly in their literal sense, whereas it yields an accuracy of 82.3% on a data set in which most usages are idiomatic. Further, they report that a classifier that predicts the idiomatic label if an expression (token) occurs in a canonical form achieves an accuracy of 53.4% on the former data set (where the majority of the MWEs occur in their literal sense) and 84.7% on the latter data set (where the majority of the MWE instances are idiomatic). This indicates that these 'canonical' forms can still be used literally. They report an overall system performance accuracy of 72.4%. We note that the use of accuracy as a measure for this work is not the most appropriate since accuracy is a measure of error rather than correctness, hence we report F-measure in addition to accuracy (to be able to compare with previous work). Their work is similar to this paper in that they explore the VNC expressions at the token level, even though they notably use type characteristics when assigning a class label to a token expression.

Fazly and Stevenson [7] correlate compositionality with idiomaticity. They measure compositionality as a combination of two similarity values: firstly, the similarity (cosine similarity) between the context of a VNC and the contexts of its constituent words; secondly, the similarity between an expression's context and that of a verb that is morphologically related to the noun in the expression, for instance, *decide* for *make a decision*. Context $context(t)$ of an expression or a word, $t$, is defined as a vector of the frequencies of nouns co-occurring with $t$ within a window of ±5 words. The resulting compositionality measure yields an $F_{\beta=1}$=0.51 on identifying literal expressions and $F_{\beta=1}$=0.42 on identifying idiomatic expressions. However, their results are not comparable to ours since it is type-based study.

## 4 Our Approach

Recognizing the significance of contextual information in MWE token classification, we explore the space of contextual modeling for the task of classifying the token instances of VNC expressions into literal versus idiomatic expressions. Inspired by works of [7, 12], our approach is to compare the context vector of a VNC, as an MWE, with the composed vector of the verb and noun (V-N) component units of the VNC when they occur in isolation of each other (i.e., not as a VNC). For example, in the case of the MWE *kick the bucket*, we compare the contexts of the instances of the VNC *kick the bucket* against the combined contexts for the verb (V) *kick*, independent of the noun *bucket*, and the contexts for the noun (N) *bucket*, independent of the verb *kick*. The intuition is that if there is a high similarity between the VNC and the combined V and N (namely, the V-N vector) contexts, then the VNC token is compositional, hence a literal instance of the MWE, otherwise the VNC token is idiomatic.

Previous work, [7], restricted context to within the boundaries of the sentences in which the tokens of interest occurred. We take a cue from that work but define '$context(t)$' as a vector with dimensions as **all** word types occurring in the same sentence as $t$, where $t$ is a verb type corresponding to the V in the VNC, noun type corresponding to N in the VNC, or VNC expression instance. Moreover our definition of context includes all nouns, verbs, adjectives and adverbs occurring in the same paragraph as $t$. This broader notion of context should help reduce sparseness effects, simply by enriching the vector with more contextual information. Further, we realize the importance of some closed class words occurring in the vicinity of $t$. [10] report the importance of determiners in identifying idiomaticity. Prepositions too should be informative of idiomaticity (or literal usage) as illustrated above in *spill the beans over the kitchen counter*. Hence, determiners and prepositions occurring in the same sentence as $t$ are also included in its context. The composed V-N contextual vector combines the co-occurrence of the verb type (aggregation of all the verb token instances in the whole corpus) as well as the noun type with this predefined set of dimensions. The VNC contextual vector is that for a specific instance of a VNC expression. Our objective is to find the best experimental settings that could yield the most accurate classification of VNC expression tokens. To that end, we explore the space of possible parameter variation. We experiment with five different parameter settings: the extent of context considered to model the vectors; the context vector dimensions for both V-N and VNC; the context content type; the vector similarity measure; and the method for combining the vectors for the verb type and the noun type to create the V-N composed contextual vector. Throughout the description below, a token ($T$) of interest could be a VNC, a (N)oun or a (V)erb. These parameters are detailed as follows:

*Context-Extent:* The definition of context for $T$ is restricted to: $Context_{Broad}$ comprises all the open class or *content* words (nouns, verbs, adjectives and adverbs) as well as the determiners and prepositions in the sentence containing $T$, in addition to the content words from the paragraph surrounding $T$. $Context_{Narrow}$ comprises all the open class words as well as the prepositions and determiners for the same sentence as $T$.

