Handling Sparsity for Verb Noun MWE Token Classification

Mona T. Diab
Center for Computational Learning Systems
Columbia University
mdiab@ccls.columbia.edu

Madhav Krishna
Computer Science Department
Columbia University
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 handling the sparsity in the possible contexts via four different parameter variations. Our approach yields state of the art performance with an overall accuracy of 75.54% on a TEST data set.

1 Introduction

A Multi-Word Expression (MWE), for our purposes, can be defined as a multi-word 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.

MWEs are pervasive in natural language, especially in 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 handling robust language meaning and use.

To date, most research has addressed the problem of MWE *type* classification for VNC expressions in English (Melamed, 1997; Lin, 1999; Baldwin et al., 2003; na Villada Moirón and Tiedemann, 2006; Fazly and Stevenson, 2007; Van de Cruys and Villada Moirón, 2007; McCarthy et al., 2007), 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 (Cook et al., 2008) of 60 idiom MWE types concluded that almost half of them had clear literal meaning 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 MWE token, 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 handle the problem of *sparsity* for MWE classification by exploring different vector space features: various vector similarity metrics, and more linguistically oriented feature sets. We evaluate our results against a standard data set from the study by (Cook et al., 2007). We achieve state of the art performance in classifying VNC tokens as either literal (F-measure: F_{β_1}=0.64) or idiomatic (F_{β_1}=0.82), corresponding to an overall accuracy of 75.54%.

This paper is organized as follows: In Section

Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, pages 96–103,
Athens, Greece, 31 March 2009.

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 in Section 7.

2 Multi-word Expressions

MWEs are typically not productive, though they allow for inflectional variation (Sag et al., 2002). 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, *fixed expressions* such as *kingdom come, ad hoc*, *non-fixed expressions* such as *break new ground, speak of the devil*. **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) such as *make a living* and Verb Particle Constructions (VPC) such as *write-up, call-up*. **Non-Idiomatic**: This category includes expressions that are semantically compositional such as *prime minister*, proper nouns such as *New York Yankees*.

3 Previous Related Work

Several researchers have addressed the problem of MWE classification (Baldwin et al., 2003; Katz and Giesbrecht, 2006; Schone and Juraksfy, 2001; Hashimoto et al., 2006; Hashimoto and Kawahara, 2008). The majority of the proposed research has been using unsupervised approaches and have addressed the problem of MWE type classification irrespective of usage in context. Only, the work by Hashimoto et al. (2006) and Hashimoto and Kawahara (2008) addressed token classification in Japanese using supervised learning.

The most comparable work to ours is the research by (Cook et al., 2007) and (Fazly and Stevenson, 2007). On the other hand, (Cook et al., 2007) develop an unsupervised technique that classifies a VNC expression as idiomatic or literal. They examine if the similarity between the context vector of the MWE, in this case the VNC, and that of its idiomatic usage is higher than the similarity between its context vector and that of its literal usage. 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 the canonical form though the meaning in this case is literal. Their work is similar to this paper in that they explore the VNC expressions at the token level. 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%.[1]

(Fazly and Stevenson, 2007) correlate compositionality with idiomaticity. They measure compositionality as a combination of two similarity values: firstly, similar to (Katz and Giesbrecht, 2006), 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.

[1] 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.

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 (Katz and Giesbrecht, 2006; Fazly and Stevenson, 2007), our approach is to compare the context vector of a VNC 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, (Fazly and Stevenson, 2007), 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. (Cook et al., 2007) 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 on the kitchen counter*. Hence, we include determiners and prepositions occurring in the same sentence as t. 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 taking into consideration the sparsity problem. To that end, we explore the space of possible parameter variation on the vectors representing our tokens of interest (VNC, V, or N). We experiment with five different parameter settings:

Context-Extent The definition of context is broad or narrow described as follows. Both $Context_{Broad}$ and $Context_{Narrow}$ comprise all the open class or *content* words (nouns, verbs, adjectives and adverbs), determiners, and prepositions in the **sentence** containing the token. Moreover, $Context_{Broad}$, additionally, includes the content words from the **paragraph** in which the token occurs.

