ISSN 2007-9737

A New Lexical Resource for Evaluating Polarity in Spanish Verbal Phrases

Belém Priego Sánchez

Universidad Autónoma Metropolitana,
Mexico

belemps@gmail.com

Abstract. Nowadays, a very small number of lexical resources that can be employed in the particular task of determining polarity for Spanish verbal phrases exist in literature and linguistic resource databases. Therefore, it is needed to increase efforts towards the construction of this kind of lexical resources since they will have a very positive impact when creating computational modules for the automatic understanding of Spanish language. In this paper, we present a new manual annotated corpus for the news genre, made up of sentences containing each a verbal phrase. The supervised corpus has been annotated with respect to polarity and the verbal phrase compositionality or non-compositionality, thus leading to have a very interesting lexical resource for identifying polarity in Spanish verbal phrases, so as to determine whether or not a given verbal phrase has a figurative or literal meaning. In addition, we present a lexicon of verbal phrases which indicates the estimate of it to be compositional.

Keywords. Polarity, verbal phrases, lexical resources.

1 Introduction

Managing information is crucial for decision making. The goal is to identify and choose alternatives based on preferences, beliefs and values of the person who takes the decision. The aim is to produce a final choice which normally may be executed immediately, but in some cases may not derive in a prompt action. Using information generated in real time by human beings creates a great opportunity of reducing the time for those final decisions to have a better impact than when the information processing takes more time.

The idea is to acquire data from social communities in internet and process the data in an automatic manner in order to generate statistical information about the collective opinion about some product, service or person. This kind of opinion is referred as opinion mining and normally it is expressed in terms of ’positive’, ’negative’ or ’neutral’.

Obtaining this kind of data for automatic analysis is highly feasible by consulting forums, blogs and social networks in which the target community publishes information, thus decreasing the cost and time involved in the opinion mining task.

Currently, many researchers in the field of information retrieval and computational linguistics have focused their research on social networks, especially on Twitter, given the large number of publications that are made daily.

In addition to these investigations, he finds the polarity analysis of texts in the journalistic domain, which, in the same way, identifies and helps to determine different characteristics of the events that occurred in the world. For this reason, efforts to identify polarity in informational texts (news) are reflected in this article. In this way, the creation of tools that can be managed, analyzed and manipulated all this available information is needed. In addition, of category according to the content generated by user, and - in relation to opinions - identify the nuances of opinion linked to the position of the users with respect to some subject. This is based on the constant monitoring of the messages produced on the Web,

doi: 10.13053/CyS-24-2-3409

from comments to scientific texts or from any other domain.

2 Sentiment Analysis

The Sentimental Analysis (SA), is a task of the text classification within of the natural language processing, its define as the computational study of the opinions, feelings and emotions express in texts [16]. The main objective consists in determinate the attitude of a writer to certain products, situations, people or organizations (*target*); identify the aspects that generate an opinion (*features*); who owns them (*holder*); and what is the type of emotion (I like it, I love it, I value it, I hate it) or its semantic orientation (positive, negative, neutral) [15]. There are different processing tasks that can be performed in the SA, a basic one is to classify the polarity of an opinion expressed in a text (document, sentence, feature or characteristic) into a binary classification, positive or negative. In addition, a more advanced classification of three classes (positive, negative or neutral). A slightly more complex task is the multi-classification of a text according to the degree of polarity of the attitude within a scale; and the most advanced task is the identification of the aspects mentioned in a text and their associated feelings.

2.1 State of the Art

This research is framed in the world of opinion mining, so that a state of the art related to work is presented below. Among the studies focused on the classification of opinions, positive or negative, is the one presented in [21], one of the first investigations on the sentiment analysis in which criticism data from films found on the Web are used; they are used in three classification algorithms, surpassing the baselines produced manually by a human. One of the pioneering works in introducing the term of Sentiment Analysis is that presented by [19]. In this publication, this task will be determined as finding expressions of feelings for a given subject and determining their polarity.

In [11], a proposal to undermine and summarize consumer reviews was presented; for which it was proposed the creation of a small list of adjectives "seed" labeled manually, if a positive or negative feeling is expressed.

Subsequently this list is augmented using WordNet [17]. A work based on a collection of blog entries, is presented in [8] that performs the sentiment analysis and opinion mining in various entries, showing the relevance of machine learning systems as a resource for the detection of opinion information.

