ISSN 2007-9737

Identification of Verbal Phraseological Units in Mexican News Stories

Belém Priego Sánchez[1,2], David Pinto[2]

[1] Université Paris 13, Sorbonne Paris Cité, LDI, France

[2] Benemérita Universidad Autónoma de Puebla, FCC, Mexico

belemps@gmail.com, dpinto@cs.buap.mx

Abstract. Verbal Phraseological Units are phrases made up of two or more words in which at least one of the words is a verb that plays the role of the predicate. One of the characteristics of this type of expression is that its global meaning rarely can be deduced from the meaning of its components. The automatic recognition of this type of linguistic structures is a very important task, since they are a standard way of expressing a concept or idea. In this paper we present the results obtained when different supervised machine learning methods are employed for determining whether or not a verbal phraseological unit is present in a given story of a newspaper. The experiments have been carried out using a supervised corpus of news stories (written in Mexican Spanish). Beside the results obtained in the experiments aforementioned, we provide access to a new lexicon having phrases as entries (instead of single words), in which each entry is associated to a real value (normalized between zero and one) indicating its probability of being a verbal phraseological unit.

Keywords. Verbal phraseological units, supervised machine learning, lexicon.

1 Introduction

The study of phraseological units has acquired a growing importance in recent years, in part because the linguistic and computational linguistic community has understood that this phenomenon covers all the sentence components [9], a fact that involves different dimensions of the natural language: linguistics, pragmatics, culturals, among others. A phraseological unit is basically one type of multiword expression, and under this denomination one assumes a wide range of linguistic constructions such as idioms (*storm in a teacup*, *sweep under the rug*), fixed phrases (*in vitro*, *by and large*, *rock'n roll*), noun compounds (*olive oil*, *laser printer*), compound verbs (*take a nap*, *bring about*), etc.

Phraseological units are multiword lexical units that are characterized by presenting a certain degree of fixation[1] or idiomaticity in its components. In other words, phraseological units are a combination of words whose meaning are not necessarily deduced from the meaning of its components, i.e., a phraseological expression can mean more than the sum of its parts [8, 10]. These linguistic structures are also known in literature as phrasemes, fixed expressions, and multiword expressions[2]. While easily mastered by native speakers, their interpretation poses a major challenge for computational systems, due to their fiexible and heterogeneous nature. Furthermore, phraseological units are not as frequent in lexical resources as they are in real-word texts, and this problem of coverage may impact the performance of many natural language processing tasks.

In this research work, we are particularly interested in studying Spanish phraseological units containing one verb as the grammar center, i.e., verbal phraseological units which present a challenged degree of fixation in comparison with other phraseological units [13], for example, *leer entre*

[1] Fixation is a inherent property of natural language that occupies a central role in the description of phraseological units.

[2] Throughout this paper we will employ the term *phraseological unit*, assuming that the terms aforementioned have a similar meaning.

líneas (to read between the lines). Actually, this paper aims to identify whether or not a verbal phraseological unit is present in a given text (news story), a process that implies an analysis of a raw text and the features in the context of a phraseological unit for creating computational algorithms that allow to fulfill the mentioned task in highly scalable environments.

The remaining part of this paper is structured as follows. Section 2 presents the research works found in literature which are associated to the topic of this paper. Section 3 describes the concept of phraseological units and presents examples of those linguistic structures, particularly, for verbal phraseological units. Section 4 shows the results obtained in the experiments carried out in this research work, presenting first the lexical resources (lexicon and corpus), then the supervised machine learning methods employed, and thereafter, the results obtained up to now. Finally, in Section 5 we give the conclusions and further work we plan to continue this research.

2 Related Work

Even though in this paper we use the term phraseological unit, we are aware that there are a number of research works in which the term multiword expression (MWE) is employed instead, thus we have added those research works as part of the related work.

