Proceedings of the 11th Conference of the European Chapter of the ACL (EACL-2006).

Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations

Afsaneh Fazly
Department of Computer Science
University of Toronto
Toronto, ON M5S 3H5
Canada
afsaneh@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, ON M5S 3H5
Canada
suzanne@cs.toronto.edu

Abstract

We investigate the lexical and syntactic flexibility of a class of idiomatic expressions. We develop measures that draw on such linguistic properties, and demonstrate that these statistical, corpus-based measures can be successfully used for distinguishing idiomatic combinations from non-idiomatic ones. We also propose a means for automatically determining which syntactic forms a particular idiom can appear in, and hence should be included in its lexical representation.

1 Introduction

The term *idiom* has been applied to a fuzzy category with prototypical examples such as *by and large*, *kick the bucket*, and *let the cat out of the bag*. Providing a definitive answer for what idioms are, and determining how they are learned and understood, are still subject to debate (Glucksberg, 1993; Nunberg et al., 1994). Nonetheless, they are often defined as phrases or sentences that involve some degree of lexical, syntactic, and/or semantic idiosyncrasy.

Idiomatic expressions, as a part of the vast family of figurative language, are widely used both in colloquial speech and in written language. Moreover, a phrase develops its idiomaticity over time (Cacciari, 1993); consequently, new idioms come into existence on a daily basis (Cowie et al., 1983; Seaton and Macaulay, 2002). Idioms thus pose a serious challenge, both for the creation of wide-coverage computational lexicons, and for the development of large-scale, linguistically plausible natural language processing (NLP) systems (Sag et al., 2002).

One problem is due to the range of syntactic idiosyncrasy of idiomatic expressions. Some idioms, such as *by and large*, contain syntactic violations; these are often completely fixed and hence can be listed in a lexicon as “words with spaces” (Sag et al., 2002). However, among those idioms that are syntactically well-formed, some exhibit limited morphosyntactic flexibility, while others may be more syntactically flexible. For example, the idiom *shoot the breeze* undergoes verbal inflection (*shot the breeze*), but not internal modification or passivization (?*shoot the fun breeze*, ?*the breeze was shot*). In contrast, the idiom *spill the beans* undergoes verbal inflection, internal modification, and even passivization. Clearly, a words-with-spaces approach does not capture the full range of behaviour of such idiomatic expressions.

Another barrier to the appropriate handling of idioms in a computational system is their semantic idiosyncrasy. This is a particular issue for those idioms that conform to the grammar rules of the language. Such idiomatic expressions are indistinguishable on the surface from compositional (non-idiomatic) phrases, but a computational system must be capable of distinguishing the two. For example, a machine translation system should translate the idiom *shoot the breeze* as a single unit of meaning (“to chat”), whereas this is not the case for the literal phrase *shoot the bird*.

In this study, we focus on a particular class of English phrasal idioms, i.e., those that involve the combination of a verb plus a noun in its direct object position. Examples include *shoot the breeze*, *pull strings*, and *push one’s luck*. We refer to these as verb+noun idiomatic combinations (VNICs). The class of VNICs accommodates a large number of idiomatic expressions (Cowie et al., 1983; Nunberg et al., 1994). Moreover, their peculiar behaviour signifies the need for a distinct treatment in a computational lexicon (Fellbaum, 2005). Despite this, VNICs have been granted relatively little attention within the computational linguistics community.

We look into two closely related problems confronting the appropriate treatment of VNICs: (i) the problem of determining their degree of flexibility; and (ii) the problem of determining their level of idiomaticity. Section 2 elaborates on the lexicosyntactic flexibility of VNICs, and how this relates to their idiomaticity. In Section 3, we propose two linguistically-motivated statistical measures for quantifying the degree of lexical and syntactic inflexibility (or fixedness) of verb+noun combinations. Section 4 presents an evaluation of the proposed measures. In Section 5, we put forward a technique for determining the syntactic variations that a VNIC can undergo, and that should be included in its lexical representation. Section 6 summarizes our contributions.