Until 2014, modified systems in the constant monitoring of messages produced on social networks were evaluated within the framework of the RepLab competition [1] for tweets in English and Spanish. Despite including the Spanish language within the competition, most of the works reported in the literature focus on the English language. Therefore, the different methods applied for the classification of student opinions in English have been applied to the Spanish language [9]. These methods have considered since the use of n-grams of words, the reduction of words to their root (stemming) and even their substitution.

The SemEval forum (International Workshop on Semantic Evaluation) is undoubtedly a space that has spread the study of polarity in Tweets and where you can find different corpora associated with the sentiment analysis. Since 2013, it has had support for polarity assessment in Tweets. In particular, in 2017, 58 teams with different proposals were presented, which are summarized in [28]. The SemEval 2018 edition has stood out for having a section (with three tasks) dedicated exclusively to the analysis of affective characteristics in Tweets. A complete description of the 75 teams that participated in SemEval task 1 called "Affect in Tweets" can be seen in [18]. They describe the way in which the corpora were constructed and labeled manually for the English, Arabic and Spanish languages, as well as the techniques used in the 319 executions presented by the teams.

In the latest edition of SemEval 2019, different tasks have been presented that cover the sentiment analysis, having a section called "Opinion, emotion and abusive language detection" and that has had four different tasks that cover the detection of contextual emotions in text [6], detection of hyper-partisan news [13], the multilingual detection of hate speech and women on Twitter [2] and the identification and categorization of offensive language in social networks [32]. The participation of the different teams in the tasks is reported individually and that is why we refer the reader to the reports that summarize the participation of all the teams.

In [29], an approach is presented for opinion mining of tweets in Spanish; based on the operation and different configurations of machine learning algorithms. Although the algorithms used present good results for the English language, this work shows how the different sizes of n-grams, the length of the corpus, the number of kinds of feelings, the balanced corpus with respect to the unbalanced corpus and the different domains (configurations) affect the accuracy of the algorithm. The generation of word lexicons, which are annotated with their corresponding polarity, is another approach in which different researchers have been oriented and that has contributed to the monitoring of opinions in the comments. In the case of Spanish, examples of these approaches are those presented in [22, 4].

3 Phraseology

The phraseology, considered as the reflection of the folkloric cultural heritage of a linguistic community, has acquired in recent decades the status of a true object of research in theoretical linguistics [14]. Some of the works focused on Spanish phraseology are [5, 33, 7, 10]. All these authors agree that the basic unit of analysis of phraseology is the Phraseological Unit (PU), also called phraseology. Some authors [31, 20, 30] dedicated to the study of Phraseological Variants reiterate that they are preset, that is, that the variation is determined and limited, so it cannot be altered and is known by the speakers.

Phraseological units have morphological, syntactic and lexical changes. A morphological change is in which one of the components of the PU undergoes some alteration; these changes can be gender, number, quantification, determination. For example: the PU *más pobre que un perro (poorer that a dog)* can change by *más pobre que los perros (poorer that dogs)*; *romper en pedazos (break into pieces)* by *romper en mil pedazos (break into a thousand pieces)*, among others. Syntactic variants occur when changes or alterations arise in the order of the elements of the PU, but which do not influence the lexicalization of the PU. For example: the PU *mover cielo y tierra (move sky and earth)* can change by *mover tierra y cielo (move earth and sky)*. The most frequent phraseological variants are those that substitute a lexical element for another. For example: *me importa un pepino (I don't care about a cucumber)* by *me importa un comino (I don't give a damn)* or by *me importa rábano (I care about radish)*.

The classifications of phraseological units have emerged, most of them, as a result of the practical problems that the lexicographer has had to face when including phraseological information in the elaboration of dictionaries [7]. In [24], a more detailed study of this type of linguistic structures can be found.

3.1 Verbal Locutions

The locutions are defined by [5] as a "stable combination of two or more terms, which functions as a sentence element and whose known unitary sense is not justified, simply, as a sum of the normal meaning of the components". The different locution definitions in Spanish have followed this characterization. The locutions have been divided according to the sentence function they perform, regardless of whether they are commutable by simple words or by phrases. In [7], the following types of locutions are distinguished: nominal, adjective, adverbial, verbal, prepositive, conjunctive and clausal.

