Verb Noun Construction MWE Token Supervised Classification

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

Pravin Bhutada
Computer Science Department
Columbia University
pb2351@columbia.edu

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. We present a supervised learning approach to the problem. We experiment with different features. Our approach yields the best results to date on MWE classification combining different linguistically motivated features, the overall performance yields an F-measure of 84.58% corresponding to an F-measure of 89.96% for idiomaticity identification and classification and 62.03% for literal identification and classification.

1 Introduction

In the literature in general a multiword expression (MWE) refers to a multiword unit or a collocation of words that co-occur together statistically more than chance. A MWE is a cover term for different types of collocations which vary in their transparency and fixedness. 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. In fact, the seminal paper (Sag et al., 2002) refers to this problem as a *key* issue for the development of high-quality NLP applications.

For our purposes, a MWE is defined as a collocation of words 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* from the meaning of its component words. An MWE meaning is transparent, i.e. predictable, in as much as the component words in the expression relay the meaning portended by the speaker compositionally. Accordingly, MWEs vary in their degree of meaning compositionality; compositionality is correlated with the level of idiomaticity. An MWE is compositional if the meaning of an MWE as a unit can be predicted from the meaning of its component words such as in *make a decision* meaning *to decide*. If we conceive of idiomaticity as being a continuum, the more idiomatic an expression, the less transparent and the more non-compositional it is. Some MWEs are more predictable than others, for instance, *kick the bucket*, when used idiomatically to mean *to die*, has nothing in common with the literal meaning of either *kick* or *bucket*, however, *make a decision* is very clearly related to *to decide*. Both of these expressions are considered MWEs but have varying degrees of compositionality and predictability. Both of these expressions belong to a class of idiomatic MWEs known as verb noun constructions (VNC). The first VNC *kick the bucket* is a non-decomposable VNC MWE, the latter *make a decision* is a decomposable VNC MWE. These types of constructions are the object of our study.

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 on the kitchen counter* is most likely a literal usage. This is given away by the use of the prepositional phrase *on the kitchen counter*, as 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 contextual usage. In a recent study by (Cook et al., 2008) of 53 idiom MWE types used in different contexts, the authors 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 VNC MWE token, to be able to determine whether it is used idiomatically or not as it could potentially have detrimental effects on the quality of the translation.

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 17–22,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

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. To date most approaches to the problem of idiomaticity classification on the token level have been unsupervised (Birke and Sarkar, 2006; Diab and Krishna, 2009b; Diab and Krishna, 2009a; Sporleder and Li, 2009). In this study we carry out a supervised learning investigation using support vector machines that uses some of the features which have been shown to help in unsupervised approaches to the problem.

This paper is organized as follows: In Section 2 we describe our understanding of the various classes of MWEs in general. Section 3 is a summary of previous related research. Section 4 describes our approach. In Section 5 we present the details of our experiments. We discuss the results in Section 6. Finally, we conclude 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*. The VNCs which we are focusing on in this paper fall into this category. **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* and collocations such as *machine translation*. These expressions are *statistically idiosyncratic*. For instance, *traffic light* is the most likely lexicalization of the concept and would occur more often in text than, say, *traffic regulator* or *vehicle light*.

3 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 (Fazly and Stevenson, 2007; Cook et al., 2007). We are aware of two supervised approaches to the problem: work by (Katz and Giesbrecht, 2006) and work by (Hashimoto and Kawahara, 2008).

In Katz and Giesbrecht (2006) (KG06) the authors carried out a vector similarity comparison between the context of an MWE and that of the constituent words using LSA to determine if the expression is idiomatic or not. The KG06 is similar in intuition to work proposed by (Fazly and Stevenson, 2007), however the latter work was unsupervised. KG06 experimented with a tiny data set of only 108 sentences corresponding to one MWE idiomatic expression.