Kenneth Church in [1], for example, discusses some of the relevant questions that arise when doing research on multiword expressions, examining how these linguistic structures are addressed in a variety of fields such as linguistics, lexicography, educational testing, and web search.

Davis et al. [4] look at the properties of support verb and nominalization in English, examining some linguistic factors for potential correlations with the acceptability of these constructions. In particular, they investigate whether or not the acceptability of support verb and nominalization pairs is linked to the membership in Levin verbal classes [7], evaluating around 2,700 (acceptable and unacceptable) combinations.

In [15], the task of MWEs identification is examined using the FXTagger tool based on conditional random fields. The authors looked at domain specificity of light verb constructions and checked the portability of the models generated from different corpora using domain adaptation techniques in order to reduce the gap between these domains.

In [11] the authors attempt to optimize and improve recall without losing precision in a fiexible search for the token identification of Italian MWEs in corpora. They propose a method for modeling the internal variation of MWEs in terms of frequent variation patterns of specific part-of-speech sequences, focusing on two types of nominal expressions.

We should emphasize that many other works associated with MWE exist in literature, mainly because of the different forums that are encouraged by the computational linguistic community[3], in which many interesting papers can be found. However, in literature there are other papers employing other terminology which refer to phraseological units, therefore, we mention some of them in what follows.

In [14] the authors propose a statistical measure for calculating the degree of acceptability of light verb constructions based on their linguistic properties. This measure shows good correlations with human ratings on unseen test data. Moreover, they have found that their measure correlates more strongly when the potential complements of the construction are separated into semantically similar classes. Their analysis demonstrates the systematic nature of the semi-productivity of these constructions.

Paul Cook et al. present the VNC-Tokens dataset, a resource of almost 3,000 English verb-noun combination usages annotated as literal or idiomatic [2]. These authors began with the dataset used by Fazly and Stevenson [6], which includes a list of idiomatic verb-noun combinations (VNCs), and found that approximately half of these expressions are attested fairly frequently in their literal sense in the British National Corpus (BNC)[4]. Their study is based on the observation that the idiomatic meaning of VNCs tends to be expressed in a small

[3] http://multiword.sourceforge.net

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

number of preferred lexico-syntactic patterns referred to as canonical forms [12].

In [6], the authors investigate the lexical and syntactic fiexibility of a class of idiomatic expressions. They develop measures that draw on linguistic properties, and demonstrate that these statistical corpus-based measures can be successfully used for distinguishing idiomatic combinations from non-idiomatic ones. They also propose a process to automatically determine which syntactic forms a particular idiom can appear in, and hence should be included in its lexical representation.

We consider that other works associated with the identification of phraseological units exist in literature, however, an exhaustive discussion of the state of the art is out of the scope of this paper. Instead, we consider it important to present in detail in the following section a description of such linguistic structures that are the aim of this research work.

3 Phraseological Units

A Phraseological Unit (PU) is basically a multiword lexical unit that is characterized by presenting a certain degree of fixation or idiomaticity. Phraseological units belong to what Coseriu [3] called "repeated discourse", and they are mainly characterized by the following three features:

1. Their poly-lexical behavior that distinguishes them from isolated words of the language, either simple or compound words.
2. Their fixation degree that presents them as if they were atomic units (inseparables) just like simple units are.
3. Their idiomaticity or lexical opacity, a feature that sometimes may be missing, as it occurs in the so-called collocations, a type of phrases that we will describe in the following paragraphs.

PUs almost never present such criteria as compositionality, substitutability, and modifiability, therefore avoiding any modification to their structure. A phraseological unit is a lexicalized, reproducible bilexemic or polylexemic word group in common use, which has relative syntactic and semantic stability, may be idiomatized, may carry connotations, and may have an emphatic or intensifying function in a text. PUs are stable word groups with partially or fully transferred meanings, for example, *Greek gift* (a gift given with the intention of tricking and causing harm to the recipient) or *to kick the bucket* (*to die*). Experts in phraseology usually consider three types of phraseological units:

1. Phraseological fusions
2. Phraseological unities
3. Phraseological combinations

A phraseological fusion is a semantically indivisible phraseological unit whose meaning is never influenced by the meanings of its components, in other words, the meaning of the components is completely absorbed by the meaning of the whole, by its expressiveness and emotional properties. For example, the phraseological unit *once in a blue moon* means *very seldom*. This type of units are also called idioms by which linguists understand a complete loss of the inner form.