A Verbal Locution (denoted, hereinafter, by VL) is a PU that contains a verb at the center of its grammar. From the syntactic point of view, they express processes and act as predicates, with or without complements. These PUS, like the verbs, combine with the subject and the complements to form a sentence. In [27], they define an VL as a group of words in which at least one is a verb that functions as the nucleus of the predicate, that is, idiomatic expressions of non-compositional meaning. They are fixed and idiomatic phrases whose interpretation is not obtained from the sum of their parts, taken separately. Definition that is considered throughout the development of this research.

In [3], a grammar study associated with VL in Spanish is carried out. This study, "collects and analyzes a sample of Spanish verbal phrases clearly distinguishing them as such from other phraseological units that are often confused with them". In [23] an analysis of morphosyntactic diversity in verbal phrases in Spanish is performed. So if you want to deepen this type of phraseological units, we suggest that the mentioned work be consulted.

The following sections describe the processes, carried out, when building lexical resources for polarity in verbal locutions.

4 Construction of Lexical Resources

The general scheme for the construction of lexical resources associated with the polarity of verbal phrases, is composed of different stages. First, we will proceed to identify contexts with a possible verbal locution(see Section 4.1). Then, the annotation of the contexts where the phenomenon of compositionality and non-compositionality can be perceived (see Section 4.2) and resulting in a manually annotated corpus of the journalistic domain, which contains a set of contexts in which a candidate verbal locution is present. Finally, and derived from the news tagging manual, it was possible to obtain a lexicon of expressions that have the probability of being a verbal locution associated (see Section 4.3).

4.1 Identifying Contexts Associated with Verbal Locutions

In this section, describe the methodology for the automatic identification of candidate verbal locutions of Mexican Spanish. The approach is based on machine learning techniques and is proposed by [27]. In this methodology, verbal locutions are called fixed verbal expressions, however, they refer to the same type of linguistic structures addressed in this work. The methodology consists of the following steps:

1. To build a knowledge base of Fixed Verbal Expressions for Spanish (FVES).

2. To gather a set of documents written in Spanish in which FVES is expected to be found.

3. To build a large FVES tagging corpus using information retrieval techniques.

4. To build a classification model to identify FVES candidates using machine learning techniques.

5. To identify FVES candidates in unlabeled texts.

After applying the proposed methodology, with a corpus of 154,182 news [24], 9,118 contexts containing a candidate verbal locution have been obtained. That is, there are contexts that reflect the compositional and non-compositional sense of a verbal phrase. This result is the one that serves for the annotation and results in the following two lexical resources.

4.2 Noting the Non-Compositional and Compositional Meaning of Verbal Locutions

Once the automatic identification of contexts associated with verbal locutions was carried out, it is continued to write it down manually in two directions. First, determining whether there is really a verbal phrase or not in context.

Second, write down the semantic orientation in two classes (positive, negative) that determine the polarity of the context and the verbal phrase. The annotation was carried out by three human annotators expert in linguistics and scholars of the phraseological units, the evaluation of this annotation can be seen in Section 4.4. The corpus of contexts with verbal locutions, in which the compositional and non-compositional meaning of these phraseological units is presented is described in Tables 1 and 2.

Table 1 shows 7,533 contexts that contain a non-compositional meaning, that is, contexts in which there is really a verbal locution and its meaning is figurative. Table 2 shows 1,585

Table 1. Non-compositional contexts of verbal locutions

Feature	Total
Instances	7,533
Tokens	435,893
Vocabulary	30,093
Minimum length	3
Maximum length	1,280
Average length	57.86

Table 2. Compositional contexts of verbal locutions

Feature	Total
Instances	1,585
Tokens	85,893
Vocabulary	11,678
Minimum length	4
Maximum length	420
Average length	54.19

Table 3. DaVeL: verbal locutions lexicon

Feature	Total
Instances	127
Tokens	361
Vocabulary	281
Minimum length	2
Maximum length	6
Average length	4

contexts that contain a compositional meaning, that is, contexts in which there is a verbal phrase and its meaning is literal being only simple words and not a set of words that determine a phraseological unit. In this sense, one of the contributions of this work is this corpus of 9,118 total contexts.

