An Unsupervised Method for Automatic Validation of Verbal Phraseological Units *

Belém Priego Sánchez1,2, David Pinto2, Darnes Vilariño2

1LDI, Université Paris 13, Sorbonne Paris Cité, 99 avenue Jean-Baptiste Clément, Paris, France, belemps@gmail.com
2FCC, Benemérita Universidad Autónoma de Puebla, Av. San Claudio y 14 Sur, Col. San Manuel, Puebla, México, {dpinto, darnes}@cs.buap.mx

Abstract. In this paper we present an unsupervised technique for validating the existence of verbal phraseological units in raw text. This technique employs the concept of internal and contextual attraction which basically considers a mathematical formula based on co-occurrence of terms inside and outside of the terms considered to be part of a verbal phraseological unit. The experiments carried out using a corpus of news stories report a 60% of accuracy, which highlights the challenging task of automatic validation of verbal phraseological units in raw texts.

Key words: Unsupervised methods, Term co-occurrence, Phraseological units

1 Introduction

The aim of this research work is to validate the occurrence of verbal phraseological units in a particular textual genre, named news stories. Validation and identification of this type of linguistic structures is considered to be a different task, because the former assumes that a set of words can be a verbal phraseological unit and thus, the aim is to validate this hypothesis, whereas the latter needs to figure out how to discover which words of a raw text made up in fact a Verbal Phraseological Unit (VPU). In [9], for example, we

* This research work has been partially supported by CONACYT.
can see experiments towards the automatic identification of VPUs. However, in this paper we aim to validate the occurrence of VPUs in raw text through a mathematical model based on two concepts: the phrasal internal attraction and the contextual attraction. One of the objectives of the proposed formula is to allow a generalized use of the model among different languages in an unsupervised manner, because it only relies on computing frequencies of word co-occurrences. In fact, this paper presents experiments using only the Spanish language, but the proposed approach can easily be employed in other languages, such as French and English, by using another corpora to estimate the frequency of word bigrams.

The study of verbal phraseological units is very important because these linguistic structures are not completely covered in linguistic resources such as lexicons and dictionaries as they are employed in real world documents. This covering problem may have a negative effect when the VPUs are required for solving some natural language processing problem. Thus, in order to support the computational linguistic community, we face the problem of automatic validation of VPUs by first constructing a lexical resource of VPUs\textsuperscript{1} which contains occurrences of VPUs in different contexts of news stories. We used the news genre because it is considered to be a type of standard writing, i.e., the news stories genre is written to be understood by almost all the people who speak the language in which the news has been written, in this case, the Mexican Spanish. Additionally, evidence exists that news stories are a good referent for the automatic analysis of this type of phraseological units [10].

The occurrence of the words belonging to a given VPU does not necessarily imply that the news story has a VPU, because in some cases, it is a literal interpretation of the words, and here it shows importance of having automatic methods for automatic validation of this type of linguistic structures. Let us take as example the two sentences shown in Table 1.

\textsuperscript{1} This lexicon is freely available by requesting it to any of the authors of this paper.
The first example given in Table 1 employs the verbal phrase “agarrar de bajada” in the literal manner (“to go down”), whereas the second example indicates another meaning, different than the literal one, which means that the person has been abused by other people.

To the best of our knowledge, there is no research work in literature reporting automatic and unsupervised methods dealing with this particular and important problem. The closest research paper presents a machine learning based method for automatic validation of VPUs (see [9]). It is very important to emphasize that there are many works reporting the automatic detection of idioms, collocations, etc [7] but none of them guarantee that the set of words found is in fact a verbal phraseological unit in the particular context in which they appear.

The remaining of this paper is structured as follows. Section 2 provides an overview of the concept of phraseological units, together with a review of some research works found in literature dealing with the extraction of this type of linguistic structures. Section 3 presents a description of the concepts of phrasal internal at-
traction and contextual attraction, both used for validating the occurrence of VPUs in raw texts. In the same section we present the proposed approach for the validation process. Section 4 shows the results obtained in the experiments carried out, presenting first the lexical resources constructed (lexicon and corpus), and later the results obtained until now. Finally, in Section 5 we give the conclusions and future work for this research work.

