Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context

Paul Cook and Afsaneh Fazly and Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, Canada
{pcook,afsaneh,suzanne}@cs.toronto.edu

Abstract

Much work on idioms has focused on type identification, i.e., determining whether a sequence of words can form an idiomatic expression. Since an idiom type often has a literal interpretation as well, token classification of potential idioms in context is critical for NLP. We explore the use of informative prior knowledge about the overall syntactic behaviour of a potentially-idiomatic expression (type-based knowledge) to determine whether an instance of the expression is used idiomatically or literally (token-based knowledge). We develop unsupervised methods for the task, and show that their performance is comparable to that of state-of-the-art supervised techniques.

1 Introduction

Identification of multiword expressions (MWEs), such as *car park*, *make a decision*, and *kick the bucket*, is extremely important for accurate natural language processing (NLP) (Sag et al., 2002). Most MWEs need to be treated as single units of meaning, e.g., *make a decision* roughly means "decide". Nonetheless, the components of an MWE can be separated, making it hard for an NLP system to identify the expression as a whole. Many researchers have recently developed methods for the automatic acquisition of various properties of MWEs from corpora (Lin, 1999; Krenn and Evert, 2001; Baldwin et al., 2003; McCarthy et al., 2003; Venkatapathy and Joshi, 2005; Villada Moirón and Tiedemann, 2006; Fazly and Stevenson, 2006). These studies look into properties, such as the collocational behaviour of MWEs, their semantic non-compositionality, and their lexicosyntactic fixedness, in order to distinguish them from similar-on-the-surface literal combinations.

Most of these methods have been aimed at recognizing MWE types; less attention has been paid to the identification of instances (tokens) of MWEs in context. For example, most such techniques (if successful) would identify *make a face* as a potential MWE. This expression is, however, ambiguous between an idiom, as in *The little girl made a funny face at her mother*, and a literal combination, as in *She made a face on the snowman using a carrot and two buttons*. Despite the common perception that phrases that can be idioms are mainly used in their idiomatic sense, our analysis of 60 idioms has shown otherwise. We found that close to half of these idioms also have a clear literal meaning; and of the expressions with a literal meaning, on average around 40% of their usages are literal. Distinguishing token phrases as MWEs or literal combinations of words is thus essential for NLP applications that require the identification of multiword semantic units, such as semantic parsing and machine translation.

Recent studies addressing MWE token classification mainly perform the task as one of word sense disambiguation, and draw on the local context of an expression to disambiguate it. Such techniques either do not use any information regarding the linguistic properties of MWEs (Birke and Sarkar, 2006), or mainly focus on their non-compositionality (Katz and Giesbrecht, 2006). Pre-

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 41–48,
Prague, June 2007. ©2007 Association for Computational Linguistics

vious work on the identification of MWE types, however, has found other properties of MWEs, such as their syntactic fixedness, to be relevant to their identification (Evert et al., 2004; Fazly and Stevenson, 2006). In this paper, we propose techniques that draw on this property to classify individual tokens of a potentially idiomatic phrase as literal or idiomatic. We also put forward classification techniques that combine such information with evidence from the local context of an MWE.

We explore the hypothesis that informative prior knowledge about the overall syntactic behaviour of an idiomatic expression (type-based knowledge) can be used to determine whether an instance of the expression is used literally or idiomatically (token-based knowledge). Based on this hypothesis, we develop unsupervised methods for token classification, and show that their performance is comparable to that of a standard supervised method.

Many verbs can be combined with one or more of their arguments to form MWEs (Cowie et al., 1983; Fellbaum, 2002). Here, we focus on a broadly documented class of idiomatic MWEs that are formed from the combination of a verb with a noun in its direct object position, as in *make a face*. In the rest of the paper, we refer to these verb+noun combinations, which are potentially idiomatic, as VNCs. In Section 2, we propose unsupervised methods that classify a VNC token as an idiomatic or literal usage. Section 3 describes our experimental setup, including experimental expressions and their annotation. In Section 4, we present a detailed discussion of our results. Section 5 compares our work with similar previous studies, and Section 6 concludes the paper.