2 Flexibility and Idiomaticity of VNICs

Although syntactically well-formed, VNICs involve a certain degree of semantic idiosyncrasy. Unlike compositional verb+noun combinations, the meaning of VNICs cannot be solely predicted from the meaning of their parts. There is much evidence in the linguistic literature that the semantic idiosyncrasy of idiomatic combinations is reflected in their lexical and/or syntactic behaviour.

2.1 Lexical and Syntactic Flexibility

A limited number of idioms have one (or more) lexical variants, e.g., *blow one's own trumpet* and *toot one's own horn* (examples from Cowie et al. 1983). However, most are lexically fixed (non-productive) to a large extent. Neither *shoot the wind* nor *fling the breeze* are typically recognized as variations of the idiom *shoot the breeze*. Similarly, *spill the beans* has an idiomatic meaning ("to reveal a secret"), while *spill the peas* and *spread the beans* have only literal interpretations.

Idiomatic combinations are also syntactically peculiar: most VNICs cannot undergo syntactic variations and at the same time retain their idiomatic interpretations. It is important, however, to note that VNICs differ with respect to the degree of syntactic flexibility they exhibit. Some are syntactically inflexible for the most part, while others are more versatile; as illustrated in 1 and 2:

1. (a) Tim and Joy shot the breeze.
   (b) ?? Tim and Joy shot a breeze.
   (c) ?? Tim and Joy shot the breezes.
   (d) ?? Tim and Joy shot the fun breeze.
   (e) ?? The breeze was shot by Tim and Joy.
   (f) ?? The breeze that Tim and Joy kicked was fun.

2. (a) Tim spilled the beans.
   (b) ? Tim spilled some beans.
   (c) ?? Tim spilled the bean.
   (d) Tim spilled the official beans.
   (e) The beans were spilled by Tim.
   (f) The beans that Tim spilled troubled Joe.

Linguists have explained the lexical and syntactic flexibility of idiomatic combinations in terms of their semantic analyzability (e.g., Glucksberg 1993; Fellbaum 1993; Nunberg et al. 1994). Semantic analyzability is inversely related to idiomaticity. For example, the meaning of *shoot the breeze*, a highly idiomatic expression, has nothing to do with either *shoot* or *breeze*. In contrast, a less idiomatic expression, such as *spill the beans*, can be analyzed as *spill* corresponding to "reveal" and *beans* referring to "secret(s)". Generally, the constituents of a semantically analyzable idiom can be mapped onto their corresponding referents in the idiomatic interpretation. Hence analyzable (less idiomatic) expressions are often more open to lexical substitution and syntactic variation.

2.2 Our Proposal

We use the observed connection between idiomaticity and (in)flexibility to devise statistical measures for automatically distinguishing idiomatic from literal verb+noun combinations. While VNICs vary in their degree of flexibility (cf. 1 and 2 above; see also Moon 1998), on the whole they contrast with compositional phrases, which are more lexically productive and appear in a wider range of syntactic forms. We thus propose to use the degree of lexical and syntactic flexibility of a given verb+noun combination to determine the level of idiomaticity of the expression.

It is important to note that semantic analyzability is neither a necessary nor a sufficient condition for an idiomatic combination to be lexically or syntactically flexible. Other factors, such as the communicative intentions and pragmatic constraints, can motivate a speaker to use a variant in place of a canonical form (Glucksberg, 1993). Nevertheless, lexical and syntactic flexibility may well be used as partial indicators of semantic analyzability, and hence idiomaticity.

3 Automatic Recognition of VNICs

Here we describe our measures for idiomaticity, which quantify the degree of lexical, syntactic, and overall fixedness of a given verb+noun combination, represented as a verb–noun pair. (Note that our measures quantify fixedness, not flexibility.)

3.1 Measuring Lexical Fixedness

A VNIC is lexically fixed if the replacement of any of its constituents by a semantically (and syntactically) similar word generally does not result in another VNIC, but in an invalid or a literal expression. One way of measuring lexical fixedness of a given verb+noun combination is thus to examine the idiomaticity of its variants, i.e., expressions generated by replacing one of the constituents by a similar word. This approach has two main challenges: (i) it requires prior knowledge about the idiomaticity of expressions (which is what we are developing our measure to determine); (ii) it needs information on "similarity" among words.