A phraseological combination, also called collocation, is a construction or an expression in which one of the components has a bound meaning while the other word has an absolutely clear independent meaning. It means that phraseological combinations contain one component used in its direct meaning while the other is used figuratively. For example, the phraseological unit *to make an attempt* means *to try*, or *to offer an apology* means *to beg pardon*.

Finally, we define the linguistic structure which is the aim of this paper. A Verbal Phraseological Unit (VPU) is a PU that contains one verb as the grammar center. For example, the PU *to come to one's sense* means *to change one's mind*, or *to fall into a rage* means *to get angry*. Taking into account the fact that verbal phrases have a paradigmatic rupture, we focus our attention on this type of phraseological units, a task that implies a very high-challenge research line in terms of semantic identification and classification of phraseological units.

The following section presents the experiments carried out attempting to detect whether or not a VPU is present in a raw text.

4 Experimental Results

Regarding our current advances in the task of automatic identification of Spanish verbal phraseological units, we have considered the Mexican newspaper domain and a number of Mexican verbal phraseological units, thus, firstly, we describe the lexical resources constructed for the proposed task. The employed VPU identification approach is based on supervised machine learning techniques, a branch of artificial intelligence that concerns the construction and study of computational systems that can learn from supervised data, and therefore, we also include a section describing the classifiers used in the experiments. Finally, we present and discuss the results obtained in the experiments we carried out.

4.1 Dataset

Supervised machine learning methods assume that we have supervised data from which the methods can learn knowledge. In this case, we need corpora manually annotated by experts indicating whether or not a certain text contains a verbal phraseological unit. Thus, we constructed a dataset for the experiments proposed in this paper by selecting a number of news stories (from a Mexican newspaper) having and not having verbal phraseological units. In order to build the dataset, firstly, we extracted all the verbal phraseological units from a dictionary named "Dictionary of Mexicanisms"[5]. In particular, we have collected 1,219 verbal phraseological units from this dictionary which have been stored in a database, considering them to be further employed for identifying their regular use in the Mexican newspaper domain. For the purpose of the experiments reported in this paper, we have selected only the most representative ones, which in this case resulted to be 69 VPUs. In order to select those VPUs, we have taken into account their frequency of occurrence in the corpus, selecting at the end the most frequent ones.

By using information retrieval techniques we have found 3,164 news stories containing at least one occurrence of some of the selected verbal phraseological units. This process considers the occurrence of original VPU in any of it morphological variants; for this purpose, we have lemmatized both, the VPU and the text in the news story, so that we can be able to find variations of the VPU in the target text. The news stories have been gathered from Mexican newspapers belonging to the Mexican Editorial Organization[6]. All the texts compiled are written in Mexican Spanish and contain news stories that occurred between the years 2007 and 2013.

As a consequence of counting the occurrence of Mexican verbal phraseological units in the corpus gathered, we were able to construct a labeled corpus which may be further used as a training corpus for supervised machine learning methods with the aim of identifying whether or not a news story contain a VPU. The context gathered has been manually annotated by 5 human annotators with an inter-annotators agreement greater than 80%. Each human annotator was asked to manually classify when a given raw text contained a VPU (Class 1), or when that text did not contained a VPU (Class 2). The description of the corpus employed is given in Table 1.