2 Unsupervised Idiom Identification

We first explain an important linguistic property attributed to idioms—that is, their syntactic fixedness (Section 2.1). We then propose unsupervised methods that draw on this property to automatically distinguish between idiomatic and literal usages of an expression (Section 2.2).

2.1 Syntactic Fixedness and Canonical Forms

Idioms tend to be somewhat fixed with respect to the syntactic configurations in which they occur (Nunberg et al., 1994). For example, *pull one's weight* tends to mainly appear in this form when used idiomatically. Other forms of the expression, such as *pull the weights*, typically are only used with a literal meaning. In their work on automatically identifying idiom types, Fazly and Stevenson (2006)—henceforth FS06—show that an idiomatic VNC tends to have one (or at most a small number of) canonical form(s), which are its most preferred syntactic patterns. The preferred patterns can vary across different idiom types, and can involve a number of syntactic properties: the voice of the verb (active or passive), the determiner introducing the noun (*the*, *one's*, etc.), and the number of the noun (singular or plural). For example, while *pull one's weight* has only one canonical form, *hold fire* and *hold one's fire* are two canonical forms of the same idiom, as listed in an idiom dictionary (Seaton and Macaulay, 2002).

In our work, we assume that in most cases, idiomatic usages of an expression tend to occur in a small number of canonical form(s) for that idiom. We also assume that, in contrast, the literal usages of an expression are less syntactically restricted, and are expressed in a greater variety of patterns. Because of their relative unrestrictiveness, literal usages may occur in a canonical idiomatic form for that expression, but usages in a canonical form are more likely to be idiomatic. Usages in alternative syntactic patterns for the expression, which we refer to as the non-canonical forms of the idiom, are more likely to be literal. Drawing on these assumptions, we develop three unsupervised methods that determine, for each VNC token in context, whether it has an idiomatic or a literal interpretation.

2.2 Statistical Methods

The following paragraphs elaborate on our proposed methods for identifying the idiomatic and literal usages of a VNC: the CForm method that uses knowledge of canonical forms only, and two Diff methods that draw on further contextual evidence as well. All three methods draw on our assumptions described above, that usages in the canonical form for an idiom are more likely to be idiomatic, and those in other forms are more likely to be literal. Thus, for all three methods, we need access to the canonical form of the idiom. Since we want our token identification methods to be unsupervised, we adopt the

unsupervised statistical method of FS06 for finding canonical forms for an idiomatic VNC. This method determines the canonical forms of an expression to be those forms whose frequency is much higher than the average frequency of all its forms.

CForm: The underlying assumption of this method is that information about the canonical form(s) of an idiom type is extremely informative in classifying the meaning of its individual instances (tokens) as literal or idiomatic. Our CForm classifies a token as idiomatic if it occurs in the automatically determined canonical form(s) for that expression, and as literal otherwise.

Diff: Our two Diff methods combine local context information with knowledge about the canonical forms of an idiom type to determine if its token usages are literal or idiomatic. In developing these methods, we adopt a distributional approach to meaning, where the meaning of an expression is approximated by the words with which it co-occurs (Firth, 1957). Although there may be fine-grained differences in meaning across the idiomatic usages of an expression, as well as across its literal usages, we assume that the idiomatic and literal usages correspond to two coarse-grained senses of the expression. Since we further assume these two groups of usages will have more in common semantically within each group than between the two groups, we expect that literal and idiomatic usages of an expression will typically occur with different sets of words. We will refer then to each of the literal and idiomatic designations as a (coarse-grained) meaning of the expression, while acknowledging that each may have multiple fine-grained senses. Clearly, the success of our method depends on the extent to which these assumptions hold.

We estimate the meaning of a set of usages of an expression e as a word frequency vector $\vec{v}_e$ where each dimension i of $\vec{v}_e$ is the frequency with which e co-occurs with word i across the usages of e. We similarly estimate the meaning of a single token of an expression t as a vector $\vec{v}_t$ capturing that usage. To determine if an instance of an expression is literal or idiomatic, we compare its co-occurrence vector to the co-occurrence vectors representing each of the literal and idiomatic meanings of the expression. We use a standard measure of distributional similarity, cosine, to compare co-occurrence vectors.