Dimension This is a pruning parameter on the words included from the Context Extent. The intuition is that salient words should have a bigger impact on the calculation of the vector similarity. This parameter is varied in three ways: $Dimension_{NoThresh}$ includes all the words that co-occur with the token under consideration 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 the following equation.

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

where N is the number of words (tokens) in the corpus and $freq(context)$ is the number of $context$s for a specific token of interest 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 indepen-

dent 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 composed vectors obtained from the combination of the verb and noun 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. 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[2]. 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 (Mitchell and Lapata, 2008) 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 defined in (Manning and Schütze, 1999). 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),[3] 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*[4] 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 (Cook et al., 2007) (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*.

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 parse them. Similar to the CFS07 set, we exclude expressions labeled *unknown* or pertaining

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

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

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

to the skewed data set as deemed by the annotators. 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. We divide the tokens between the DEV and TEST maintaining the same proportions of IDM to LIT as recommended in CFS07: DEV is 61.5% and TEST is 63.7%.

5.2 Experimental Set-up

We vary four of the experimental parameters: Context-Extent {sentence only narrow (N), sentence + paragraph broad(B)}, Context-Content {Words (W), Words+NER (NE)}, Dimension {no threshold (nT), frequency (F), ratio (R)}, and V-N compositionality {Additive (A), Multiplicative (M)}. We present the results for all similarity measures. 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. Based on tuning on the DEV set, we empirically set the value of the threshold on F to be 188 and for R to be 175 across all experimental conditions. We present results here for 10 experimental conditions based on the four experimental parameters: {**nT-A-W-N, nT-M-W-N, F-A-W-N, F-M-W-N, R-A-W-N, R-M-W-N, R-A-W-B, R-M-W-B, R-A-NE-B, R-M-NE-B**}. For instance, **R-A-W-N**, the Dimension parameter is set to the Ratio $Dimension_{Ratio}$ (R), the V-N compositionality mode is addition (A), and the Context-Content is set to $Context - Content_{Words}$ (W), and, Context-Extent is set to $Context_{Narrow}$ (N).

5.3 Results

We use $F_{\beta=1}$ (F-measure) as 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 the DEV and TEST data set for all four similarity measures.

6 Discussion

As shown in Table 2, we obtain the best classification accuracy of 75.54% (R-A-NE-B) on TEST using the Overlap similarity measure, with $F_{\beta=1}$ values for the IDM and LIT classes being 0.82 and 0.64, respectively. These results are generally 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 3% absolute.

In the DEV set, the highest results (F-measures for IDM and LIT, as well as accuracy scores) are obtained for all conditions consistently using the Overlap similarity measure. We also note that our approach tends to fare better overall in classifying IDM than LIT. The best performance is obtained in experimental setting **R-A-NE-B** at 78.53% accuracy corresponding to an IDM classification F-measure of 0.83 and LIT classification F-measure of 0.71.

In the TEST set, we note that Overlap similarity yields the highest overall results, however inconsistently across all the experimental conditions. The highest scores are yielded by the same experimental condition R-A-NE-B. In fact, comparable to previous work, the Cosine similarity measure significantly outperforms the other similarity measures when the Dimension parameter is set to no threshold (nT) and with a set threshold on frequency (F). However, Cosine is outperformed by Overlap when we apply a threshold to the Ratio Dimension. It is worth noting that across all experimental conditions (except in one case, **nT-A-W-N** using Overlap similarity), IDM F-measures are consistently higher than LIT F-measures, suggesting that our approach is more reliable in detecting idiomatic VNC MWE rather than not.