*Dimension:* This parameter is varied in three ways: $Dimension_{NoThresh}$ includes all the words that co-occur with $T$ in the specified Context-Extent. $Dimension_{Freq}$ sets a threshold on the co-occurrence frequency for the words to include in the dimensions thereby reducing the dimensionality of the vectors. $Dimension_{Ratio}$ is inspired by the utility of the *tf-idf* measure in information retrieval, we devise a threshold scheme that takes into consideration the salience of the word in context as a function of its relative frequency. Hence the raw frequencies of the words in context are converted to a ratio of two probabilities as per equation (1).

$$ratio = \frac{p(word|context)}{p(word)} = \frac{\frac{freq(word\ in\ context)}{freq(context)}}{\frac{freq(word\ in\ corpus)}{N}} \tag{1}$$

In equation (1), $N$ is the number of words (tokens) in the corpus and $freq(context)$ is the number of *context*s for a specific $T$ occurs. The numerator of the ratio is the probability that the word occurs in a particular context. The denominator is the probability of occurrence of the word in the corpus. Here, more weight is placed on words that are frequent in a certain context but rarer in the entire corpus. In case of the V and N contexts, a suitable threshold, which is independent of data size, is determined on this ratio in order to prune context words.

The latter two pruning techniques, $Dimension_{Freq}$ and $Dimension_{Ratio}$, are not performed for a VNC token's context, hence, all the words in the VNC token's contextual window are included. These thresholding methods are only applied to V-N vectors.

*Context-Content:* This parameter had two settings: words as they occur in the corpus, $Context-Content_{Words}$; or some of the words are collapsed into named entities, $Context-Content_{Words+NER}$. $Context-Content_{Words+NER}$ attempts to perform dimensionality reduction and sparsity reduction by collapsing named entities in the context of the VNC as well as those in the context of the V-N vectors. The intuition is that if we reduce the dimensions in semantically salient ways we will not adversely affect performance.

We employ BBN's *IdentiFinder* Named Entity Recognition (NER) System[1]. The NER system reduces all proper names, months, days, dates and times to NE tags. NER tagging is done on the corpus before the context vectors are extracted. For our purposes, it is not important that ***John*** *kicked the bucket on* ***Friday*** *in* ***New York City*** – neither the specific actor of the action, nor the place where is occurs is of relevance. The sentence ***PERSON*** *kicked the bucket on* ***DAY*** *in* ***PLACE*** conveys the same amount of information.

*IdentiFinder* identifies 24 NE types. We deem 5 of these inaccurate based on our observation, and exclude them. We retain 19 NE types: *Animal, Contact Information, Disease, Event, Facility, Game, Language, Location (merged with Geo-political Entity), Nationality, Organization, Person, Product, Date, Time, Quantity, Cardinal, Money, Ordinal* and *Percentage.* The written-text portion of the BNC contains 6.4M named entities in 5M sentences (at least one NE per sentence). The average number of words per NE is 2.56, the average number of words per sentence is 18.36. Thus, we estimate that by using NER, we reduce vector dimensionality by at least 14% without introducing the negative effects of sparsity.

*V-N Combination:* In order to create a single vector from the units of a VNC expression, we need to combine the vectors pertaining to the verb type (V) and the noun type (N). After combining the word types in the vector dimensions, we need to handle their co-occurrence frequency values. Hence we have two methods: *addition* where we simply add the frequencies in the cases of the shared dimensions which amounts to a union where the co-occurrence frequencies are added; or *multiplication* which amounts to an intersection of the vector dimensions where the co-occurrence frequencies are multiplied, hence giving more weight to the shared dimensions than in the *addition* case. In a study by [14] on a sentence similarity task, a multiplicative combination model performs better than the additive one.