2 Verbal Phraseological Units

Textual information can be decomposed in different components, for example a textual document contains paragraphs, each of these paragraphs is made up of sentences which are also decomposed in small units named phrases. Each phrase has information that can be used as an independent unit [13]. The majority of these phrases are made up of one verb and one or more variables. The verb requires a rigorous selection of the subjects and components around it. These phrases are fused in the sentence to state something more broadly, but when they are seen in a separate way from the sentence, they have a full sense, i.e., they have a semantic meaning by themselves. This is another reason for studying this kind of linguistic structures.

The automatic analysis (including identification and validation) of verbal phraseological units is a very important and complex task for the particular area of computational linguistics, because they are used in real life by human beings since they are a standard manner to express a concept or an idea. In general, the approaches reported in literature dealing with the problem of automatic extraction of phraseological units can be classified into three different categories: a) Statistical approaches based on the computing of frequency of co-occurrence of these linguistic units; b) Knowledge based approaches that employ parsers, lexicons, or rules created by experts in the target language; c) Hybrid approaches that combine the two aforementioned methods.

In [3, 5, 12] we can observe some examples of the first category, in which the language is modelled as a stochastic process,
estimating the probabilities of the language model by means of a reference corpus. In [6, 1, 4] we can see examples of research works using approaches that can be classified into the second category. They study the internal structure of the phraseological unit and the relationship among its components. Finally, in [11, 8, 9] we can have some examples of those works falling down into the third category, i.e., hybrid approaches combining the two previous ones, either employing first statistical methods followed by knowledge based methods or vice versa.

The subject of this research study are the phraseological units, which are defined in [2] as a stable combination of two or more terms that functions as a phrase whose meaning cannot be justified as the compound normal meaning of its components. The phraseological units are characterized by at least the following three linguistic features:

1. **Polylexical meaning**: This feature distinguishes these linguistic structures from the isolated words, either single or compound words of the language.
2. **Fixation**: This feature means that some or all the components of the phraseological units are inseparable, i.e., those components cannot be replaced, for example, by quasi-synonyms.
3. **Idiomaticity or lexical opacity**: This feature refers when the meaning of a compound word or an idiom cannot be inferred from its parts or morphemes, i.e., there is not a semantic transparency. However, in some cases, this feature cannot be present as it happens with the so called collocations.

The traditional classification of phraseological units is based upon the criterion of the role of this unit in the sentence, taking into account the nuclear component of the unit: noun, adjective, adverb, verb, preposition, etc. In our case, we have been studying those units having a verb playing the main role of the predicate, named “verbal phraseological units”.

From a syntactic point of view, verbal phraseological units express processes and act as predicates with or without complements.
These phraseological units are combined with the subject and complements in order to construct a sentence. We consider them as idioms with non-compositional meaning whose interpretation is obtained by analysing all their components at the same time, i.e., the meaning of the VPU cannot be obtained by the sum of its components meaning. Some examples of these linguistic structures are: *to come to one’s sense* that means *to change one’s mind*, or *to fall into a rage* that means *to get angry*.

In the following section we describe the methodology proposed for validating verbal phraseological units in raw texts.

### 3 Methodology proposed

The rationale of the methodology proposed is given in this section. First, we must understand that the words that belong to a real VPU generally will appear together in a given context, a fact can be statistically measured by using large corpora. We can employ term co-occurrence methods for determining how frequent two given words appear together, and generalize this measure for all the words inside of the candidate VPU. The sum of the co-occurrence of the terms inside of the VPU will determine what we call the “internal attraction”.

On the other hand, we consider that the words belonging to the VPU will have a low level of co-occurrence with those words in the context of the VPU. Consider the example given in Table 1, in this case, the word “bajada (down)” will have a high level of co-occurrence with words like “subida (up)” when the meaning of the VPU “agarrar de bajada” is literal. So, we can propose a measure named “contextual attraction” considering this fact.