In supervised approaches, such as that of Katz and Giesbrecht (2006), co-occurrence vectors for literal and idiomatic meanings are formed from manually-annotated training data. Here, we propose unsupervised methods for estimating these vectors. We use one way of estimating the idiomatic meaning of an expression, and two ways for estimating its literal meaning, yielding two methods for token classification.

Our first Diff method draws further on our expectation that canonical forms are more likely idiomatic usages, and non-canonical forms are more likely literal usages. We estimate the idiomatic meaning of an expression by building a co-occurrence vector, $\vec{v}_{I\text{-}CF}$, for all uses of the expression in its automatically determined canonical form(s). Since we hypothesize that idiomatic usages of an expression tend to occur in its canonical form, we expect these co-occurrence vectors to be largely representative of the idiomatic usage of the expression. We similarly estimate the literal meaning by constructing a co-occurrence vector, $\vec{v}_{L\text{-}NCF}$, of all uses of the expression in its non-canonical forms. We use the term $\text{Diff}_{I\text{-}CF,L\text{-}NCF}$ to refer to this method.

Our second Diff method also uses the vector $\vec{v}_{I\text{-}CF}$ to estimate the idiomatic meaning of an expression. However, this approach follows that of Katz and Giesbrecht (2006) in assuming that literal meanings are compositional. The literal meaning of an expression is thus estimated by composing (summing and then normalizing) the co-occurrence vectors for its component words. The resulting vector is referred to as $\vec{v}_{L\text{-}Comp}$, and this method as $\text{Diff}_{I\text{-}CF,L\text{-}Comp}$.

For both Diff methods, if the meaning of an instance of an expression is determined to be more similar to its idiomatic meaning (e.g., $\text{cosine}(\vec{v}_t, \vec{v}_{I\text{-}CF}) > \text{cosine}(\vec{v}_t, \vec{v}_{L\text{-}NCF})$), then we label it as an idiomatic usage. Otherwise, it is labeled as literal.[1]

[1]We also performed experiments using a KNN classifier in which the co-occurrence vector for a token was compared against the co-occurrence vectors for the canonical and non-canonical forms of that expression, which were assumed to be idiomatic and literal usages respectively. However, performance was generally worse using this method.

Note that all three of our proposed techniques for token identification depend on how accurately the canonical forms of an expression can be acquired. FS06's canonical form acquisition technique, which we use here, works well if the idiomatic usage of a VNC is sufficiently frequent compared to its literal usage. In our experiments, we examine the performance of our proposed classification methods for VNCs with different proportions of idiomatic-to-literal usages.

3 Experimental Setup

3.1 Experimental Expressions and Annotation

We use data provided by FS06, which consists of a list of VNCs and their canonical forms. From this data, we discarded expressions whose frequency in the British National Corpus[2] (BNC) is lower than 20, in an effort to make sure that there would be literal and idiomatic usages of each expression. The frequency cut-off further ensures an accurate estimate of the vectors representing each of the literal and idiomatic meanings of the expression. We also discarded expressions that were not found in at least one of two dictionaries of idioms (Seaton and Macaulay, 2002; Cowie et al., 1983). This process resulted in the selection of 60 candidate expressions.

For each of these 60 expressions, 100 sentences containing its usage were randomly selected from the automatically parsed BNC (Collins, 1999), using the automatic VNC identification method described by FS06. For an expression which occurs less than 100 times in the BNC, all of its usages were extracted. Our primary judge, a native English speaker and an author of this paper, then annotated each use of each candidate expression as one of literal, idiomatic, or unknown. When annotating a token, the judge had access to only the sentence in which it occurred, and not the surrounding sentences. If this context was insufficient to determine the class of the expression, the judge assigned the unknown label.