*Similarity Measures:* We experiment with several standard similarity measures: Cosine Similarity, Overlap similarity, Dice Coefficient and Jaccard Index, as

[1] http://www.bbn.com/technology/identifinder

defined in [15]. A context vector is converted to a set by using the dimensions of the vector as members of the set.

## 5 Experiments and Results

### 5.1 Data

We use the British National Corpus (BNC),[2] which contains 100M words, because it draws its text from a wide variety of domains and the existing gold standard data sets are derived from it. The BNC contains multiple genres including written text and transcribed speech. We only experiment with the written-text portion. We syntactically parse the corpus with the *Minipar*[3] parser in order to identify all VNC expression tokens in the corpus. We exploit the lemmatized version of the text in order to reduce dimensionality and sparseness.

The standard data used in [10] (henceforth CFS07) is derived from a set comprising 2920 unique VNC-Token expressions drawn from the whole BNC. In this set, VNC token expressions are manually annotated as *idiomatic*, *literal* or *unknown*. The annotators were presented with the sentence that contained the VNC token only. The *unknown* class was used only in cases when the context did not seem enough to discern idiomaticity. The 2920 VNC expressions correspond to 53 VNC expression types, 28 of which have $\sim$ 60% of their token instances labeled idiomatic while $\sim$ 40% are labeled literal. The remaining 25 VNC expression types (corresponding to 1309 VNC tokens) are skewed in their distribution, almost all instances of a given expression are either idiomatic or literal.

For our purposes, we discard 127 of the 2920 token gold standard data set either because they are derived from the speech transcription portion of the BNC, or because *Minipar* could not find them. Similar to the CFS07 set, we exclude expressions labeled *unknown* by the annotators or pertaining to the skewed data set. Therefore, our resulting data set comprises 1125 VNC token expressions (CFS07 has 1180). We then split them into a development (DEV) set and a test (TEST) set. The DEV set comprises 564 token expressions corresponding to 346 idiomatic (IDM) expressions and 218 literal (LIT) ones (CFS07 dev has 573). The TEST set comprises 561 token expressions corresponding to 356 IDM expression tokens and 205 LIT ones (CFS07 test has 607). There is a complete overlap in types between our DEV and CFS07's dev set and our TEST and CFS07's test set. They each comprise 14 VNC type expressions with no overlap in type between the TEST and DEV sets. That means that the techniques developed and tested to address MWE problem in both our work and the work of CFS07 are robust and generalizable since no tuning of parameters is done on any VNC types that are present in the TEST data. We divide the tokens between the DEV and TEST maintaining the same proportions recommended in CFS07. Our DEV set has 61.5% while CFS07 has 60.9% idiomatic expressions. Our TEST set has 63.7% idiomatic expressions compared to 63.3% reported in CFS07. Even though

[2] http://www.natcorp.ox.ac.uk/

[3] http://www.cs.ualberta.ca/ lindek/minipar.htm

**Table 1.** VNC Expression types in OPAQUE data set

*move goalpost, pull weight, pull leg, make hay, hit roof, hold horse, blow smoke, kick heel, get sack, give sack, blow whistle, blow trumpet, get drift, get wind*

the number of instances is less in our TEST set, we believe that the results are generally comparable with those obtained by CFS07.

Following the intuition that idiomaticity is not a binary property, we create a new test data set, OPAQUE. The OPAQUE data set comprises those expressions that are at the high idiomaticity extreme of the spectrum. An opaque expression is one whose *idiomatic* meaning is highly non-compositional; [1] have called such expressions 'non-decomposable'. For instance, the expression *kick the bucket* is non-decomposable as its idiomatic meaning is completely unrelated to its literal meaning. On the other hand, *make a face*, though idiomatic, is decomposable and transparent to a certain degree.