Thus, the process of automatic validation of verbal phraseological units is based on the following linguistic hypothesis:

*The greater the internal attraction and the lower the contextual attraction in a verbal phrase are, the higher it is the likelihood of the verbal phrase of being a verbal phraseological unit.*
In order to validate the hypothesis previously presented, we have employed statistical methods for determining the level of internal attraction and contextual attraction between the terms of the Verbal Phrase (VP) and those of its context.

Thus, for each word $p_i$ belonging to the verbal phrase, VP, we may estimate its Internal Level of Attraction ($ILA$) as it is shown in Eq.(1).

$$ILA(VP) = \frac{2}{n \ast (n-1)} \sum_{\forall p_i \neq p_j; p_i, p_j \in VP} P(p_i, p_j)$$  (1)

where $n$ is the number of words in the verbal phrase, and $P(p_i, p_j)$ is the probability of two words $p_i$ and $p_j$ to appear together.

On the other hand, the Contextual Attraction Level ($CAL$) of a text $T$ made up of: Left Context ($LC$), a VP, and the Right Context ($RC$) can be estimate using Eq.(2).

$$CAL(T) = \frac{1}{n_1 \ast n_2} \sum_{\forall p_i \in VP, p_j \in \{LC \cup RC\}} P(p_i, p_j)$$  (2)

where, $n_1$ is the number of words in the VP, and $n_2$ is the number of words in the context of the VP.

It can be said that, if $ILA(VP)$ is greater than $CAL(T)$, then it is highly the likelihood that the verbal phrase is indeed a verbal phraseological unit.

There can be plenty of manners of estimate $P(p_i, p_j)$ to measure the level of co-relationship of the words inside of the VP, however, in this paper we employ an estimate based on theory of language models. Thus, $P(p_i, p_j)$ can be estimate by using the frequency of the biword “$p_ip_j$” and the frequency of each single word ($p_i$ and $p_j$).

Formally, given two words $p_1$ and $p_2$, we estimate the likelihood of these two words to be together as it is shown in Eq.(3).

$$P(p_i, p_j) = \frac{freq(p_i, p_j)}{freq(p_i)}$$  (3)
where, \(freq(p_i, p_j)\) is the frequency of occurrence of the two words \(p_i\) and \(p_j\), in a context determined by a window size (usually, a sentence).

We are aware that the co-occurrence frequency may be calculated by means of other co-occurrences formulae, for instance, point-wise mutual information, t-score, etc., however, the aim of this paper is not to present an exhaustive manner of calculating co-occurrence between pair of words, but to propose a valid hypothesis that can be employed for validating verbal phraseological units in raw texts. An example is shown in Section 6 in order to illustrate how to calculate the proposed formula for a given verbal phraseological unit.

In the following section we present the experimental results that we have obtained by using the proposed formula, a description of the verbal phraseological units used and the description of the dataset of news stories written in Mexican Spanish employed in the experiments.

4 Experimental results

In this section we present the experiments carried out for validating verbal phraseological units for one diatopic variant of the Spanish language, named Mexican. As lexical resources we have employed a database of verbal phraseological units, one corpus of the news genre, both for the same language, i.e., Mexican Spanish. The obtained results are presented together with a discussion of findings when the formulae presented in Section 3 were used.

4.1 Dataset

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”\(^2\). In particular, we

\(^2\) http://www.academia.org.mx/
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.

By using information retrieval techniques, we have found 3,050 news stories containing at least one occurrence of some of the verbal phraseological units selected. This process considers the occurrence of the original VPU or any of its morphological variants; for this purpose, we have lemmatized both, the VPU and the text in the news story, so that we can find the variations of the VPU in the target text. The news stories have been gathered from Mexican newspapers belonging to the Mexican Editorial Organization\(^3\). All the texts compiled are written in Mexican Spanish and contain news stories that occurred between the years 2007 and 2013.

Thus, in order to validate the experiments, we have manually annotated the corpus gathered for determining whether or not the candidate verbal phrase is in fact a verbal phraseological unit.

### 4.2 Obtained results

In order to estimate the probabilities of biwords and single words in the given dataset, we have calculated the frequency of their occurrence in the corpus without considering the class tag, i.e., the training method only counts frequencies, but in any case it does know if a given sentence contains or does not contain a verbal phraseological unit.