4.3 DaVeL: Verbal Locution Lexicon

The construction of DaVeL is designed under an automatic identification of candidate verbal locutions in sentences taken from news [24]; subsequently, a manual labeling is carried out, which consists of reviewing said identification, establishing whether there is absence or presence of a verbal locution (VL), and finally evaluating whether the VL is positive or negative in the surrounding context. That is, the sentence that really has an VL is positive or negative, in order to polarize the use of verbal phrases in the journalistic domain.

For the identification of verbal locutions, 1,198 different phraseological units were used, of which 27% were identified in the news corpus. The most representative was selected, that is, those with the highest frequency in the corpus.

Additionally, they tried to balance the identified contexts, so that there were the same number of contexts for each verbal locution selected. Obtaining as a result a lexicon of 127 entries, Table 3 contains the characteristics of the lexical resource obtained.

Within the applications, of the created resource, there is [26] where different experiments have been carried out that allow identifying the polarity of contexts, of the journalistic genre, that contain candidate verbal locutions using different machine learning algorithms; the results show that the use of multi-word expressions benefit from the use of simple words. Another work is the one presented in [25] where, in addition to using a lexicon of phraseological units, different lexicons of simple words are used, allowing the union of the different types of words that are found (simple and compound).

DaVeL is a linguistic resource of great interest for the analysis of the polarity of texts; in addition, it helps to identify and evaluate whether an VL exists automatically and through machine learning processes.

4.4 Annotation Evaluation

Pearson's correlation coefficient, in statistics, is a linear measure between two quantitative random variables [12]. Less formally, Pearson's correlation coefficient can be defined as an index that can be used to measure the degree of relationship of two variables if both are quantitative and continuous. The correlation coefficient is given by equation 1, which refers to the average of the cross products of the standardized scores of X and Y; its value ranges, in absolute terms, between 0 and 1:

$$r_{xy} = \frac{\sum Z_x Z_y}{N}. \qquad (1)$$

For the type of annotation that was made, in this article, the correlation coefficient allowed to determine the degree of agreement between the scorers. The contexts compiled (9,118 news) manually annotated, by three human scorers, have a degree of agreement between scorers greater than 60% (60.8%) in the case of the contexts associated with verbal locutions; however, in the case of polarity the degree of agreement decreased by around 10%, being 51%. This is due to the fact that the classes (positive, negative) that determine the semantic orientation of the context are more difficult to determine and, as is known, is a slightly more complicated task.

5 Conclusions

The phraseology and study of the language have increased its interest in recent years, being important for various areas of natural language processing. The analysis and study of phraseological units, specifically verbal locutions, highlights the complexity and richness of the Mexican language. This article has focused mainly on the creation of lexical resources that allow addressing tasks within sentiment analysis. Specifically, identifying the semantic orientation of a text (polarity) from multi-word expressions and not simple words, as is normally reported in the literature.

The construction of lexical resources initially is designed under an automatic identification of candidate verbal locutions in sentences extracted from news. Subsequently, a manual labeling of each one of the contexts, of the journalistic genre; resulting in a corpus of 9,118 instances in Spanish, distributed in 7,534 non-compositional contexts and 1,585 compositional contexts. Additionally, DaVeL was created, a lexicon consisting of a collection of 127 verbal phrases in Spanish with probabilities of being or not being a verbal phrase in the journalistic domain.

The lexical resources, presented in this work, are of great interest for the analysis of the polarity of texts. In addition, it helps to identify and evaluate whether there is a verbal locution automatically and through machine learning processes.

References

1. **Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., de Rijke, M., & Spina, D. (2013).** Overview of replab 2013: Evaluating online reputation monitoring systems. **Forner, P., Müller, H., Paredes, R., Rosso, P., & Stein, B.**, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 333–352.
2. **Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019).** Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 54–63.
3. **Bobes Soler, E. (2016).** *Información gramatical asociada a las locuciones verbales del español*. Ph.D. thesis, Universidad de Barcelona, España.
4. **Brooke, J., Tofiloski, M., & Taboada, M.**, . Cross-linguistic sentiment analysis: From English to Spanish. *International Conference Recent Advances in Natural Language Processing, RANLP*.
5. **Casares, J. (1950).** *Introducción a la lexicología moderna*. C.S.I.C, Madrid.
6. **Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019).** SemEval-2019 task 3: EmoContext contextual emotion detection in text. *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 39–48.
7. **Corpas Pastor, G. (1996).** *Manual de fraseología española*. Gredos, Madrid.
8. **Fernández, J., Boldrini, E., Soriano, J. M. G., & Martínez-Barco, P. (2011).** Análisis de sentimientos y minería de opiniones: el corpus emotiblog. *Procesamiento del Lenguaje Natural*, Vol. 47, pp. 179–187.
9. **Fernández Anta, A., Núñez Chiroque, L., Morere, P., & Santos Méndez, A. (2013).** *Sentiment analysis and topic detection of Spanish tweets: a comparative study of NLP techniques*.
10. **García-Page Sánchez, M. (2008).** *Introducción a la fraseología española. Estudio de las locuciones.* Antrhopos, Barcelona.