To this end, we create a set of OPAQUE expressions from the VNC gold standard data set identified in work by [9]. The OPAQUE set comprises 14 VNC expression types listed in Table 1. The set was created in the following manner. All 53 VNC expression types in the gold set are judged by two annotators independently according to two diagnostics: if the verb and noun in the VNC expression are not indicative of their idiomatic meaning; if the idiomatic and literal meaning of the expression are completely distinct.[4] The resulting set of 14 expressions is the intersection between the two annotators, i.e. 100% agreement between the two annotators. Five of the OPAQUE expressions overlap with the skewed set in the gold standard, i.e. they are not in either our DEV or TEST sets: *hold horse, blow smoke, get drift, give sack*, and *move goalpost.*

The opaque set comprises 428 tokens (224 literal and 204 idiomatic) corresponding to 9 VNC types from the DEV and TEST sets, in addition to the 5 VNC types added from the the skewed data set. In our evaluation, we exclude the opaque expressions that come from the DEV set and the skewed data set, leaving only 282 expressions. Accordingly, the final OPAQUE set includes 142 idiomatic and 140 literal tokens. In order to maintain the 61-64% ratio (idiomatic to total number of tokens) as in CFS07, we employ a *bootstrapping* scheme. With the number of idiomatic tokens fixed at 142, 83 literal tokens are selected at random from the set of 140 literal expressions to form a data set containing 225 tokens with the ratio of IDM expressions to total number of tokens being 63.1%. This random selection process is repeated 1000 times. The results reported below are obtained after averaging over a 1000 trials.

### 5.2 Experimental Set-Up

We vary four of the experimental parameters: Context-Extent, Context-Content, Dimension and V-N compositionality, to create 9 experimental conditions. In the

[4] All meanings are looked up in the Cambridge Dictionaries Online, http://dictionary.cambridge.org/

following experiments, the thresholds (for $Dimension_{Freq}$ and $Dimension_{Ratio}$) are tuned on all the similarity measures collectively. It is observed that the performance of all the measures improved/worsened together, illustrating the same trends in performance, over the various settings of the thresholds evaluated on the DEV data set. Once the thresholds are tuned using the DEV set, they are applied to the TEST and OPAQUE data sets with no further tuning. Optimal thresholds for the similarity measures are also tuned on DEV. We note that different experimental conditions warranted different frequency and ratio thresholds. The experimental conditions are detailed as follows:

**nT-A-W-N:** The Dimension parameter is set to $Dimension_{NoThresh}$ (nT) and the V-N compositionality is addition (A). Context-Content is set to $Context-Content_{Words}$ (W) and Context-Extent is set to $Context_{Narrow}$ (N).
**nT-M-W-N:** The Dimension parameter is set to $Dimension_{NoThresh}$ (nT), and the V-N compositionality used is multiplication (M). Context-Content is set to $Context-Content_{Words}$ (W) and Context-Extent is set to $Context_{Narrow}$ (N).
**F-M-W-N:** $Dimension_{Freq}$ (F) is set to a threshold on the raw co-occurrence frequency of a word with the V-N composed vector. The optimal threshold is determined empirically on the DEV set to be 175. Multiplicative compositionality (M) is used, and Context-Content is set to $Context-Content_{Words}$ (W). Context-Extent is set to $Context_{Narrow}$ (N).
**R-M-W-N:** The Dimension parameter is set to the Ratio $Dimension_{Ratio}$ (R). The ratio is from Equation (1) is used and its optimal threshold value is 27 by tuning on the DEV set. Multiplicative compositionality mode for the V-N vectors (M), and the Context-Content is set to $Context-Content_{Words}$ (W). Context-Extent is set to $Context_{Narrow}$ (N).
**F-A-W-N:** The Dimension parameter $Dimension_{Freq}$ (F) is set to a threshold on the raw co-occurrence frequency of a word with the V-N composed vector. The optimal threshold is determined empirically on the DEV set to be 200. The V-N compositionality mode used is addition (A), and the Context-Content is set to $Context-Content_{Words}$ (W). Context-Extent is set to $Context_{Narrow}$ (N).
**R-A-W-N:** The Dimension parameter is set to the Ratio $Dimension_{Ratio}$ (R). An optimal threshold value for the ratio is determined as 75 based empirically on the DEV set. The V-N compositionality mode is addition (A), and the Context-Content is set to $Context-Content_{Words}$ (W). Again, Context-Extent is set to $Context_{Narrow}$ (N).
**R-A-W-B:** The parameter settings are the same as R-A-W-N except for Context-Extent which is set to $Context_{Broad}$ (B). The optimal threshold for the ratio is 265.
**R-M-W-B:** The parameter settings are the same as R-M-W-N except for Context-Extent which is set to $Context_{Broad}$ (B). The optimal threshold for the ratio is 150.
**R-A-NE-B:** Similar to the R-A-W-B experimental set up except that the Context-Content is set to $Context-Content_{Words+NER}$ (NE). The optimal value for the threshold on the ratio values is 275.