Table 1. Description of the manually annotated corpus

Feature	Class 1 (VPU)	Class 2 (¬VPU)	Total
Instances	1,959	1,205	3,164
Tokens	117,715	63,600	181,315
Vocabulary	16,359	10,817	20,953
Minimum length	3	3	3
Maximum length	2,291	302	2,291
Average length	60.09	52.78	57.31

In the performed experiments, all the texts were represented by means of a vector of n-gram frequencies, with $n = 1, 2,$ and 3. Frequencies greater than two for the n-grams were only considered for the vector features. The corpus was used as both training and test corpus by means of a v-fold cross

[5] http://www.academia.org.mx/

[6] http://www.oem.com.mx/

validation process (v=10). The results obtained in the experiments are shown in Section 4.3.

4.2 Description of the Classifiers Employed

Supervised machine learning techniques are able to learn the human process of identifying verbal phraseological units based on features fed in the classifier by means of a manually annotated corpus. In order to have a perspective of the type of classifier that can best deal with the problem of automatic detection of VPUs, we have selected one learning algorithms from four different types of classifiers: Bayes, Lazy, Functions, and Trees. The following four learning algorithms were chosen:

NaïveBayes: a standard probabilistic Naïve Bayes classifier.

K-Star: a k-nearest neighbor classifier with a generalized distance function.

SMO: a sequential minimal optimization algorithm for support vector classification.

J48: a C4.5 decision tree learner which implements the revision 8 of C4.5.

The results we obtained identifying VPUs in news stories are as follows.

4.3 Obtained Results

In this section we present the accuracy obtained by each classifier when attempting to identify whether or not a VPU occurs in a given raw text. We have used the following standard measures for the evaluation: TP Rate (True Positive Rate), FP Rate (False Positive Rate), Precision, Recall, and F-Measure [5].

Tables 2, 3, 4, and 5 show the detailed accuracy by class using the Naïve Bayes, KStar, SMO, and J48 supervised classifiers, respectively. As it can be seen, in all the cases the identification of Class 1 (when the set of words is a VPU) obtained a better performance than the identification of Class 2 (when the set of words is not a VPU). Even if the difference is not so significant, this issue is important. As future work, we need to provide better features for improving the obtained results. On the one hand, the KStar classifier was the one that obtained the worst results with a weighted average F-Measure of 0.712. On the other hand, the J48 classifier obtained the best results with a weighted average F-Measure of 0.766. In this case, J48 obtained a TP Rate, Precision, Recall, and F-Measure accuracy greater than 0.8 for Class 1, which we consider acceptable.

Table 2. Detailed accuracy by class using the Naïve Bayes classifier

Class	Precision	Recall	F-Measure
Class 1 (VPU)	0.788	0.795	0.791
Class 2 (¬VPU)	0.662	0.651	0.657
Weighted Avg.	0.740	0.741	0.740

Table 3. Detailed accuracy by class using the KStar classifier

Class	Precision	Recall	F-Measure
Class 1 (VPU)	0.771	0.760	0.765
Class 2 (¬VPU)	0.618	0.633	0.626
Weighted Avg.	0.713	0.711	0.712

Table 4. Detailed accuracy by class using the SMO classifier

Class	Precision	Recall	F-Measure
Class 1 (VPU)	0.800	0.801	0.801
Class 2 (¬VPU)	0.676	0.675	0.676
Weighted Avg.	0.753	0.753	0.753

Table 5. Detailed accuracy by class using the J48 classifier

Class	Precision	Recall	F-Measure
Class 1 (VPU)	0.801	0.831	0.816
Class 2 (¬VPU)	0.707	0.664	0.685
Weighted Avg.	0.765	0.767	0.766

In Table 6 we show the percentage of instances classified correctly and incorrectly. Basically, this table summarizes the weighted average results in the previously shown result tables. Actually, we have included a graph (see Figure 1) with the aim of showing in a more clear way the summarized average results.