Idiomaticity is not a binary property, rather it is known to fall on a continuum from completely semantically transparent, or literal, to entirely opaque, or idiomatic. The human annotators were required to pick the label, literal or idiomatic, that best fit the usage in their judgment; they were not to use the unknown label for intermediate cases. Figurative extensions of literal meanings were classified as literal if their overall meaning was judged to be fairly transparent, as in *You turn right when we **hit the road** at the end of this track* (taken from the BNC). Sometimes an idiomatic usage, such as *had words* in *I was in a bad mood, and he kept pestering me, so we **had words***, is somewhat directly related to its literal meaning, which is not the case for more semantically opaque idioms such as *hit the roof*. The above sentence was classified as idiomatic since the idiomatic meaning is much more salient than the literal meaning.

Based on the primary judge's annotations, we removed expressions with fewer than 5 instances of either of their literal or idiomatic meanings, leaving 28 expressions. The remaining expressions were then split into development (DEV) and test (TEST) sets of 14 expressions each. The data was divided such that DEV and TEST would be approximately equal with respect to the frequency, and proportion of idiomatic-to-literal usages, of their expressions. Before consensus annotation, DEV and TEST contained a total of 813 and 743 tokens, respectively.

A second human judge, also a native English-speaking author of this paper, then annotated DEV and TEST. The observed agreement and unweighted kappa score on TEST were 76% and 0.62 respectively. The judges discussed tokens on which they disagreed to achieve a consensus annotation. Final annotations were generated by removing tokens that received the unknown label as the consensus annotation, leaving DEV and TEST with a total of 573 and 607 tokens, and an average of 41 and 43 tokens per expression, respectively.

3.2 Creation of Co-occurrence Vectors

We create co-occurrence vectors for each expression in our study from counts in the BNC. We form co-occurrence vectors for the following items.

- Each token instance of the target expression
- The target expression in its automatically determined canonical form(s)
- The target expression in its non-canonical form(s)

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

- The verb in the target expression
- The noun in the target expression

The co-occurrence vectors measure the frequency with which the above items co-occur with each of 1000 *content bearing words* in the same sentence.[3] The content bearing words were chosen to be the most frequent words in the BNC which are used as a noun, verb, adjective, adverb, or determiner. Although determiners are often in a typical stoplist, we felt it would be beneficial to use them here. Determiners have been shown to be very informative in recognizing the idiomaticity of MWE types, as they are incorporated in the patterns used to automatically determine canonical forms (Fazly and Stevenson, 2006).[4]

3.3 Evaluation and Baseline

Our baseline for comparison is that of always predicting an idiomatic label, the most frequent class in our development data. We also compare our unsupervised methods against the supervised method proposed by Katz and Giesbrecht (2006). In this study, co-occurrence vectors for the tokens were formed from uses of a German idiom manually annotated as literal or idiomatic. Tokens were classified in a leave-one-out methodology using k-nearest neighbours, with $k = 1$. We report results using this method (1NN) as well as one which considers a token's 5 nearest neighbours (5NN). In all cases, we report the accuracy macro-averaged across the experimental expressions.

4 Experimental Results and Analysis

In Section 4.1, we discuss the overall performance of our proposed unsupervised methods. Section 4.2 explores possible causes of the differences observed in the performance of the methods. We examine our estimated idiomatic and literal vectors, and compare them with the actual vectors calculated from manually-annotated data. Results reported in Sections 4.1 and 4.2 are on TEST (results on DEV have very similar trends). Section 4.3 then examines the performance of the unsupervised methods on expressions with different proportions of idiomatic-to-literal usages. This section presents results on TEST and DEV combined, as explained below.

Method		$\%Acc$	($\%RER$)
Baseline		61.9	-
Unsupervised	Diff $_{I\text{-}CF,\ L\text{-}Comp}$	67.8	(15.5)
	Diff $_{I\text{-}CF,\ L\text{-}NCF}$	70.1	(21.5)
	CForm	**72.4**	(**27.6**)
Supervised	1NN	72.4	(27.6)
	5NN	**76.2**	(**37.5**)

Table 1: Macro-averaged accuracy ($\%Acc$) and relative error reduction ($\%RER$) over TEST.

4.1 Overall Performance

Table 4.1 shows the macro-averaged accuracy on TEST of our three unsupervised methods, as well as that of the baseline and the two supervised methods for comparison (see Section 3.3). The best supervised performance and the best unsupervised performance are indicated in boldface. As the table shows, all three unsupervised methods outperform the baseline, confirming that the canonical forms of an expression, and local context, are both informative in distinguishing literal and idiomatic instances of the expression.