**Table 2.** Evaluation on of different experimental conditions on DEV

| Experiment | Dice Coefficient | | | Jaccard Index | | | Overlap | | | Cosine | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F-measure | | Accuracy | F-measure | | Accuracy | F-measure | | Accuracy | F-measure | | Accuracy |
| | IDM | LIT | | IDM | LIT | | IDM | LIT | | IDM | LIT | |
| nT-A-W-N | 0.45 | 0.44 | 44.39% | 0.47 | 0.43 | 44.92% | **0.50** | **0.56** | **53.30%** | 0.49 | 0.42 | 45.63% |
| nT-M-W-N | 0.48 | 0.46 | 46.88% | 0.48 | 0.46 | 46.88% | **0.58** | **0.57** | **57.78%** | 0.46 | 0.47 | 46.52% |
| F-M-W-N | 0.48 | 0.48 | 47.77% | 0.48 | 0.48 | 47.59% | **0.58** | **0.57** | **57.75%** | 0.49 | 0.49 | 49.19% |
| R-M-W-N | 0.63 | 0.60 | 61.50% | 0.63 | 0.60 | 61.50% | **0.72** | **0.63** | **68.45%** | 0.65 | 0.61 | 63.10% |
| F-A-W-N | 0.48 | 0.48 | 47.95% | 0.48 | 0.48 | 47.95% | **0.54** | **0.53** | **53.65%** | 0.52 | 0.52 | 51.69% |
| R-A-W-N | 0.66 | 0.63 | 64.17% | 0.66 | 0.63 | 64.17% | **0.78** | **0.68** | **73.80%** | 0.76 | 0.64 | 71.12% |
| R-M-W-B | 0.50 | 0.61 | 56.33% | 0.62 | 0.61 | 61.50% | **0.84** | **0.73** | **79.68%** | 0.66 | 0.66 | 65.78% |
| R-A-W-B | 0.70 | 0.63 | 67.20% | 0.70 | 0.63 | 67.20% | **0.84** | **0.76** | **81.10%** | 0.77 | 0.69 | 73.62% |
| R-A-NE-B | 0.72 | 0.66 | 69.05% | 0.72 | 0.66 | 69.05% | **0.85** | **0.75** | **81.22%** | 0.78 | 0.71 | 75.00% |

### 5.3 Results

We use $F_{\beta=1}$ (F-measure) which is the harmonic mean between precision and recall, as well as accuracy to report the results. We report the results separately for the two classes IDM and LIT on all the data sets, DEV, TEST and OPAQUE. As mentioned above, throughout the experiments, all the thresholds are tuned on the DEV set. The tables below illustrate the results obtained using all four similarity measures.

