Towards the Automatic Identification of Spanish Verbal Phraseological Units

Belém Priego Sánchez[1,2], David Pinto[2] and Salah Mejri[1]

[1]LDI, Université Paris 13, Sorbonne Paris Cité
Paris, France
belemps@gmail.com, smejri@ldi.univ-paris13.fr

[2]Benemérita Universidad Autónoma de Puebla, FCC
Puebla, Pue., México
dpinto@cs.buap.mx

Abstract. Verbal Phraseological Units are expressions made up of two or more words in which at least one of these words is a verb that plays the role of the predicate. Their main attribute is that this form of expression has taken on a more specific meaning than the expression itself. The automatic recognition of this type of linguistic structures is a very important task, since they are the standard way of expressing a concept or idea. This paper describes the outgoing advances of a PhD research work in which it is attempted to construe a methodology which allows to automatically identify these linguistic structures for the Mexican Spanish language. It is presented a set of hypotheses which will allow to produce novel proposals in the way of automatically identifying a verbal phraseological unit in a raw text. Additionally, we have presented experiments carried out in this sense, for example, by employing machine learning methods. Finally, we show a lexical resource which is product of the current advances in this PhD thesis.

Keywords: Verbal phraseological units, Automatic identification, Corpus linguistics.

1 Introduction

Some concepts are expressed in language through set of words or phrases, which intuitively are employed native speakers, thus characterizing different cultural communities. Phraseology, considered a cultural heritage of the linguistic community [8], aims to study these blocks of words, which are usually refered as phraseological units.

The study of phraseological units has 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 [11], a fact that involves different dimensions of the natural language: linguistics, pragmatics, culturals, among others [12]. 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.

In this research work, we are particularly interested in studying mexican spanish phraseological units containing one verb as the grammar centre, i.e., Verbal Phraseological Units (VPU) which present a challenged degree of fixation in comparison with other phraseological units [18], for example, "*Leer entre líneas (To read between the lines)*". Actually, this paper aims to present an overview of the doctoral research that mainly focuses on identifying whether or not a verbal phraseological unit is present in a given text, a process that implies to analyze raw text and the features in the context of the phraseological unit for creating computational algorithms that allow to fullfil the mentioned task in environments highly scalable.

The remaining of this paper is structured as follows. Section 2 describes the motivation and shows a number of hypotheses we are proposing which we consider useful for the development of this research work. Section 3 presents some relevant works found in literature. Section 4 shows the results obtained in the experiments carried out in this research work, presenting first the lexical resources (lexicon and corpus), and later the results obtained up to now. Finally, in Section 5 the conclusions and perspectives of this research work are given.

2 Motivation and Advances

Phraseological Units (PU) are multiword lexical units that are characterized for presenting 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., the words together can mean more than their sum of parts [10, 6].

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 flexible and heterogeneous nature. Furthermore, phraseological units are not nearly as frequent in lexical resources as they are in real-world text, and this problem of coverage may impact the performance of many natural language processing tasks.

Phraseological units are widely employed by human beings. In [7], it is said that the number of phraseological units (there expressed as multiwords expressions, in the terminolgy used by the author), is in the same order of magnitude as the number of simple or isolated words.

A Verbal Phraseological Unit (VPU) is a PU that contains one verb as the grammar centre. For example, the PU *to come to one's sense* means *to change*

[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.

one's mind, or *to fall into a rage* means *to get angry*. Verbal phraseological units perfectly illustrate the overall saturation as indicated in [13]. Taking this characteristic into account and also the fact that verbal phrases have a paradigmatic rupture, make us focus our attention in 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.

In this way, is it very important to study the nature of such linguistic structures, so that we can be able to construe automatic methods for dealing with those units. In particular, we have centered our research in a set of hypotheses associated to the behaviour of verbal phraseological units. These hypotheses will determine the reseach path of this PhD thesis.

- *Fixation hypothesis.* The fixer a verbal phrase is, the higher it is its likelihood of being a verbal phraseological unit. We will substitute each component of a target verbal phrase (candidate VPU) with their near synonyms in order to verify if the new verbal phrase loses the meaning. In such case, it will be considered that a fixation phenomenon exist, thus verifying that the candidate VPU is a real VPU. In order to verify the meaning of the new verbal phrase, we are considering to use a reference corpus in which we may search the evidence of such phrase.
- *Translation hypothesis.* The more literal the translation of a verbal phrase is, the lower it is its likelihood of being a verbal phraseological unit. We will translate the verbal phrase from one language to another one. Thereafter, we will look for evidence of such translation in a reference corpus written in the target language.
- *Internal attraction and contextual co-reference hypothesis.* The greater the internal attraction and the lower the contextual co-reference in a verbal phrase are, the higher it is the likelihood of the verbal phrase of being a verbal phraseological unit. We will employ statistical methods for determining the level of internal attraction and contextual co-reference between the terms of the verbal phrase and those of their context.
- *Terminology domain hypothesis.* The greater the number of vocabulary terms out of the current domain for the verbal phrase is, the higher it is its likelihood of being a verbal phraseological unit. Is quite common to employ terms out of the current domain in a real VPU, therefore, we will identify terminology out of the current domain in order to determine whether or not, such verbal phrase is a real VPU.