The table also shows that Diff $_{I\text{-}CF,\ L\text{-}NCF}$ performs better than Diff $_{I\text{-}CF,\ L\text{-}Comp}$. This suggests that estimating the literal meaning of an expression using the non-canonical forms is more accurate than using the composed vector, $\vec{v}_{L\text{-}Comp}$. In Section 4.2 we find more evidence for this. Another interesting observation is that CForm has the highest performance (among unsupervised methods), very closely followed by Diff $_{I\text{-}CF,\ L\text{-}NCF}$. These results confirm our hypothesis that canonical forms—which reflect the overall behaviour of a VNC type—are strongly informative about the class of a token, perhaps even more so than the local context of the token. Importantly, this is the case even though the canonical forms that we use are imperfect knowledge obtained automatically through an unsupervised method.

Our results using 1NN, 72.4%, are comparable

[3] We also considered 10 and 20 word windows on either side of the target expression, but experiments on development data indicated that using the sentence as a window performed better.

[4] We employed singular value decomposition (Deerwester et al., 1990) to reduce the dimensionality of the co-occurrence vectors. This had a negative effect on the results, likely because information about determiners, which occur frequently with many expressions, is lost in the dimensionality reduction.

Vectors	cosine	Vectors	cosine
$\vec{a}_{idm}$ and $\vec{a}_{lit}$	.55		
$\vec{v}_{I\text{-}CF}$ and $\vec{a}_{lit}$	.70	$\vec{v}_{I\text{-}CF}$ and $\vec{a}_{idm}$	.90
$\vec{v}_{L\text{-}NCF}$ and $\vec{a}_{lit}$	.80	$\vec{v}_{L\text{-}NCF}$ and $\vec{a}_{idm}$	.60
$\vec{v}_{L\text{-}Comp}$ and $\vec{a}_{lit}$	.72	$\vec{v}_{L\text{-}Comp}$ and $\vec{a}_{idm}$	.76

Table 2: Average similarity between the actual vectors ($\vec{a}$) and the estimated vectors ($\vec{v}$), for the idiomatic and literal meanings.

to those of Katz and Giesbrecht (2006) using this method on their German data (72%). However, their baseline is slightly lower than ours at 58%, and they only report results for 1 expression with 67 instances. Interestingly, our best unsupervised results are in line with the results using 1NN and not substantially lower than the results using 5NN.

4.2 A Closer Look into the Estimated Vectors

In this section, we compare our estimated idiomatic and literal vectors with the actual vectors for these usages calculated from manually-annotated data. Such a comparison helps explain some of the differences we observed in the performance of the methods. Table 4.2 shows the similarity between the estimated and actual vectors representing the idiomatic and literal meanings, averaged over the 14 TEST expressions. Actual vectors, referred to as $\vec{a}_{idm}$ and $\vec{a}_{lit}$, are calculated over idiomatic and literal usages of the expressions as determined by the human annotations. Estimated vectors, $\vec{v}_{I\text{-}CF}$, $\vec{v}_{L\text{-}CF}$, and $\vec{v}_{L\text{-}Comp}$, are calculated using our methods described in Section 2.2.

For comparison purposes, the first row of Table 4.2 shows the average similarity between the actual idiomatic and literal vectors, $\vec{a}_{idm}$ and $\vec{a}_{lit}$. These vectors are expected to be very dissimilar, hence the low average cosine between them serves as a baseline for comparison. We now look into the relative similarity of each estimated vector, $\vec{v}_{I\text{-}CF}$, $\vec{v}_{L\text{-}CF}$, $\vec{v}_{L\text{-}Comp}$, with these two vectors.