## 6 Discussion

As shown in Table 3, we obtain a classification accuracy of 77.56% (R-A-W-N) on TEST using the Overlap similarity measure, with $F_{\beta=1}$ values for the IDM and LIT classes being 0.83 and 0.69, respectively. These results are comparable to state-of-the-art results obtained by CFS07 who report an overall system accuracy of 72.4% on their test set. Hence, we improve over state-of-the-art results by 5.16% absolute. Even if we compare the results yielded by the Cosine measure (as this was the measure used in CFS07, we note an increase of 4.22% absolute improvement at an accuracy of 76.66% for our classification approach. It is worth noting that the differences among all possible pairs of mean F-measure and accuracy values (within each experiment), except for the cases when the Dice and Jaccard measures perform equivalently, are found to be statistically significant.

The highest accuracy figures across all experimental conditions are obtained using the overlap similarity measure across all three data sets. It is also worth noting that for each similarity measure, the highest accuracy values are associated with the highest F-measure performance on IDM and LIT classification. Moreover, in those conditions, our system is always yielding better performance of identifying IDM expressions than literal expressions with significantly large difference in performance. Contrary to previous work, we note that the Cosine similarity is outperformed by the Overlap measure.

Comparing the different experimental conditions, results suggest that $Dimension_{Ratio}$ outperforms $Dimension_{Freq}$ and $Dimension_{NoThresh}$ within all data sets. We recognize that in the $Dimension_{Ratio}$, we vary the ratio threshold value depending on experimental condition which might render the results

**Table 3.** Evaluation of different experimental conditions on TEST

| Experiment | Dice Coefficient | | | Jaccard Index | | | Overlap | | | Cosine | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F-measure | | Accuracy | F-measure | | Accuracy | F-measure | | Accuracy | F-measure | | Accuracy |
| | IDM | LIT | | IDM | LIT | | IDM | LIT | | IDM | LIT | |
| nT-A-W-N | 0.58 | 0.48 | 53.50% | 0.62 | 0.49 | 56.37% | 0.43 | 0.50 | 46.32% | **0.63** | **0.48** | **56.37%** |
| nT-M-W-N | 0.58 | 0.46 | 52.60% | 0.53 | 0.48 | 50.45% | 0.53 | 0.50 | 51.71% | **0.55** | **0.51** | **52.78%** |
| F-M-W-N | 0.56 | 0.48 | 52.06% | 0.56 | 0.48 | 52.06% | 0.50 | 0.44 | 47.04% | **0.60** | **0.51** | **47.04%** |
| R-M-W-N | 0.64 | 0.61 | 62.48% | 0.64 | 0.61 | 62.48% | **0.72** | **0.61** | **68.04%** | 0.68 | 0.62 | 65.35% |
| F-A-W-N | **0.57** | **0.46** | **52.24%** | **0.57** | **0.46** | **52.24%** | 0.44 | 0.39 | 41.29% | 0.57 | 0.45 | 51.53% |
| R-A-W-N | 0.73 | 0.65 | 69.30% | 0.73 | 0.65 | 69.30% | **0.83** | **0.69** | **77.56%** | 0.82 | 0.66 | 76.66% |
| R-M-W-B | 0.51 | 0.60 | 55.54% | 0.64 | 0.58 | 61.07% | **0.80** | **0.64** | **74.46%** | 0.66 | 0.60 | 63.04% |
| R-A-W-B | 0.69 | 0.57 | 64.11% | 0.69 | 0.57 | 64.11% | **0.82** | **0.66** | **76.07%** | 0.76 | 0.61 | 70.00% |
| R-A-NE-B | 0.70 | 0.58 | 64.93% | 0.70 | 0.58 | 64.93% | **0.83** | **0.65** | **76.62%** | 0.76 | 0.62 | 70.86% |