Once we estimate these probabilities we are able to validate whether or not a verbal phrase is a verbal phraseological unit. The accuracy and precision of the approach is reported in Table 2, so basically we have employed formulae shown in Eq.(4) and Eq.(5), using the following values obtained in the experiments carried out:

- True Positives = correctly identified = 906
- False Positives = incorrectly identified = 533

\(^3\) http://www.oem.com.mx/
– True Negatives = correctly rejected = 926
– False Negatives = incorrectly rejected = 685

\[
\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}} \quad (4)
\]

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (5)
\]

**Table 2.** The results obtained in a Mexican News corpus

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>60.07%</td>
</tr>
<tr>
<td>Precision for class 1 (VPU)</td>
<td>62.96%</td>
</tr>
<tr>
<td>Precision for class 2 (NOT VPU)</td>
<td>57.48%</td>
</tr>
</tbody>
</table>

The accuracy obtained is 60% which we consider could be improved by gathering more contexts for the verbal phraseological units, so that we may better estimate the probabilities of occurrence of biwords and single words. However, given the complexity of the task, we consider that the accuracy obtained in our experiments as a good result.

As can be seen, the proposed method is able to detect true positives with a 62.96% of precision and true negatives with a 57.48% of precision. Both classes obtain similar values of precision, which is also interesting, because in some cases, this class of balanced results is not obtained, resulting in giving benefit to one class over the other one.

5 Conclusions and future work

In this paper we have proposed a formula for validating verbal phraseological units. This formula employs the concept of level
of internal attraction and the concept of contextual attraction. By bringing these two concepts together we were able to validate whether or not a given verbal phrase is a verbal phraseological unit with a 60% of accuracy.

As future work we would like to use a bigger corpus than the one used to estimate the probabilities required for computing the proposed formulas. We would also like to analyse the precision considering each verbal phraseological unit alone. Finally, even if it is a very hard work task because of the manual annotation, we would like to experiment with other languages.

References

In this section we present one example that illustrates how the proposed formulas work.

Consider the text $T$ = “Si me mientes, te parto tu mandarina en gajos, me dijo mi amigo” (which means “If you lie to me, i will kick your ass, said my friend”).

In $T$ we can identify the two contexts (left and right) and the verbal phrase:

- $VP$ = “parto tu mandarina en gajos” (“i will kick your ass”)
- $LC$ = “Si me mientes te” (“If you lie to me”)
- $RC$ = “me dijo mi amigo” (“said my friend”)

Taking into consideration the components detected, we may proceed to evaluate the hypothesis. In this exercise we will use only the words tagged as one of the following Part of Speech (PoS) tags: verbs, adjectives, nouns. Thus, the aforementioned components will be filtered and they will be expressed as follows:

- $VP$ = \{parto, mandarina, gajos\}
- $LC$ = \{mientes\}
- $RC$ = \{dijo, amigo\}

In order to calculate the Internal Level of Attraction, $ILA(VP)$, it is needed to estimate the following probabilities: $P(\text{parto}, \text{mandarina})$, $P(\text{parto}, \text{gajos})$ and $P(\text{mandarina}, \text{gajos})$. So, let us consider the following values:
- \(freq(\text{parto}) = 49,500,00\)
- \(freq(\text{mandarina}) = 8,010,000\)
- \(freq(\text{parto, mandarina}) = 54,450\)
- \(freq(\text{parto, gajos}) = 14,389,650\)
- \(freq(\text{mandarina, gajos}) = 1,649,259\)

Thus,
- \(P(\text{parto, mandarina}) = \frac{freq(\text{parto, mandarina})}{freq(\text{parto})} = 0.0011\)
- \(P(\text{parto, gajos}) = \frac{freq(\text{parto, gajos})}{freq(\text{parto})} = 0.2907\)
- \(P(\text{mandarina, gajos}) = \frac{freq(\text{mandarina, gajos})}{freq(\text{mandarina})} = 0.2059\)

With the previous estimate of probabilities, we can compute the final value for the internal attraction as

\[
ILA(VP) = \left( 0.0011 + 0.2907 + 0.2059 \right) / 10 = 0.04977
\]