Table 6. Percentage of correctly vs. incorrectly classified instances

Classifier	Type	Correct (%)	Incorrect (%)
Naïve Bayes	Bayes	74.05	25.95
K-Star	Lazy	71.14	28.86
SMO	Functions	75.32	24.68
J48	Trees	76.74	23.26

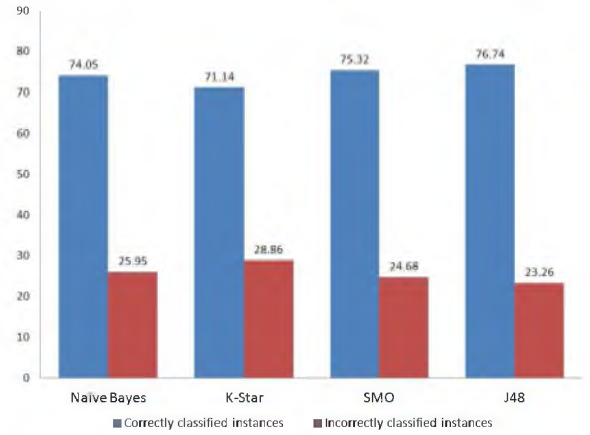

Fig. 1. Comparison among the different classifiers employed in the experiments

4.4 A Lexicon of VPUs with Probabilities

The news stories were collected from the web by means of an information retrieval system, employing the candidate VPUs as input query. Thus, we obtained texts from Internet which may or may not contain a VPU (see Table 1). In other words, the distribution of occurrence of a given VPU can be approximated by counting the number of times the candidate phrase is a VPU, and the number of times this sequence of words it is not a VPU. By doing so, it is possible to estimate the probability of a given sequence of words (candidate VPU) of being a VPU in real texts. This lexical resource may be of high benefit for the computational linguistic community since, to the best of our knowledge, no restricted domain corpora have been constructed, or at least the existing corpora have not been considered with that amount of data. We provide the community with a public access to this lexical resource by requesting it from any of the authors of this paper. Up to now, this lexicon contains only 69 entries, because we have selected only the most frequent VPUs from the total we collected from the above mentioned dictionary of mexicanisms; however, as further work we plan to apply exactly the same methodology for introducing more entries to this lexicon. A sample of the entries of this lexicon is shown in Table 7.

Table 7. Lexicon of VPUs with probabilities of being vs. not being VPU in the news domain context

Verbal phraseological unit	P(VPU)	$P(\neg$VPU)
darse por vencido (*to give up*)	0.49	0.51
salir a flote (*to keep one's head above water*)	0.83	0.17
comer el mandado (*to take advantage of*)	0.94	0.06
pegar su chicle (*to catch somebody's eye*)	0.95	0.05
dar el ancho / no dar el ancho (*to be (not be) capable*)	0.95	0.05
no quitar el dedo del renglón (*to take 'no' for an answer*)	0.99	0.01
dejar colgado (*to let someone down*)	0.59	0.41
ponerse la camiseta (*to put one's back into it*)	0.57	0.43
valer madre (*to be worthless*)	0.98	0.02
echar porras (*to encourage someone*)	0.52	0.48

5 Conclusions and Further Work

In this paper we presented a set of experiments towards the identification of the presence of verbal phraseological units in raw texts. We compared four different supervised classifiers with the aim of determining whether or not there exist significant differences among the results obtained by applying each supervised classifier. The revision 8 of the C4.5 decision tree learner obtained the best results for the task presented in this paper, obtaining 0.766 of F-measure. We are still interested in improving the obtained performance by analyzing other features which can be used in the classification process, this issue is considered as future work.

An additional interesting contribution was the construction of a lexicon of 69 VPUs, each one annotated with an estimate of its probability of being

a VPU in the news story domain[7]. As future work, we plan to increase the number of entries in this interesting lexicon.

Acknowledgements

This paper has been partially supported by the CONACyT grant with reference #218862/314461 and CONACyT Project #225784.