**Table 4.** Evaluation of different experimental conditions on OPAQUE

| Experiment | Dice Coefficient | | | Jaccard Index | | | Overlap | | | Cosine | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F-measure | | Accuracy | F-measure | | Accuracy | F-measure | | Accuracy | F-measure | | Accuracy |
| | IDM | LIT | | IDM | LIT | | IDM | LIT | | IDM | LIT | |
| nT-A-W-N | 0.51 | 0.53 | 52.09% | 0.53 | 0.52 | 52.40% | 0.48 | 0.48 | 47.96% | **0.55** | **0.52** | **53.45%** |
| nT-M-W-N | 0.29 | 0.46 | 38.87% | 0.29 | 0.48 | 39.70% | **0.73** | **0.56** | **66.45%** | 0.32 | 0.50 | 42.57% |
| F-M-W-N | 0.34 | 0.50 | 42.71% | 0.34 | 0.50 | 42.71% | **0.66** | **0.40** | **56.88%** | 0.44 | 0.50 | 47.47% |
| R-M-W-N | 0.69 | 0.65 | 67.22% | 0.69 | 0.65 | 67.22% | **0.77** | **0.68** | **72.85%** | 0.69 | 0.65 | 66.97% |
| F-A-W-N | 0.41 | 0.37 | 38.65% | 0.41 | 0.37 | 38.65% | 0.46 | 0.23 | 36.15% | **0.44** | **0.38** | **41.18%** |
| R-A-W-N | 0.71 | 0.66 | 68.42% | 0.71 | 0.66 | 68.42% | **0.85** | **0.75** | **81.10%** | 0.80 | 0.72 | 76.76% |
| R-M-W-B | 0.24 | 0.54 | 42.69% | 0.56 | 0.54 | 54.95% | **0.79** | **0.70** | **75.00%** | 0.46 | 0.58 | 52.72% |
| R-A-W-B | 0.58 | 0.60 | 58.82% | 0.58 | 0.60 | 58.82% | **0.83** | **0.76** | **80.23%** | 0.69 | 0.66 | 67.83% |
| R-A-NE-B | 0.61 | 0.61 | 61.06% | 0.61 | 0.61 | 61.06% | **0.86** | **0.76** | **81.66%** | 0.70 | 0.65 | 67.71% |

not directly comparable across the different $R$ conditions in the same data set. We would argue that the results are directly comparable however, since *Ratio* as characterized by our definition is a relative threshold that will have to depend on the other parameters, for example, using *addition* warrants a very different ratio threshold from using *multiplication*, therefore it is more condition dependent. Hence we can grossly compare across conditions that apply dimensionality reduction using *some* ratio threshold. However, we emphasize that the results are directly comparable across data sets with the same condition.

Accordingly, for vector compositionality for the V-N vector, *addition* clearly outperforms *multiplication* for the task of MWE classification. This indicates that union is better than intersection for combining the V and N vectors for this task. *multiplication* seems to increase vector sparsity. We note that $Context_{Narrow}$ does better than $Context_{Broad}$ on the TEST and OPAQUE data sets, though this is clearly the opposite in the DEV data set where the $Context_{Broad}$ conditions outperform their $Context_{Narrow}$ counterparts, R-A-W-B yields better results than R-A-W-N. This may indicate that the parameter tuned for the $Context_{Broad}$ conditions on DEV is overfitted for the DEV data set. Comparing results (accuracy, and F-measure on IDM and LIT) using $Context - Content_{Words}$ versus $Context - Content_{Words+NER}$, in R-A-W-B against R-A-NE-B, we note that in all data sets, $Context - Content_{Words+NER}$ is closely comparable or even outperforms using $Context - Content_{Words}$ across all similarity measures. This strongly suggests that dimensionality reduction using NER has a significant positive impact on MWE classification.

The best performing condition for the DEV and OPAQUE data is R-A-NE-B across all similarity measures. However, this does not hold for the TEST data set. For the latter data set, we note that R-A-W-N yields the best performance for all the measures followed closely by R-A-NE-B. These results suggest that R-A-W-N and R-A-NE-B are the best experimental conditions for classification. R-A-W-N is not the 2nd best condition for all similarity measures, in the case of DEV. The variation in experimental results between the DEV, TEST and OPAQUE sets may be attributed to the fact that the tuning parameters are tuned on the DEV data and that there are no shared MWE types between the DEV and OPAQUE and TEST data sets.