Hashimoto and Kawahara (2008) (HK08) is the first large scale study to our knowledge that addressed token classification into idiomatic versus literal for Japanese MWEs of all types. They apply a supervised learning framework using support vector machines based on TinySVM with a quadratic kernel. They annotate a web based corpus for training data. They identify 101 idiom types each with a corresponding 1000 examples, hence they had a corpus of 102K sentences of annotated data for their experiments. They experiment with 90 idiom types only for which they had more than 50 examples. They use two types of features: word sense disambiguation (WSD) features and idiom features. The WSD features comprised some basic syntactic features such as POS, lemma information, token n-gram features, in addition to hypernymy information on words as well as domain information. For the idiom features they were mostly inflectional features such as voice, negativity, modality, in addition to adjacency and adnominal features. They report results in terms of accuracy and rate of error reduction. Their overall accuracy is of 89.25% using all the features.

4 Our Approach

We apply a supervised learning framework to the problem of both identifying and classifying a MWE expression token in context. We specifically focus on VNC MWE expressions. We use the annotated data by (Cook et al., 2008). We adopt a chunking approach to the problem using an Inside Outside Beginning (IOB) tagging framework for performing the identification of MWE VNC tokens and classifying them as idiomatic or literal in context. For chunk tagging, we use the Yam-

Cha sequence labeling system.[1] YamCha is based on Support Vector Machines technology using degree 2 polynomial kernels.

We label each sentence with standard IOB tags. Since this is a binary classification task, we have 5 different tags: B-L (Beginning of a literal chunk), I-L (Inside of a literal chunk), B-I (Beginning an Idiomatic chunk), I-I (Inside an Idiomatic chunk), O (Outside a chunk). As an example a sentence such as *J*ohn kicked the bucket last Friday will be annotated as follows: *J*ohn O, kicked B-I, the I-I, bucket I-I, last O, Friday O. We experiment with some basic features and some more linguistically motivated ones.

We experiment with different window sizes for context ranging from $-/+1$ to $-/+5$ tokens before and after the token of interest. We also employ linguistic features such as character n-gram features, namely last 3 characters of a token, as a means of indirectly capturing the word inflectional and derivational morphology (NGRAM). Other features include: Part-of-Speech (POS) tags, lemma form (LEMMA) or the citation form of the word, and named entity (NE) information. The latter feature is shown to help in the unsupervised setting in recent work (Diab and Krishna, 2009b; Diab and Krishna, 2009a). In general all the linguistic features are represented as separate feature sets explicitly modeled in the input data. Hence, if we are modeling the POS tag feature for our running example the training data would be annotated as follows: {John *NN* O, kicked *VBD* B-I, the *Det* I-I, bucket *NN* I-I, last *ADV* O, Friday *NN* O }. Likewise adding the NGRAM feature would be represented as follows: {John *NN ohn* O, kicked *VBD ked* B-I, the *Det the* I-I, bucket *NN ket* I-I, last *ADV ast* O, Friday *NN day* O.} and so on.

With the NE feature, we followed the same representation as the other features as a separate column as expressed above, referred to as Named Entity Separate (NES). For named entity recognition (NER) we use the BBN Identifinder software which identifies 19 NE tags.[2] We have two settings for NES: one with the full 19 tags explicitly identified (NES-Full) and the other where we have a binary feature indicating whether a word is a NE or not (NES-Bin). Moreover, we added another experimental condition where we changed the words' representation in the input to their NE class, Named Entity InText (NEI). For example for the NEI condition, our running example is represented as follows: {PER *NN ohn* O, kicked *VBD ked* B-I, the *Det the* I-I, bucket *NN ket* I-I, last *ADV ast* O, DAY *NN day* O}, where *John* is replaced by the NE "PER" .