We are currently working in the different methods for validating or rejecting the aforementioned hypotheses. In the following sections we will show the current advances with respect to this research work.

3 Related Work

There exist at least three different methods for recognizing phraseological units in a raw text: the application of constructed "local" grammars, the employment

of dictionaries, and the use of statistical processes [1]. Dictionaries are the common method for localizing phraseological units, but it is insufficient when we are considering to discover new fixed expressions, i.e., those that were not stored in the dictionary. Constructing local grammars is a knowledge based method that provides a broad reach because it may find new PU's which have similar linguistic structures than those considered when the grammars were constructed. Statistical methods usually employ document term frequencies for searching evidence of linguistic phenomena. For example, in [16] it is presented a statistical-based method for detecting verbal phraseological units; this paper includes a particular procedure for constructing lexical resources which may be later employed for machine learning methods.

Eventhough we are using the term phraseological unit in this paper, we are aware that there are a number of research works in which the term multiword expression (MWE) is employed. The following references are examples of such works [2, 4, 9, 20, 14].

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 we can be able to find many interesting papers. However, in literature there are other papers employing other terminology which refer to phraseological units, therefore, we mention some of them as follows.

In [19] 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 find 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. presents the VNC-Tokens dataset, a resource of almost 3000 English verb-noun combination usages annotated as to whether they are literal or idiomatic [3]. This authors began with the dataset used by Fazly and Stevenson [5], which includes a list of idiomatic verb-noun combinations (VNCs), and they 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 a VNC tends to be expressed in a small number of preferred lexico-syntactic patterns, referred to as canonical forms [17].

In [5], the authors investigate the lexical and syntactic flexibility of a class of idiomatic expressions. They 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. They also propose a process for automatically determining which syntactic

[3] http://multiword.sourceforge.net
[4] http://www.natcorp.ox.ac.uk/

forms a particular idiom can appear in, and hence should be included in its lexical representation.

We consider that other works associated to the identification of phraseological units exist in literature, for example, in [15] a set of experiment towards the identification of polarity has been presented, however, the exhaustive discussion of the state of the art is out of the scope of this paper.

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

4 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 VPU identification approach employed is based in supervised machine learning techniques, a branch of artificial intelligence that concerns the construction and study of computational systems that can learn from supervised data.

The 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 the 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 algorithm from four different types of classifiers: Bayes, Lazy, Functions and Trees. The obtained results are discussed in Section 4.2, and the lexical resource obtained up to now is presented in Section 4.3.

4.1 Dataset

Supervised machine learning methods assume that we have supervised data from which they 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 do so, 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 they to be further employed for identifying their regular use in the Mexican newspaper domain. For the purpose of the experiments carried out in this paper, we have selected only the most representative ones, which in this case resulted to be 56 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.

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

By using information retrieval techniques we have found 3,164 news stories containing at least one occurrence of some of the verbal phraseological units selected. This process considers the occurrence of original VPU 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 contains a VPU. The context gathered have been manually annotated by 5 human annotators with an agreement inter-annotators greater than 80%. Each human annotator was asked to manually clasify 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 shown 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 experiments carried out, all the texts were represented by means of a vector of n-grams 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 validation process (v=10). The results obtained in the experiments are shown in Section 4.2.

4.2 Obtained Results

In this section we are presenting the accuracy obtained by each classifier when attempting to identify whether or not a VPU exist in a given raw text.

In Table 2 we show the percentage of instances classified correctly and incorrectly. Basically, this table summarizes the weighted average results of the previously shown result tables. As it can be seen, the results obtained are highly enough to be seriously considered in the process of automatic detection of verbal

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

phraseological units in raw texts. All the classifiers have obtained a percentage above 71%. It is the J48 implementation of C4.5 that has generated a decision tree able to classify correctly the 76.74% of instances.

Table 2. Percentage of correctly vs. incorrectly instances classified

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

4.3 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 contain or not a real VPU inside (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 really a VPU, and the number of times this sequence of words is not a real VPU. By doing so, it is possible to estimate the probability of a given sequence of words (candidate VPU) of being a real VPU in real texts. This lexical resource may be of high benefit for the computational linguistic community since, up to our knowledge, they have not been constructed for restricted domain corpora, or at least they have not been considered with that amount of data. We then, provide public access to this lexical resource to the community, by requesting it to any of the authors of this paper. Up to now, this lexicon contains only 56 entries, because we have selected only the most frequent VPUs from the total we have collected from the above mentioned dictionary of mexicanisms; however, as further work we are planning 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 3.