Inspired by Lin (1999), we examine the strength of association between the verb and noun constituents of the target combination and its variants, as an indirect cue to their idiomaticity. We use the automatically-built thesaurus of Lin (1998) to find similar words to the noun of the target expression, in order to automatically generate variants. Only the noun constituent is varied, since replacing the verb constituent of a VNIC with a semantically related verb is more likely to yield another VNIC, as in *keep/lose one's cool* (Nunberg et al., 1994).

Let $\mathcal{S}_{sim}(n) = \{n_m \mid 1 \leq m \leq M\}$ be the set of the M most similar nouns to the noun n of the target pair $\prec v, n \succ$. We calculate the association strength for the target pair, and for each of its variants, $\prec v, n_m \succ$, using pointwise mutual information (PMI) (Church et al., 1991):

$$\begin{aligned} \text{PMI}(v,\, n_j) &= \log \frac{P(v,\, n_j)}{P(v)\, P(n_j)} \\ &= \log \frac{|\mathcal{V} \times \mathcal{N}|\, f(v,\, n_j)}{f(v,\, *)\, f(*,\, n_j)} \quad (1) \end{aligned}$$

where $0 \leq j \leq M$ and n_0 is the target noun; $\mathcal{V}$ is the set of all transitive verbs in the corpus; $\mathcal{N}$ is the set of all nouns appearing as the direct object of some verb; $f(v,\, n_j)$ is the frequency of v and n_j occurring as a verb–object pair; $f(v,\, *)$ is the total frequency of the target verb with any noun in $\mathcal{N}$; $f(*,\, n_j)$ is the total frequency of the noun n_j in the direct object position of any verb in $\mathcal{V}$.

Lin (1999) assumes that a target expression is non-compositional if and only if its PMI value is significantly different from that of any of the variants. Instead, we propose a novel technique that brings together the association strengths (PMI values) of the target and the variant expressions into a single measure reflecting the *degree* of lexical fixedness for the target pair. We assume that the target pair is lexically fixed to the extent that its PMI deviates from the average PMI of its variants. Our measure calculates this deviation, normalized using the sample's standard deviation:

$$\text{Fixedness}_{\text{lex}}(v,\, n) = \frac{\text{PMI}(v,\, n) - \overline{\text{PMI}}}{s} \quad (2)$$

$\overline{\text{PMI}}$ is the mean and s the standard deviation of the sample; $\text{Fixedness}_{\text{lex}}(v,\, n) \in [-\infty, +\infty]$.

3.2 Measuring Syntactic Fixedness

Compared to compositional verb+noun combinations, VNICs are expected to appear in more restricted syntactic forms. To quantify the syntactic fixedness of a target verb–noun pair, we thus need to: (i) identify relevant syntactic patterns, i.e., those that help distinguish VNICs from literal verb+noun combinations; (ii) translate the frequency distribution of the target pair in the identified patterns into a measure of syntactic fixedness.

3.2.1 Identifying Relevant Patterns

Determining a unique set of syntactic patterns appropriate for the recognition of all idiomatic combinations is difficult indeed: exactly which forms an idiomatic combination can occur in is not entirely predictable (Sag et al., 2002). Nonetheless, there are hypotheses about the difference in behaviour of VNICs and literal verb+noun combinations with respect to particular syntactic variations (Nunberg et al., 1994). Linguists note that semantic analyzability is related to the referential status of the noun constituent, which is in turn related to participation in certain morphosyntactic forms. In what follows, we describe three types of variation that are tolerated by literal combinations, but are prohibited by many VNICs.

Passivization There is much evidence in the linguistic literature that VNICs often do not undergo passivization.[1] Linguists mainly attribute this to the fact that only a referential noun can appear as the surface subject of a passive construction.

[1] There are idiomatic combinations that are used only in a passivized form; we do not consider such cases in our study.

Determiner Type A strong correlation exists between the flexibility of the determiner preceding the noun in a verb+noun combination and the overall flexibility of the phrase (Fellbaum, 1993). It is however important to note that the nature of the determiner is also affected by other factors, such as the semantic properties of the noun.

Pluralization While the verb constituent of a VNIC is morphologically flexible, the morphological flexibility of the noun relates to its referential status. A non-referential noun constituent is expected to mainly appear in just one of the singular or plural forms. The pluralization of the noun is of course also affected by its semantic properties.