[1] http://www.tado-chasen.com/yamcha

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

5 Experiments and Results

5.1 Data

We use the manually annotated standard data set identified in (Cook et al., 2008). This data comprises 2920 unique VNC-Token expressions drawn from the entire British National Corpus (BNC).[3] The BNC contains 100M words of multiple genres including written text and transcribed speech. In this set, VNC token expressions are manually annotated as *idiomatic*, *literal* or *unknown*. We exclude those annotated as *unknown* and those pertaining to the Speech part of the data leaving us with a total of 2432 sentences corresponding to 53 VNC MWE types. This data has 2571 annotations,[4] corresponding to 2020 Idiomatic tokens and 551 literal ones. Since the data set is relatively small we carry out 5-fold cross validation experiments. The results we report are averaged over the 5 folds per condition. We split the data into 80% for training, 10% for testing and 10% for development. The data used is the tokenized version of the BNC.

5.2 Evaluation Metrics

We use $F_{\beta=1}$ (F-measure) as the harmonic mean between (P)recision and (R)ecall, as well as accuracy to report the results.[5] We report the results separately for the two classes IDM and LIT averaged over the 5 folds of the TEST data set.

5.3 Results

We present the results for the different features sets and their combination. We also present results on a simple most frequent tag baseline (FREQ) as well as a baseline of using no features, just the tokenized words (TOK). The baseline is basically tagging all *identified* VNC tokens in the data set as idiomatic. It is worth noting that the baseline has the advantage of gold identification of MWE VNC token expressions. In our experimental conditions, identification of a potential VNC MWE is part of what is discovered automatically, hence our system is penalized for identifying other VNC MWE

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

[4] A sentence can have more than one MWE expression hence the number of annotations exceeds the number of sentences.

[5] We do not think that accuracy should be reported in general since it is an inflated result as it is not a measure of error. All words identified as O factor into the accuracy which results in exaggerated values for accuracy. We report it only since it the metric used by previous work.

tokens that are not in the original data set.[6]

In Table 2 we present the results yielded per feature and per condition. We experimented with different context sizes initially to decide on the optimal window size for our learning framework, results are presented in Table 1. Then once that is determined, we proceed to add features.

Noting that a window size of $-/+3$ yields the best results, we proceed to use that as our context size for the following experimental conditions. We will not include accuracy since it above 96% for all our experimental conditions.

All the results yielded by our experiments outperform the baseline FREQ. The simple tokenized words baseline (TOK) with no added features with a context size of $-/+3$ shows a significant improvement over the very basic baseline FREQ with an overall F measure of 77.04%.

Adding lemma information or POS or NGRAM features all independently contribute to a better solution, however combining the three features yields a significant boost in performance over the TOK baseline of 2.67% absolute F points in overall performance.

Confirming previous observations in the literature, the overall best results are obtained by using NE features. The NEI condition yields slightly better results than the NES conditions in the case when no other features are being used. NES-Full significantly outperforms NES-Bin when used alone especially on literal classification yielding the highest results on this class of phenomena across the board. However when combined with other features, NES-Bin fares better than NES-Full as we observe slightly less performance when comparing NES-Full+L+N+P and NES-Bin+L+N+P.

Combining NEI+L+N+P yields the highest results with an overall F measure of 84.58% a significant improvement over both baselines and over the condition that does not exploit NE features, L+N+P. Using NEI may be considered a form of dimensionality reduction hence the significant contribution to performance.

6 Discussion

The overall results strongly suggest that using linguistically interesting features explicitly has a positive impact on performance. NE features help the most and combining them with other features yields the best results. In general performance on the classification and identification of idiomatic expressions yielded much better results. This may be due to the fact that the data has a lot more idiomatic token examples for training. Also we note that precision scores are significantly higher than recall scores especially with performance on literal token instance classification. This might be an indication that identifying when an MWE is used literally is a difficult task.

We analyzed some of the errors yielded in our best condition NEI+L+N+P. The biggest errors are a result of identifying other VNC constructions not annotated in the training and test data as VNC MWEs. However, we also see errors of confusing idiomatic cases with literal ones 23 times, and the opposite 4 times.