The overall results strongly suggest that using intelligent dimensionality reduction, such as a threshold on the ratio, significantly outperforms no thresholding (nT) and simple frequency thresholding (F) comparing across different similarity measures and all experimental conditions. Recall that R was employed to maintain the salient signals in the context and exclude those that are irrelevant.

Experiment	Dice Coefficient			Jaccard Index			Overlap			Cosine		
	F-measure		Acc. %	F-measure		Acc. %	F-measure		Acc. %	F-measure		Acc. %
	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-A-W-N	0.47	0.47	46.70	0.47	0.47	46.70	0.58	0.53	**55.62**	0.50	0.50	50.09
F-M-W-N	0.48	0.49	48.31	0.48	0.49	48.31	0.58	0.57	**57.40**	0.54	0.50	52.05
R-A-W-N	0.79	0.62	72.73	0.79	0.62	72.73	0.79	0.63	**73.44**	0.79	0.62	72.73
R-M-W-N	0.76	0.06	62.21	0.76	0.06	62.21	0.77	0.06	**62.39**	0.77	0.06	**62.39**
R-A-W-B	0.59	0.57	58.11	0.59	0.57	58.11	0.80	0.72	**76.47**	0.67	0.65	65.78
R-M-W-B	0.67	0.63	65.06	0.67	0.63	65.06	0.80	0.71	**76.65**	0.71	0.66	68.81
R-A-NE-B	0.58	0.58	58.14	0.58	0.58	58.14	0.83	0.71	**78.53**	0.70	0.64	67.08
R-M-NE-B	0.63	0.63	62.79	0.63	0.63	62.79	0.76	0.69	**73.17**	0.73	0.67	70.13

Table 1: Evaluation on of different experimental conditions on DEV

Experiment	Dice Coefficient			Jaccard Index			Overlap			Cosine		
	F-measure		Acc. %	F-measure		Acc. %	F-measure		Acc. %	F-measure		Acc. %
	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-A-W-N	0.60	0.48	55.12	0.60	0.48	55.12	0.46	0.36	41.47	0.60	0.46	**54.04**
F-M-W-N	0.56	0.48	52.07	0.56	0.48	52.07	0.49	0.45	47.04	0.62	0.49	**56.19**
R-A-W-N	0.81	0.57	73.61	0.81	0.57	73.61	0.82	0.57	**74.51**	0.81	0.57	73.61
R-M-W-N	0.78	0.09	**64.99**	0.78	0.09	**64.99**	0.78	0.08	64.81	0.78	0.08	64.81
R-A-W-B	0.69	0.57	64.11	0.62	0.56	59.11	0.78	0.66	**73.04**	0.68	0.60	64.64
R-M-W-B	0.64	0.60	61.79	0.64	0.60	61.79	0.78	0.64	**72.86**	0.69	0.62	65.89
R-A-NE-B	0.61	0.56	58.45	0.61	0.56	58.45	0.82	0.64	**75.54**	0.68	0.58	63.37
R-M-NE-B	0.59	0.58	58.63	0.59	0.58	58.63	0.76	0.65	**71.40**	0.69	0.61	65.29

Table 2: Evaluation of different experimental conditions on TEST

The results suggest some interaction between the vector combination method, A or M, and the Dimensionality pruning parameters. Experimental conditions that apply the multiplicative compositionality on the component vectors V and N yield higher results in the nT and F conditions across all the similarity measures. Yet once we apply R dimensionality pruning, we see that the additive vector combination, A parameter setting, yields better results. This indicated that the M condition already prunes too much in addition to the R dimensionality hence leading to slightly lower performance.

For both DEV and TEST, we note that the R parameter settings coupled with the A parameter setting. For DEV, we observe that the results yielded from the Broad context extent, contextual sentence and surrounding paragraph, yield higher results than those obtained from the narrow N, context sentence only, across M and A conditions. This trend is not consistent with the results on the TEST data set. R-A-W-N, outperforms R-A-W-B, however, R-M-W-B outperforms R-M-W-N.