Merging the three variation types results in a pattern set, $\mathcal{PS}$, of 11 distinct syntactic patterns, given in Table 1.[2]

3.2.2 Devising a Statistical Measure

The second step is to devise a statistical measure that quantifies the degree of syntactic fixedness of a verb–noun pair, with respect to the selected set of patterns, $\mathcal{PS}$. We propose a measure that compares the "syntactic behaviour" of the target pair with that of a "typical" verb–noun pair. Syntactic behaviour of a typical pair is defined as the prior probability distribution over the patterns in $\mathcal{PS}$. The prior probability of an individual pattern $pt \in \mathcal{PS}$ is estimated as:

$$P(pt) \quad = \quad \frac{\displaystyle\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} f(v_i,\, n_j,\, pt)}{\displaystyle\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} \sum_{pt_k \in \mathcal{PS}} f(v_i,\, n_j,\, pt_k)}$$

The syntactic behaviour of the target verb–noun pair $\prec v, n \succ$ is defined as the posterior probability distribution over the patterns, given the particular pair. The posterior probability of an individual pattern pt is estimated as:

$$\begin{aligned} P(pt|\, v,\, n) \quad &= \quad \frac{P(v,\, n,\, pt)}{P(v,\, n)} \\ &= \quad \frac{f(v,\, n,\, pt)}{\displaystyle\sum_{pt_k \in \mathcal{PS}} f(v,\, n,\, pt_k)} \end{aligned}$$

The degree of syntactic fixedness of the target verb–noun pair is estimated as the divergence of its syntactic behaviour (the posterior distribution over the patterns), from the typical syntactic behaviour (the prior distribution). The divergence of the two probability distributions is calculated using a standard information-theoretic measure, the Kullback Leibler (KL-)divergence:

Patterns					
v	det:NULL	n_{sg}	v	det:NULL	n_{pl}
v	det:*a/an*	n_{sg}			
v	det:*the*	n_{sg}	v	det:*the*	n_{pl}
v	det:DEM	n_{sg}	v	det:DEM	n_{pl}
v	det:POSS	n_{sg}	v	det:POSS	n_{pl}
v	det:OTHER	[$\mathrm{n}_{sg,pl}$]	det:ANY [$\mathrm{n}_{sg,pl}$] be $\mathrm{v}_{passive}$		

Table 1: Patterns for syntactic fixedness measure.

$$\begin{aligned} &\mathrm{Fixedness_{syn}}\,(v,\, n) \\ &\quad = \quad D(P(pt|v,n)\,||\,P(pt)) \\ &\quad = \quad \sum_{pt_k \in \mathcal{PS}} P(pt_k|\, v,\, n) \log \frac{P(pt_k|\, v,\, n)}{P(pt_k)} \qquad (3) \end{aligned}$$

KL-divergence is always non-negative and is zero if and only if the two distributions are exactly the same. Thus, $\mathrm{Fixedness_{syn}}(v,\, n) \in [0, +\infty]$.

KL-divergence is argued to be problematic because it is not a symmetric measure. Nonetheless, it has proven useful in many NLP applications (Resnik, 1999; Dagan et al., 1994). Moreover, the asymmetry is not an issue here since we are concerned with the relative distance of several posterior distributions from the same prior.

3.3 A Hybrid Measure of Fixedness

VNICs are hypothesized to be, in most cases, both lexically and syntactically more fixed than literal verb+noun combinations (see Section 2). We thus propose a new measure of idiomaticity to be a measure of the overall fixedness of a given pair. We define $\mathrm{Fixedness_{overall}}\,(v,\, n)$ as:

$$\begin{aligned} &\mathrm{Fixedness_{overall}}\,(v,\, n) \\ &\quad = \quad \alpha\; \mathrm{Fixedness_{syn}}\,(v,\, n) \\ &\qquad\quad + (1-\alpha)\; \mathrm{Fixedness_{lex}}\,(v,\, n) \qquad (4) \end{aligned}$$

where α weights the relative contribution of the measures in predicting idiomaticity.