5 Conclusions

In this paper we have presented advances towards the automatic identification of the presence of verbal phraseological units in raw texts. We consider particularly important, the set of hypotheses proposed, because they will lead the current research of this PhD thesis. We are very interested in obtaining feedback about these hypotheses, and thus this is the reason of presenting this paper in this forum.

Additionally, as a manner of example, we have presented an experiment in which we compared four different supervised classifiers with the aim of determining whether or not exist significant differences among the results obtained

Table 3. Lexicon of VPUs with probabilities of being vs not being VPU in the news domain context

Verbal phraseological unit	**Probability of being a real VPU** P**(VPU)**	**Probability of not being a real 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
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

by applying each supervised classifier in the process of automatic identification of VPU's in raw texts. The revision 8 of the C4.5 decision tree learner obtained the best results for the task executed in this paper, obtaining 76.74% of accuracy. We still interested in improving the performance obtained by analyzing other features which can be used in the classification process, this issue will be considered as future work.

An additional interesting contribution was the construction of a lexicon of 56 VPUs, each one containing an estimate of its probability of being a real VPU in a news stories domain[7]. As future work, we are planning to increase the number of entries to this interesting lexicon.

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

References

1. Buvet, P.A.: Vers l'elaboration d'un dictionnaire unique des prédicats du français : DEESSE. Dictionnaire Electronique Syntactico-Sémantique. In: Description linguistique pour le traitement automatique du français. pp. 23–42 (2008)
2. Church, K.: How many multiword expressions do people know? TSLP 10(2), 4:1–4:13 (2013)

[7] The lexicon has been provided freely available for research purposes to any people that request it to any of the authors of this paper, considering this paper as the corresponding reference for every one that use the lexical resource.

3. Cook, P., Fazly, A., Stevenson, S.: The VNC-tokens dataset. In: Proceedings of the MWE workshop ACL. pp. 19–22 (2008)
4. Davis, A.R., Barrett, L.: Lexical semantic factors in the acceptability of english support-verb-nominalization constructions. ACM Trans. Speech Lang. Process. 10(2), 5:1–5:15 (2013)
5. Fazly, A., Stevenson, S.: Automatically constructing a lexicon of verb phrase idiomatic combinations. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL). pp. 337–344 (2006)
6. Huerta, P.M.: Estudio contrastivo lingüístico y semántico de las construcciones verbales fijas diatópicas mexicanas/españolas. In: Las construcciones verbo-nominales libres y fijas. pp. 179–198 (2010)
7. Jackendoff, R.: The Architecture of the Language Faculty, vol. 28. MIT Press (1997)
8. Lamiroy, B.: Les expressions figées: á la recherche d'une définition. In: Blumental et Mejri 2008. pp. 85–98 (2008)
9. Levin, B.: English verb classes and alternations : a preliminary investigation. Chicago Press, University (1993)
10. Martinez-Blasco, I.: Verbos soporte y fijación lexica. In: Las construcciones verbo-nominales libres y fijas. pp. 47–59 (2008)
11. Mejri, S.: Le figement lexical. Descriptions linguistiques et structuration sémantique. In: Publications de la faculté des lettres de Manouba, Tunis (1997)
12. Mejri, S.: Catégories linguistiques et étiquetage de corpus. In: L′information grammaticale, Peeters, Paris (2007)
13. Mejri, S.: Constructions à verbes supports, collocations et locutions verbales. In: La traduction des Mejri Salah (2008)
14. Nissim, M., Zaninello, A.: Modeling the internal variability of multiword expressions through a pattern-based method. ACM Trans. Speech Lang. Process. 10(2), 7:1–7:26 (2013)
15. Priego Sánchez, B., Pinto, D., Mejri, S.: Evaluating polarity for verbal phraseological units. In: Human-Inspired Computing and Its Applications - 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I. pp. 191–200 (2014)
16. Priego Sánchez, B., Pinto, D., Mejri, S.: Metodología para la identificación de secuencias verbales fijas. Research in Computer Science 85(1), 45–56 (2014)
17. Riehemann, Z., Wasow, T., Copestake, A.A., Clark, E.V., Zwicky, A.M.: A constructional approach to idioms and word formation. Tech. rep., Stanford University. Dept. of Linguistics (2001)
18. Sfar, I.: Polylexicalite et continuite prédicative: le cas des locutions verbales figées. In: Las construcciones verbo-nominales libres y fijas. Aproximación contrastiva y traductológica. pp. 213–221 (2008)
19. Stevenson, S., Fazly, A., North, R.: Statistical measures of the semi-productivity of light verb constructions. In: Proceedings of the Workshop on Multiword Expressions: Integrating Processing. pp. 1–8. MWE '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
20. Vincze, V., T., I.N., Zsibrita, J.: Learning to detect english and hungarian light verb constructions. ACM Transactions on Speech and Language Processing 10(2), 6:1–6:25 (2013)