Some of the errors where the VNC should have been classified as literal however the system classified them as idiomatic are *k*ick heel, find feet, make top. Cases of idiomatic expressions erroneously classified as literal are for MWE types *h*it the road, blow trumpet, blow whistle, bit a wall.

The system is able to identify new VNC MWE constructions. For instance in the sentence *On the other hand Pinkie seemed to have* **lost his head** *to a certain extent perhaps some prospects of* **making his mark** *by bringing in something novel in the way of business*, the first MWE **lost his head** is annotated in the training data, however **making his mark** is newly identified as idiomatic in this context.

Also the system identified **hit the post** as a literal MWE VNC token in *As the ball* **hit the post** *the referee* **blew the whistle**, where **blew the whistle** is a literal VNC in this context and it identified **hit the post** as another literal VNC.

7 Conclusion

In this study, we explore a set of features that contribute to VNC token expression binary supervised classification. The use of NER significantly improves the performance of the system. Using NER as a means of dimensionality reduction yields the best results. We achieve a state of the art performance of an overall F measure of 84.58%. In the future we are looking at ways of adding more sophisticated syntactic and semantic features from WSD. Given the fact that we were able to get more interesting VNC data automatically, we are currently looking into adding the new data to the annotated pool after manual checking.

[6] We could have easily identified all VNC syntactic configurations corresponding to verb object as a potential MWE VNC assuming that they are literal by default. This would have boosted our literal score baseline, however, for this investigation, we decided to strictly work with the gold standard data set exclusively.

	IDM-F	LIT-F	Overall F	Overall Acc.
$-/+1$	77.93	48.57	71.78	96.22
$-/+2$	85.38	55.61	79.71	97.06
$-/+3$	**86.99**	**55.68**	**81.25**	96.93
$-/+4$	86.22	55.81	80.75	97.06
$-/+5$	83.38	50	77.63	96.61

Table 1: Results in %s of varying context window size

	IDM-P	IDM-R	IDM-F	LIT-P	LIT-R	LIT-F	Overall F
FREQ	70.02	89.16	78.44	0	0	0	69.68
TOK	81.78	83.33	82.55	71.79	43.75	54.37	77.04
(L)EMMA	83.1	84.29	83.69	69.77	46.88	56.07	78.11
(N)GRAM	83.17	82.38	82.78	70	43.75	53.85	77.01
(P)OS	83.33	83.33	83.33	77.78	43.75	56.00	78.08
L+N+P	86.95	83.33	85.38	72.22	45.61	55.91	79.71
NES-Full	85.2	87.93	86.55	79.07	**58.62**	**67.33**	82.77
NES-Bin	84.97	82.41	83.67	73.49	52.59	61.31	79.15
NEI	89.92	85.18	87.48	81.33	52.59	63.87	82.82
NES-Full+L+N+P	89.89	84.92	87.34	76.32	50	60.42	81.99
NES-Bin+L+N+P	90.86	84.92	87.79	76.32	50	60.42	82.33
NEI+L+N+P	**91.35**	**88.42**	**89.86**	**81.69**	50	62.03	**84.58**

Table 2: Final results in %s averaged over 5 folds of test data using different features and their combinations

8 Acknowledgement

The first author was partially funded by DARPA GALE and MADCAT projects. The authors would like to acknowledge the useful comments by two 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.

J. Birke and A. Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL*, volume 6, pages 329–336.

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.

Mona Diab and Madhav Krishna. 2009a. Handling sparsity for verb noun MWE token classification. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 96–103, Athens, Greece, March. Association for Computational Linguistics.

Mona Diab and Madhav Krishna. 2009b. Unsupervised classification for vnc multiword expressions tokens. In *CICLING*.

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, Honolulu, 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.

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.

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.

C. Sporleder and L. Li. 2009. Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762. Association for Computational Linguistics.

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.