4 Evaluation of the Fixedness Measures

To evaluate our proposed fixedness measures, we determine their appropriateness as indicators of idiomaticity. We pose a classification task in which idiomatic verb–noun pairs are distinguished from literal ones. We use each measure to assign scores

[2] We collapse some patterns since with a larger pattern set the measure may require larger corpora to perform reliably.

to the experimental pairs (see Section 4.2 below). We then classify the pairs by setting a threshold, here the median score, where all expressions with scores higher than the threshold are labeled as idiomatic and the rest as literal.

We assess the overall goodness of a measure by looking at its accuracy (*Acc*) and the relative reduction in error rate (*RER*) on the classification task described above. The *RER* of a measure reflects the improvement in its accuracy relative to another measure (often a baseline).

We consider two baselines: (i) a random baseline, Rand, that randomly assigns a label (literal or idiomatic) to each verb–noun pair; (ii) a more informed baseline, PMI, an information-theoretic measure widely used for extracting statistically significant collocations.[3]

4.1 Corpus and Data Extraction

We use the British National Corpus (BNC; "http://www.natcorp.ox.ac.uk/") to extract verb–noun pairs, along with information on the syntactic patterns they appear in. We automatically parse the corpus using the Collins parser (Collins, 1999), and further process it using TGrep2 (Rohde, 2004). For each instance of a transitive verb, we use heuristics to extract the noun phrase (NP) in either the direct object position (if the sentence is active), or the subject position (if the sentence is passive). We then use NP-head extraction software[4] to get the head noun of the extracted NP, its number (singular or plural), and the determiner introducing it.

4.2 Experimental Expressions

We select our development and test expressions from verb–noun pairs that involve a member of a predefined list of (transitive) "basic" verbs. Basic verbs, in their literal use, refer to states or acts that are central to human experience. They are thus frequent, highly polysemous, and tend to combine with other words to form idiomatic combinations (Nunberg et al., 1994). An initial list of such verbs was selected from several linguistic and psycholinguistic studies on basic vocabulary (e.g., Pauwels 2000; Newman and Rice 2004). We further augmented this initial list with verbs that are semantically related to another verb already in the list; e.g., *lose* is added in analogy with *find*. The final list of 28 verbs is:

blow, bring, catch, cut, find, get, give, have, hear, hit, hold, keep, kick, lay, lose, make, move, place, pull, push, put, see, set, shoot, smell, take, throw, touch

From the corpus, we extract all verb–noun pairs with minimum frequency of 10 that contain a basic verb. From these, we semi-randomly select an idiomatic and a literal subset.[5] A pair is considered idiomatic if it appears in a credible idiom dictionary, such as the Oxford Dictionary of Current Idiomatic English (ODCIE) (Cowie et al., 1983), or the Collins COBUILD Idioms Dictionary (CCID) (Seaton and Macaulay, 2002). Otherwise, the pair is considered literal. We then randomly pull out 160 development and 200 test pairs (half idiomatic and half literal), ensuring both low and high frequency items are included. Sample idioms corresponding to the extracted pairs are: *kick the habit*, *move mountains*, *lose face*, and *keep one's word*.

4.3 Experimental Setup

Development expressions are used in devising the fixedness measures, as well as in determining the values of the parameters M in Eqn. (2) and α in Eqn. (4). M determines the maximum number of nouns similar to the target noun, to be considered in measuring the lexical fixedness of a given pair. The value of this parameter is determined by performing experiments over the development data, in which M ranges from 10 to 100 by steps of 10; M is set to 50 based on the results. We also experimented with different values of α ranging from 0 to 1 by steps of .1. Based on the development results, the best value for α is .8 (giving more weight to the syntactic fixedness measure).

Test expressions are saved as unseen data for the final evaluation. We further divide the set of all test expressions, $\mathrm{TEST}_{\mathrm{all}}$, into two sets corresponding to two frequency bands: $\mathrm{TEST}_{f_{\mathrm{low}}}$ contains 50 idiomatic and 50 literal pairs, each with total frequency between 10 and 40 ($10 \leq \mathit{freq}(v,\, n,\, *) < 40$); $\mathrm{TEST}_{f_{\mathrm{high}}}$ consists of 50 idiomatic and 50 literal pairs, each with total frequency of 40 or greater ($\mathit{freq}(v,\, n,\, *) \;\geq\; 40$). All frequency counts are over the entire BNC.

