Evaluating Polarity for Verbal Phraseological Units

Belém Priego Sánchez\textsuperscript{1,2}, David Pinto\textsuperscript{2}, and Salah Mejri\textsuperscript{1}

\textsuperscript{1} LDI, Université Paris 13, Sorbonne Paris Cité,
99 avenue Jean-Baptiste Clément, Paris, France
\{abpriegosanchez,smejri\}@ldi.univ-paris13.fr
\textsuperscript{2} FCC, Benemérita Universidad Autónoma de Puebla,
Av. San Claudio y 14 Sur, Col. San Manuel, Puebla, México
dpinto@cs.buap.mx

Abstract. Fixation in linguistic expressions is an inherent property of natural language that plays a central role in their description. Verbal phraseological units are phrases made up of two or more words characterized for presenting certain degree of fixation or idiomaticity (at least one of these words is a verb that plays the role of the predicate).

Phraseological units do not appear so frequently in manually constructed lexical resources as they do in real-word text, and this problem of coverage may impact the performance of many natural language processing tasks. Therefore, the construction of automatic understanding systems for these types of linguistic structures is very important, since they are a standard way of expressing a concept or idea. In this paper we present a set of experiments towards the automatic identification of the polarity of verbal phraseological units. We obtained a maximum performance of 80\% for this particular task when the contextual information of a phraseological unit is considered, in comparison with a 62\% when the VPU alone is only used. These results highlight the importance of analyzing automatically this type of linguistic structures. It should be stressed at the outset that these experiments are intended as a preliminary study rather than as a comprehensive analysis or solution of the aforementioned problem.

Keywords: Verbal phraseological units, Text polarity, Machine learning.

1 Introduction

Phraseological Units (PU), also known as phrasemes or fixed expressions, are basically multi-word lexical units that are characterized for presenting certain degree of fixation or idiomaticity. In other words, PU’s 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 \cite{1}.

\footnote{This paper has been partially supported by the CONACYT grant \#218862.}
Thus, PU’s almost never presents the following criteria: compositionality, substitutability and modifiability, therefore, avoiding any modification to its structure. A phraseological unit is a lexicalized, reproducible b lexemic or poly lexemic word group in common use, which has a relative syntactic and semantic stability. This type of linguistic structure may be idiomatized, may carry connotations, and may have an emphatic or intensifying function in a text.

Phraseological units are a stable group of words with partially or fully transferred meanings, for example, “Greek gift” (a gift given with the intention of tricking and causing harm to the recipient), “to kick the bucket” (to die) or “it is raining cats and dogs” (it is raining very hard).

Phraseological Units (PU) belong to what Coseriu [2] called “repeated discourse”, and they are mainly characterized by the following three features:

1. Their poly-lexical behaviour that distinguish them from isolated words of the language, either simple or compound words.
2. Their fixation degree, that presents them as 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.

The study of this type of linguistic expressions 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 [3], a fact that involves different dimensions of the natural language: linguistics, pragmatics, culturals, among others [4]. PU’s are not nearly as frequent in lexical resources as they are in real-word text, and this problem of coverage may impact the performance of many natural language processing tasks.

In this paper we focus our experiments on detecting the polarity of Verbal Phraseological Units (VPU’s). For this purpose, we have executed several experiments using various supervised classifiers, so that we can have an idea of the best performance we can obtain for determining the polarity of such linguistic expressions. The experiments were carried out using a corpus of VPU’s manually tagged with three possible classes (positive, negative and neutral).

The remaining of this paper is structured as follows. Section 2 presents the state of the art in determining polarity of verbal phraseological units. Section 3 describes the experiments carried out in this paper towards the automatic identification of the polarity of VPU’s. Finally, in Section 4 we give the conclusions and findings of this research work.

2 Verbal Phraseological Units

In this research work, we are particularly interested in studying spanish phraseological units containing one verb as the grammar nucleus, i.e., verbal phraseological units which present a high degree of fixation in comparison with other phraseological units [5], for example, Leer entre líneas (To read between the lines).