References

1. **Church, K.** (**2013**). How many multiword expressions do people know? *ACM Trans. Speech Lang. Process.*, Vol. 10, No. 2, pp. 4:1–4:13.
2. **Cook, P., Fazly, A., & Stevenson, S.** (**2008**). The VNC-Tokens Dataset. *Proceedings of the LREC Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco.
3. **Coseriu, E.** (**1977**). *Principios de semántica estructural*. Gredos.
4. **Davis, A. R. & Barrett, L.** (**2013**). Lexical semantic factors in the acceptability of English support-verb-nominalization constructions. *ACM Trans. Speech Lang. Process.*, Vol. 10, No. 2, pp. 5:1–5:15.
5. **Fawcett, T.** (**2004**). Roc graphs: Notes and practical considerations for researchers. Technical report, HP Labs.
6. **Fazly, A. & Stevenson, S.** (**2006**). Automatically constructing a lexicon of verb phrase idiomatic combinations. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 337–344.
7. **Levin, B.** (**1993**). *English verb classes and alternations: a preliminary investigation*. Chicago Press, University.
8. **Martinez-Blasco, I.** (**2008**). Verbos soporte y fijación lexica. *Las construcciones verbo-nominales libres y fijas*, pp. 47–59.
9. **Mejri, S.** (**1997**). *Le figement lexical. Descriptions linguistiques et structuration sémantique*. Publications de la faculté des lettres de Manouba, Tunis.
10. **Mogorron Huerta, P.** (**2010**). Estudio contrastivo lingüístico y semántico de las construcciones verbales fijas diatópicas mexicanas/españolas. *Quaderns de filología de estudis linguistics*, pp. 179–198. Universitat de València.
11. **Nissim, M. & Zaninello, A.** (**2013**). Modeling the internal variability of multiword expressions through a pattern-based method. *ACM Trans. Speech Lang. Process.*, Vol. 10, No. 2, pp. 7:1–7:26.
12. **Riehemann, Z., Wasow, T., Copestake, A. A., Clark, E. V., & Zwicky, A. M.** (**2001**). A constructional approach to idioms and word formation. Technical report, Stanford University. Dept. of Linguistics.
13. **Sfar, I.** (**2008**). Polylexicalite et continuite prédicative: le cas des locutions verbales figées. *Las construcciones verbo-nominales libres y fijas. Aproximación contrastiva y traductológica*, pp. 213–221.
14. **Stevenson, S., Fazly, A., & North, R.** (**2004**). Statistical measures of the semi-productivity of light verb constructions. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–8.
15. **Vincze, V., Nagy T., I., & Zsibrita, J.** (**2013**). Learning to detect English and Hungarian light verb constructions. *ACM Trans. Speech Lang. Process.*, Vol. 10, No. 2, pp. 6:1–6:25.

[7]The lexicon has been provided freely available for research purposes to any person who requests it from any of the authors of this paper, considering this paper as the corresponding reference for any user of the lexical resource.

Belem Priego Sánchez obtained her Master degree in Computer Science from the Benemérita Universidad Autónoma de Puebla, México, in 2012. She is actually a Ph.D. student at the LDI laboratory of the University of Paris XIII, France, and strongly collaborates with the LKE research group of the BUAP university in Mexico. Her areas of interest include computer science, phraseology, lexical acquisition of multiword expressions (MWEs) for natural language processing applications, corpus linguistics, and computational linguistics in general.

David Eduardo Pinto Avendaño obtained his Ph.D. in Computer Science in the area of Artificial Intelligence and Pattern Recognition from the Polytechnic University of Valencia, Spain, in 2008. He is actually a full time professor at the Faculty of Computer Science of the Benemérita Universidad Autónoma de Puebla in which he is the current

leader of the Language & Knowledge Enginering Research Group. His areas of interest include clustering, information retrieval, crosslingual NLP tasks, and computational linguistics in general.

Article received on 29/01/2015; accepted on 05/03/2015.
Corresponding author is Belem Priego Sánchez.