4.4 Results

We first examine the performance of the individual fixedness measures, $\mathrm{Fixedness}_{\mathrm{lex}}$ and

[3] As in Eqn. (1), our calculation of PMI here restricts the verb–noun pair to the direct object relation.

[4] We use a modified version of the software provided by Eric Joanis based on heuristics from (Collins, 1999).

[5] In selecting literal pairs, we choose those that involve a physical act corresponding to the basic semantics of the verb.

Data Set:	TEST$_{\text{all}}$	
	%Acc	*%RER*
Rand	50	-
PMI	64	28
Fixedness$_{\text{lex}}$	65	30
Fixedness$_{\text{syn}}$	**70**	**40**

Table 2: Accuracy and relative error reduction for the two fixedness and the two baseline measures over all test pairs.

Data Set:	TEST$_{f_{\text{low}}}$		TEST$_{f_{\text{high}}}$	
	%Acc	*%RER*	*%Acc*	*%RER*
Rand	50	-	50	-
PMI	56	12	70	40
Fixedness$_{\text{lex}}$	68	36	66	32
Fixedness$_{\text{syn}}$	**72**	**44**	**82**	**64**

Table 3: Accuracy and relative error reduction for all measures over test pairs divided by frequency.

Data Set:	TEST$_{\text{all}}$	
	%Acc	*%RER*
Fixedness$_{\text{lex}}$	65	30
Fixedness$_{\text{syn}}$	70	40
Fixedness$_{\text{overall}}$	**74**	**48**

Table 4: Performance of the hybrid measure over TEST$_{\text{all}}$.

Fixedness$_{\text{syn}}$, as well as that of the two baselines, Rand and PMI; see Table 2. (Results for the overall measure are presented later in this section.) As can be seen, the informed baseline, PMI, shows a large improvement over the random baseline (28% error reduction). This shows that one can get relatively good performance by treating verb+noun idiomatic combinations as collocations.

Fixedness$_{\text{lex}}$ performs as well as the informed baseline (30% error reduction). This result shows that, as hypothesized, lexical fixedness is a reasonably good predictor of idiomaticity. Nonetheless, the performance signifies a need for improvement. Possibly the most beneficial enhancement would be a change in the way we acquire the similar nouns for a target noun.

The best performance (shown in boldface) belongs to Fixedness$_{\text{syn}}$, with 40% error reduction over the random baseline, and 20% error reduction over the informed baseline. These results demonstrate that syntactic fixedness is a good indicator of idiomaticity, better than a simple measure of collocation (PMI), or a measure of lexical fixedness. These results further suggest that looking into deep linguistic properties of VNICs is both necessary and beneficial for the appropriate treatment of these expressions.

PMI is known to perform poorly on low frequency data. To examine the effect of frequency on the measures, we analyze their performance on the two divisions of the test data, corresponding to the two frequency bands, TEST$_{f_{\text{low}}}$ and TEST$_{f_{\text{high}}}$. Results are given in Table 3, with the best performance shown in boldface.

As expected, the performance of PMI drops substantially for low frequency items. Interestingly, although it is a PMI-based measure, Fixedness$_{\text{lex}}$ performs slightly better when the data is separated based on frequency. The performance of Fixedness$_{\text{syn}}$ improves quite a bit when it is applied to high frequency items, while it improves only slightly on the low frequency items. These results show that both Fixedness measures perform better on homogeneous data, while retaining comparably good performance on heterogeneous data. These results reflect that our fixedness measures are not as sensitive to frequency as PMI. Hence they can be used with a higher degree of confidence, especially when applied to data that is heterogeneous with regard to frequency. This is important because while some VNICs are very common, others have very low frequency.

Table 4 presents the performance of the hybrid measure, Fixedness$_{\text{overall}}$, repeating that of Fixedness$_{\text{lex}}$ and Fixedness$_{\text{syn}}$ for comparison. Fixedness$_{\text{overall}}$ outperforms both lexical and syntactic fixedness measures, with a substantial improvement over Fixedness$_{\text{lex}}$, and a small, but notable, improvement over Fixedness$_{\text{syn}}$. Each of the lexical and syntactic fixedness measures is a good indicator of idiomaticity on its own, with syntactic fixedness being a better predictor. Here we demonstrate that combining them into a single measure of fixedness, while giving more weight to the better measure, results in a more effective predictor of idiomaticity.