In order to estimate the probability of two words to appear together in a given context, we can use Eq.(3), counting the number of times that both words appear together, and divide this value between the number of times that the first word appears (in the complete corpus). If an information retrieval system is available, we can estimate the frequency of the biword by just using the NEAR operator. In other case, we can use use commercial information retrieval systems such as google, yahoo, msn, etc, employing Eq.(3) for estimate the probability. For instance, \(P(\text{parto, mandarina})\) can be estimated by considering the phrase “parto su mandarina”, as follows:

\[
P(\text{“parto su mandarina”}) = \frac{freq(\text{“parto su mandarina”})}{freq(\text{“parto su”})}
\]

We also need to calculate the contextual attraction, \(CAL(T)\). In this case, the estimated value is calculated by using each one of the VP words against those of the VP context considering the number of words expected to be between these two words. For example, if we need to estimate the probability of \(P(\text{parto, mientes})\), it is needed to calculate both, the frequency of co-occurrence of the
two words “parto” and “mientes”, and the frequency of the word “parto” in the whole corpus. According to the structure of the text $T$ given, it is expected that the window of co-occurrence of the two words is about one (“mientes te parto”), so we need to take into account the window size when we compute the frequency.

So, let us consider the following values:

- $freq(\text{parto}) = 49,500,00$
- $freq(\text{mandarina}) = 8,010,000$
- $freq(\text{gajos}) = 560,000$
- $freq(\text{parto, mientes}) = 562$
- $freq(\text{mandarina, mientes}) = 61$
- $freq(\text{gajos, mientes}) = 1$
- $freq(\text{parto, dijo}) = 1,297$
- $freq(\text{mandarina, dijo}) = 3,204$
- $freq(\text{gajos, dijo}) = 20$
- $freq(\text{parto, amigo}) = 0$
- $freq(\text{mandarina, amigo}) = 0$
- $freq(\text{gajos, amigo}) = 37$

Then, we can estimate the probabilities of the co-occurrence between two words as follows:

- $P(\text{parto, mientes}) = \frac{freq(\text{parto, mientes})}{freq(\text{parto})} = 1.14 \times 10^{-5}$
- $P(\text{mandarina, mientes}) = \frac{freq(\text{mandarina, mientes})}{freq(\text{mandarina})} = 7.62 \times 10^{-6}$
- $P(\text{gajos, mientes}) = \frac{freq(\text{gajos, mientes})}{freq(\text{gajos})} = 1.79 \times 10^{-6}$
- $P(\text{parto, dijo}) = \frac{freq(\text{parto, dijo})}{freq(\text{parto})} = 2.62 \times 10^{-5}$
- $P(\text{mandarina, dijo}) = \frac{freq(\text{mandarina, dijo})}{freq(\text{mandarina})} = 0.0004$
- $P(\text{gajos, dijo}) = \frac{freq(\text{gajos, dijo})}{freq(\text{gajos})} = 3.57 \times 10^{-5}$
- $P(\text{parto, amigo}) = \frac{freq(\text{parto, amigo})}{freq(\text{parto})} = 0$
- $P(\text{mandarina, amigo}) = \frac{freq(\text{mandarina, amigo})}{freq(\text{mandarina})} = 0$
- $P(\text{gajos, amigo}) = \frac{freq(\text{gajos, amigo})}{freq(\text{gajos})} = 6.61 \times 10^{-5}$
And, therefore, the contextual attraction, can be computed as follows:

\[
CAL(T) = \frac{(1.1 \times 10^{-5} + 7.6 \times 10^{-6} + 1.7 \times 10^{-6} + 2.6 \times 10^{-7} + 0.0004 + 3.5 \times 10^{-5} + 0 + 0 + 6.6 \times 10^{-5})}{40}
\]

\[
CAL(T) = 0.0000137
\]

In summary, \(ILA(VP) = 0.04977\) and \(CAL(T) = 0.0000137\), therefore, it is possible to say that there exist a greater internal attraction among the terms of the verbal phrase, than the contextual attraction of those terms with respect to others that exist in the context of the verbal phrase. Thus, the verbal phrase “te parto tu mandarina en gajos” has a high likelihood to be a verbal phraseological unit.