We would like to point out that R-M-W-N has very low values for the LIT F-measure, this is attributed to the use of a unified R threshold value of 175. We experimented with different optimal thresholds for R depending on the parameter setting combination and we discovered that for R-M-W-N, the fine-tuned optimal threshold should have been 27 as tuned on the DEV set, yielding LIT F-measures of 0.68 and 0.63, for DEV and TEST, respectively. Hence when using the unified value of 175, more of the compositional vectors components of V+N are pruned away leading to similarity values between the V+N vector and the VNC vector of 0 (across all similarity measures). Accordingly, most of the expressions are

mis-classified as IDM.

The best results overall are yielded from the NE conditions. This result strongly suggests that using class based linguistic information and novel ways to keep the relevant tokens in the vectors such as R yields better MWE classification.

Qualitatively, we note the best results are obtained on the following VNCs from the TEST set in the Overlap similarity measure for the **R-A-W-B** experimental setting (percentage of tokens classified correctly): *make hay(94%), make mark(88%), pull punch (86%), have word(81%), blow whistle (80%), hit wall (79%), hold fire (73%)*. While we note the highest performance on the following VNCs in the corresponding **R-A-NE-B** experimental setting: *make hay(88%), make mark(87%), pull punch (91%), have word(85%), blow whistle (84%), hold fire (82%)*. We observe that both conditions performed the worse on tokens from the following VNCs *lose thread, make hit*. Especially, *make hit* is problematic since it mostly a literal expression, yet in the gold standard set we see it marked inconsistently. For instance, the literal sentence *He bowled it himself and Wilfred Rhodes* **made** *the winning* **hit** while the following annotates *make hit* as idiomatic: *It was the TV show Saturday Night Live which originally* **made** *Martin a huge hit in the States*.

We also note the difference in performance in the hard cases of VNCs that are relatively transparent, only the **R-A-W-B and R-A-NE-B** experimental conditions were able to classify them correctly with high F-measures as either IDM or LIT, namely: *have word, hit wall, make mark*. For **R-A-W-B**, the yielded accuracies are 81%, 79% and 88% respectively, and for **R-A-NE-B**, the accuracies are 85%, 65%, and 87%, respectively. However, in the **nT-A-W-N** condition *have word* is classified incorrectly 82% of the time and in **F-A-W-N** it is classified incorrectly 85% of the time. *Make mark* is classified incorrectly 77% of the time, *make hay* (77%) and *hit wall* (57%) in the **F-A-W-N** experimental setting. This may be attributed to the use of the Broader context, or the use of R in the other more accurate experimental settings.

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. Moreover we empirically show that the Overlap similarity measure is a better measure to use for MWE classification.

8 Acknowledgement

The first author was partially funded by DARPA GALE and MADCAT projects. The authors would like to acknowledge the useful comments by three anonymous reviewers who helped in making this publication more concise and better presented.

References

Timothy Baldwin, Collin Bannard, Takakki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 89–96, Morristown, NJ, USA.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. 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*, pages 41–48, Prague, Czech Republic, June. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop on Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, June.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 992–1001, Hon-

olulu, Hawaii, October. Association for Computational Linguistics.

Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006. Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 353–360, Sydney, Australia, July. Association for Computational Linguistics.

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, Sydney, Australia, July. Association for Computational Linguistics.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324, Univeristy of Maryland, College Park, Maryland, USA.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, June.

Diana McCarthy, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379, Prague, Czech Republic, June. Association for Computational Linguistics.

Dan I. Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*, pages 97–108, Providence, RI, USA, August.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June. Association for Computational Linguistics.

Bego na Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL-06 Workshop on Multiword Expressions in a Multilingual Context*, pages 33–40, Morristown, NJ, USA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15, London, UK. Springer-Verlag.

Patrick Schone and Daniel Juraksfy. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, pages 100–108, Pittsburg, PA, USA.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.