5 Determining the Canonical Forms

Our evaluation of the fixedness measures demonstrates their usefulness for the automatic recognition of idiomatic verb–noun pairs. To represent such pairs in a lexicon, however, we must determine their canonical form(s)—Cforms henceforth. For example, the lexical representation of $\prec$*shoot, breeze*$\succ$ should include *shoot the breeze* as a Cform.

Since VNICs are syntactically fixed, they are mostly expected to have a single Cform. Nonetheless, there are idioms with two or more acceptable forms. For example, *hold fire* and *hold one's fire* are both listed in CCID as variations of the same idiom. Our approach should thus be capable of predicting all allowable forms for a given idiomatic verb–noun pair.

We expect a VNIC to occur in its Cform(s) more frequently than it occurs in any other syntactic patterns. To discover the Cform(s) for a given idiomatic verb–noun pair, we thus examine its frequency of occurrence in each syntactic pattern in $\mathcal{PS}$. Since it is possible for an idiom to have more than one Cform, we cannot simply take the most dominant pattern as the canonical one. Instead, we calculate a z-score for the target pair $\prec v,\, n\succ$ and each pattern $pt_k \in \mathcal{PS}$:

$$z_k(v,\, n) \quad = \quad \frac{f(v,\, n,\, pt_k) - \overline{f}}{s}$$

in which $\overline{f}$ is the mean and s the standard deviation over the sample $\{f(v,\, n,\, pt_k)\,|\,pt_k \in \mathcal{PS}\}$.

The statistic $z_k(v,\, n)$ indicates how far and in which direction the frequency of occurrence of the pair $\prec v,\, n\succ$ in pattern pt_k deviates from the sample's mean, expressed in units of the sample's standard deviation. To decide whether pt_k is a canonical pattern for the target pair, we check whether $z_k(v,\, n) > T_z$, where T_z is a threshold. For evaluation, we set T_z to 1, based on the distribution of z and through examining the development data.

We evaluate the appropriateness of this approach in determining the Cform(s) of idiomatic pairs by verifying its predicted forms against ODCIE and CCID. Specifically, for each of the 100 idiomatic pairs in TEST$_{\text{all}}$, we calculate the precision and recall of its predicted Cforms (those whose z-scores are above T_z), compared to the Cforms listed in the two dictionaries. The average precision across the 100 test pairs is 81.7%, and the average recall is 88.0% (with 69 of the pairs having 100% precision and 100% recall). Moreover, we find that for the overwhelming majority of the pairs, 86%, the predicted Cform with the highest z-score appears in the dictionary entry of the pair. Thus, our method of detecting Cforms performs quite well.

6 Discussion and Conclusions

The significance of the role idioms play in language has long been recognized. However, due to their peculiar behaviour, idioms have been mostly overlooked by the NLP community. Recently, there has been growing awareness of the importance of identifying non-compositional multiword expressions (MWEs). Nonetheless, most research on the topic has focused on compound nouns and verb particle constructions. Earlier work on idioms have only touched the surface of the problem, failing to propose explicit mechanisms for appropriately handling them. Here, we provide effective mechanisms for the treatment of a broadly documented and crosslinguistically frequent class of idioms, i.e., VNICs.

Earlier research on the lexical encoding of idioms mainly relied on the existence of human annotations, especially for detecting which syntactic variations (e.g., passivization) an idiom can undergo (Villavicencio et al., 2004). We propose techniques for the automatic acquisition and encoding of knowledge about the lexicosyntactic behaviour of idiomatic combinations. We put forward a means for automatically discovering the set of syntactic variations that are tolerated by a VNIC and that should be included in its lexical representation. Moreover, we incorporate such information into statistical measures that effectively predict the idiomaticity level of a given expression. In this regard, our work relates to previous studies on determining the compositionality (inverse of idiomaticity) of MWEs other than idioms.