The highest yielded results are obtained on the OPAQUE data set, at an accuracy of 81.66%, an IDM F-measure classification of 86%, and a LIT F-measure of 76% in the R-A-NE-B experimental condition using overlap similarity. These results are even higher than those obtained in the best performing condition on the DEV set. These results are significantly higher than those obtained on the best condition of the TEST set. This suggests that the OPAQUE set indeed has a naturally clear distinction between idiomatic and literal usages of MWE expressions.

## 7 Conclusion

In this study, we explored a set of features that contribute to VNC token expression binary classification. We applied dimensionality reduction heuristics inspired by information retrieval (*tf-idf* like ratio measure) and linguistics (named-entity recogniiton). These contributions improve significantly over experimental conditions that do not manipulate context and dimensions. Our system achieves state-of-the-art performance on a set that is very close to a standard data set. Different from previous studies, we classify VNC token expressions in context. We include function words in modeling the VNC token contexts as well as using the whole paragraph in which it occurs as context. We also designate a new data set, OPAQUE, that reflects the more non-decompositional aspect of VNC MWEs' idiomaticity. The results suggest that our approach is able to reliably capture the discriminatory features for MWE classification. As expected the results on the OPAQUE set outperform the results yielded on the TEST set due to the clear separability of the idiomatic senses from the literal ones for the VNC tokens. Further, these results reaffirm the notion that idiomaticity is not a discrete binary property.

## References

1. Sag, I.A., Baldwin, T., Bond, F., Copestake, A.A., Flickinger, D.: Multiword expressions: A pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002)
2. Villavicencio, A., Copestake, A.: On the nature of idioms. In: LinGO Working Paper No. 2002-2004 (2002)

3. Melamed, D.I.: Automatic discovery of non-compositional compounds in parallel data. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP 1997), Providence, RI, USA, pp. 97–108 (1997)
4. Lin, D.: Automatic identification of noncompositional phrases. In: Proceedings of ACL 1999, Univeristy of Maryland, College Park, Maryland, USA, pp. 317–324 (1999)
5. Baldwin, T., Bannard, C., Tanaka, T., Widdows, D.: An empirical model of multiword expression decomposability. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions, Morristown, NJ, USA, pp. 89–96 (2003)
6. na Villada Moirón, B., Tiedemann, J.: Identifying idiomatic expressions using automatic word-alignment. In: Proceedings of the EACL 2006 Workshop on Multiword Expressions in a Multilingual Context, Morristown, NJ, USA, pp. 33–40 (2006)
7. Fazly, A., Stevenson, S.: Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, Czech Republic, Association for Computational Linguistics, pp. 9–16 (2007)
8. Van de Cruys, T., Villada Moirón, B.n.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, Czech Republic, Association for Computational Linguistics, pp. 25–32 (2007)
9. Cook, P., Fazly, A., Stevenson, S.: The VNC-Tokens Dataset. In: Proceedings of the LREC Workshop on Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco (2008)
10. Cook, P., Fazly, A., Stevenson, S.: Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, Czech Republic, Association for Computational Linguistics, pp. 41–48 (2007)
11. Resnik, P.: Selectional preference and sense disambiguation. In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, Washington, DC, USA (1997)
12. Katz, G., Giesbrecht, E.: 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, Sydney, Australia, Association for Computational Linguistics, pp. 12–19 (2006)
13. Schone, P., Juraksfy, D.: Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In: Proceedings of Empirical Methods in Natural Language Processing, Pittsburg, PA, USA, pp. 100–108 (2001)
14. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: Proceedings of ACL 2008: HLT, Columbus, Ohio, Association for Computational Linguistics, pp. 236–244 (2008)
15. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)