The second row of the table shows that, as desired, our estimated idiomatic vector, $\vec{v}_{I\text{-}CF}$, is notably more similar to the actual idiomatic vector than to the actual literal vector. Also, $\vec{v}_{L\text{-}NCF}$ is more similar to the actual literal vector than to the actual idiomatic vector (third row). Surprisingly, however, $\vec{v}_{L\text{-}Comp}$ is somewhat similar to both actual literal and idiomatic vectors (in fact it is slightly more similar to the latter). These results suggest that the vector composed of the context vectors for the constituents of an expression may not always be the best estimate of the literal meaning of the expression.[5] Given this observation, the overall better-than-baseline performance of $\text{Diff}_{I\text{-}CF,L\text{-}Comp}$ might seem unjustified at a first glance. However, we believe this performance is mainly due to an accurate estimate of $\vec{v}_{I\text{-}CF}$.

4.3 Performance Based on Class Distribution

We further divide our 28 DEV and TEST expressions according to their proportion of idiomatic-to-literal usages, as determined by the human annotators. In order to have a sufficient number of expressions in each group, here we merge DEV and TEST (we refer to the new set as DT). $\text{DT}_{\text{I}_{high}}$ contains 17 expressions with 65%–90% of their usages being idiomatic—i.e., their idiomatic usage is dominant. $\text{DT}_{\text{I}_{low}}$ contains 11 expressions with 8%–58% of their occurrences being idiomatic—i.e., their idiomatic usage is not dominant.

Table 4.3 shows the average accuracy of all the methods on these two groups of expressions, with the best performance on each group shown in boldface. On $\text{DT}_{\text{I}_{high}}$, both $\text{Diff}_{I\text{-}CF,\ L\text{-}NCF}$ and CForm outperform the baseline, with CForm having the highest reduction in error rate. The two methods perform similarly to each other on $\text{DT}_{\text{I}_{low}}$, though note that the error reduction of CForm is more in line with its performance on $\text{DT}_{\text{I}_{high}}$. These results show that even for VNCs whose idiomatic meaning is not dominant—i.e., those in $\text{DT}_{\text{I}_{low}}$—automatically-acquired canonical forms can help with their token classification.

An interesting observation in Table 4.3 is the inconsistent performance of $\text{Diff}_{I\text{-}CF,\ L\text{-}Comp}$: the method has a very poor performance on $\text{DT}_{\text{I}_{high}}$, but outperforms the other two unsupervised methods on $\text{DT}_{\text{I}_{low}}$. As we noted earlier in Section 2.2, the more frequent the idiomatic meaning of an expression, the more reliable the acquired canonical forms for that expression. Since the performance of CForm and $\text{Diff}_{I\text{-}CF,\ L\text{-}NCF}$ depends highly on the accuracy of the automatically acquired canonical forms, it is not surprising that these two methods perform

[5] This was also noted by Katz and Giesbrecht (2006) in their second experiment.

Method		$DT_{I_{high}}$	$DT_{I_{low}}$
Baseline		81.4 (-)	35.0 (-)
Unsupervised	$Diff_{I\text{-}CF,\ L\text{-}Comp}$	73.1 (-44.6)	58.6 (36.3)
	$Diff_{I\text{-}CF,\ L\text{-}NCF}$	82.3 (4.8)	52.7 (27.2)
	CForm	**84.7** (**17.7**)	53.4 (28.3)
Supervised	1NN	78.3 (-16.7)	65.8 (47.4)
	5NN	82.3 (4.8)	**72.4** (**57.5**)

Table 3: Macro-averaged accuracy over DEV and TEST, divided according to the proportion of idiomatic-to-literal usages.

worse than $Diff_{I\text{-}CF,\ L\text{-}Comp}$ on VNCs whose idiomatic usage is not dominant.

The high performance of the supervised methods on $DT_{I_{low}}$ also confirms that the poorer performance of the unsupervised methods on these VNCs is likely due to the inaccuracy of the canonical forms extracted for them. Interestingly, when canonical forms can be extracted with a high accuracy (i.e., for VNCs in $DT_{I_{high}}$) the performance of the unsupervised methods is comparable to (or even slightly better than) that of the best supervised method. One possible way of improving the performance of unsupervised methods is thus to develop more accurate techniques for the automatic acquisition of canonical forms.