Most previous work on compositionality of MWEs either treat them as collocations (Smadja, 1993), or examine the distributional similarity between the expression and its constituents (McCarthy et al., 2003; Baldwin et al., 2003; Bannard et al., 2003). Lin (1999) and Wermter and Hahn (2005) go one step further and look into a linguistic property of non-compositional compounds—their lexical fixedness—to identify them. Venkatapathy and Joshi (2005) combine aspects of the above-mentioned work, by incorporating lexical fixedness, collocation-based, and distributional similarity measures into a set of features which are used to rank verb+noun combinations according to their compositionality.

Our work differs from such studies in that it carefully examines several linguistic properties of VNICs that distinguish them from literal (compositional) combinations. Moreover, we suggest novel techniques for translating such characteristics into measures that predict the idiomaticity level of verb+noun combinations. More specifically, we propose statistical measures that quantify the degree of lexical, syntactic, and overall fixedness of such combinations. We demonstrate that these measures can be successfully applied to the task of automatically distinguishing idiomatic combinations from non-idiomatic ones. We also show that our syntactic and overall fixedness measures substantially outperform a widely used measure of collocation, PMI, even when the latter takes syntactic relations into account.

Others have also drawn on the notion of syntactic fixedness for idiom detection, though specific to a highly constrained type of idiom (Widdows and Dorow, 2005). Our syntactic fixedness measure looks into a broader set of patterns associated with a large class of idiomatic expressions. Moreover, our approach is general and can be easily extended to other idiomatic combinations.

Each measure we use to identify VNICs captures a different aspect of idiomaticity: PMI reflects the statistical idiosyncrasy of VNICs, while the fixedness measures draw on their lexicosyntactic peculiarities. Our ongoing work focuses on combining these measures to distinguish VNICs from other idiosyncratic verb+noun combinations that are neither purely idiomatic nor completely literal, so that we can identify linguistically plausible classes of verb+noun combinations on this continuum (Fazly and Stevenson, 2005).

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proc. of the ACL-SIGLEX Workshop on Multiword Expressions*, 89–96.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL-SIGLEX Workshop on Multiword Expressions*, 65–72.

Cristina Cacciari and Patrizia Tabossi, editors. 1993. *Idioms: Processing, Structure, and Interpretation*. Lawrence Erlbaum Associates, Publishers.

Cristina Cacciari. 1993. The place of idioms in a literal and metaphorical world. In Cacciari and Tabossi (Cacciari and Tabossi, 1993), 27–53.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115–164. Lawrence Erlbaum.

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.

Ido Dagan, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proc. of ACL'94*, 272–278.

Afsaneh Fazly and Suzanne Stevenson. 2005. Automatic acquisition of knowledge about multiword predicates. In *Proc. of PACLIC'05*.

Christiane Fellbaum. 1993. The determiner in English idioms. In Cacciari and Tabossi (Cacciari and Tabossi, 1993), 271–295.

Christiane Fellbaum. 2005. The ontological loneliness of verb phrase idioms. In Andrea Schalley and Dietmar Zaefferer, editors, *Ontolinguistics*. Mouton de Gruyter. Forthcomming.

Sam Glucksberg. 1993. Idiom meanings and allusional content. In Cacciari and Tabossi (Cacciari and Tabossi, 1993), 3–26.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL'98*.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proc. of ACL'99*, 317–24.

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

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.

John Newman and Sally Rice. 2004. Patterns of usage for English SIT, STAND, and LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics*, 15(3):351–396.

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

Paul Pauwels. 2000. *Put, Set, Lay and Place: A Cognitive Linguistic Approach to Verbal Meaning*. LINCOM EUROPA.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *JAIR*, (11):95–130.

Douglas L. T. Rohde. 2004. TGrep2 User Manual.

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

Maggie Seaton and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, 2nd edition.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *CL*, 19(1):143–177.

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

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of MWEs. In *Proc. of the ACL'04 Workshop on Multiword Expressions*, 80–87.

Joachim Wermter and Udo Hahn. 2005. Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proc. of HLT-EMNLP'05*, 843–850.

Dominic Widdows and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proc. of ACL'05 Workshop on Deep Lexical Acquisition*, 48–56.