5 Related Work

Various properties of MWEs have been exploited in developing automatic identification methods for MWE types (Lin, 1999; Krenn and Evert, 2001; Fazly and Stevenson, 2006). Much research has addressed the non-compositionality of MWEs as an important property related to their idiomaticity, and has used it in the classification of both MWE types and tokens (Baldwin et al., 2003; McCarthy et al., 2003; Katz and Giesbrecht, 2006). We also make use of this property in an MWE token classification task, but in addition, we draw on other salient characteristics of MWEs which have been previously shown to be useful for their type classification (Evert et al., 2004; Fazly and Stevenson, 2006).

The idiomatic/literal token classification methods of Birke and Sarkar (2006) and Katz and Giesbrecht (2006) rely primarily on the local context of a token, and fail to exploit specific linguistic properties of non-literal language. Our results suggest that such properties are often more informative than the local context, in determining the class of an MWE token.

The supervised classifier of Patrick and Fletcher (2005) distinguishes between compositional and non-compositional English verb-particle construction tokens. Their classifier incorporates linguistically-motivated features, such as the degree of separation between the verb and particle. Here, we focus on a different class of English MWEs, verb+noun combinations. Moreover, by making a more direct use of their syntactic behaviour, we develop unsupervised token classification methods that perform well. The unsupervised token classifier of Hashimoto et al. (2006) uses manually-encoded information about allowable and non-allowable syntactic transformations of Japanese idioms—that are roughly equivalent to our notions of canonical and non-canonical forms. The rule-based classifier of Uchiyama et al. (2005) incorporates syntactic information about Japanese compound verbs (JCVs), a type of MWE composed of two verbs. In both cases, although the classifiers incorporate syntactic information about MWEs, their manual development limits the scalability of the approaches.

Uchiyama et al. (2005) also propose a statistical token classification method for JCVs. This method is similar to ours, in that it also uses type-based knowledge to determine the class of each token in context. However, their method is supervised, whereas our methods are unsupervised. Moreover, Uchiyama et al. (2005) evaluate their methods on a set of JCVs that are mostly monosemous. Here, we intentionally exclude such cases from consideration, and focus on those MWEs that have two clear idiomatic and literal meanings, and that are frequently used with either meaning.

6 Conclusions

While a great deal of research has focused on properties of MWE types, such as their compositionality, less attention has been paid to issues surrounding MWE tokens. In this study, we have developed techniques for a semantic classification of tokens of a potential MWE in context. We focus on a broadly documented class of English MWEs that are formed from the combination of a verb and a noun in its direct object position, referred to as VNCs. We annotated a total of 1180 tokens for 28 VNCs accord-

ing to whether they are a literal or idiomatic usage, and we found that approximately 40% of the tokens were literal usages. These figures indicate that automatically determining whether a VNC token is used idiomatically or literally is of great importance for NLP applications. In this work, we have proposed three unsupervised methods that perform such a task. Our proposed methods incorporate automatically acquired knowledge about the overall syntactic behaviour of a VNC type, in order to do token classification. More specifically, our methods draw on the syntactic fixedness of VNCs—a property which has been largely ignored in previous studies of MWE tokens. Our results confirm the usefulness of this property as incorporated into our methods. All our methods outperform the baseline of always predicting the most frequent class. Moreover, considering our approach is unsupervised, our best accuracy of 72.4% is not substantially lower than the accuracy of a standard supervised approach at 76.2%.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-06*, 329–336.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Anthony P. Cowie, Ronald Mackin, and Isabel R. McCaig. 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2. Oxford University Press.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Stefan Evert, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings LREC-04*.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of EACL-06*, 337–344.

Christiane Fellbaum. 2002. VP idioms in the lexicon: Topics for research using a very large corpus. In S. Busemann, editor, *Proceedings of the KONVENS-02 Conference*.

John R. Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1–32. The Philological Society, Oxford.

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*, 353–360.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, 12–19.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL-01 Workshop on Collocations*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, 317–324.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Jon Patrick and Jeremy Fletcher. 2005. Classifying verb-particle constructions by verb arguments. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in Computational Linguistics Formalisms and Applications*, 200–209.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing-02*, 1–15.

Maggie Seaton and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, second edition.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):497–512.

Sriram Venkatapathy and Aravid Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of HLT/EMNLP-05*, 899–906.

Begoña 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*, 33–40.