11. **Hu, M. & Liu, B. (2004).** Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, ACM, New York, NY, USA, pp. 168–177.

12. **Jones, D. H. (1994).** Book review: Statistical methods, 8th edition George W. Snedecor and William G. Cochran Ames: Iowa State University Press, 1989. *Journal of Educational and Behavioral Statistics*, Vol. 19, No. 3, pp. 304–307.

13. **Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., & Potthast, M. (2019).** SemEval-2019 task 4: Hyperpartisan news detection. *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 829–839.

14. **Lamiroy, B., René Compas-Kuleuven, J., & Klein, J. (2005).** Le probleme central du figement est le semi-figement. *Linx*, Vol. 53.

15. **Liu, B. (2010).** Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*.

16. **Liu, B. (2012).** *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

17. **Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990).** Wordnet: An on-line lexical database. *International Journal of Lexicography*, Vol. 3, pp. 235–244.

18. **Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018).** SemEval-2018 task 1: Affect in tweets. *Proceedings of The 12th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 1–17.

19. **Nasukawa, T. & Yi, J. (2003).** Sentiment analysis: capturing favorability using natural language processing. **Gennari, J. H., Porter, B. W., & Gil, Y.**, editors, *K-CAP*, ACM, pp. 70–77.

20. **Ortega, G. & Gonzáles, A. (2005).** En torno a la variación de las unidades fraseológicas. *Fraseología contrastiva: con ejemplos del alemán, español, francés e italiano*, pp. 91–109.

21. **Pang, B., Lee, L., & Vaithyanathan, S. (2002).** Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 79–86.

22. **Pérez-Rosas, V., Banea, C., & Mihalcea, R. (2012).** Learning sentiment lexicons in spanish. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*.

23. **Priego Sánchez, B. (2015).** Análisis de la diversidad morfosintáctica en las locuciones verbales. *Research in Computing Science*, Vol. 97, pp. 113–125.

24. **Priego Sánchez, B. & Pinto, D. (2015).** Identification of verbal phraseological units in mexican news stories. *Computación y Sistemas*, Vol. 19, No. 4.

25. **Priego Sánchez, B. & Pinto, D. (2018).** Idiom polarity identification using contextual information. *Computación y Sistemas*, Vol. 22, No. 1.

26. **Priego Sánchez, B., Pinto, D., & Mejri, S. (2014).** Evaluating polarity for verbal phraseological units. *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.

27. **Priego Sánchez, B., Pinto, D., & Mejri, S. (2014).** Metodología para la identificación de secuencias verbales fijas. *Research in Computing Science*, Vol. 85, pp. 45–56.

28. **Rosenthal, S., Farra, N., & Nakov, P. (2017).** SemEval-2017 task 4: Sentiment analysis in twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Association for Computational Linguistics, Vancouver, Canada, pp. 502–518.

29. **Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J. (2013).** Empirical study of machine learning based approach for opinion mining in tweets. **Batyrshin, I. & González Mendoza, M.**, editors, *Advances in Artificial Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–14.

30. **Soler, N. & Batista-Rodríguez, J. (2008).** Unidades fraseológicas y variación. *Ogigia. Revista Electrónica de Estudios Hispánicos*, Vol. 3.

31. **Wotjak, G. (1985).** *Estudios de fraseología y fraseografía actual*. Iberoamericana, Frankfurt: Vervuert.

32. **Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019).** SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *Proceedings*

of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 75–86.

33. **Zuloaga, A. (1980).** *Introducción al estudio de las expresiones fijas*. Verlang Peter Lang, Frankfurt.

Article received on 29/10/2019; accepted on 07/03/2020.
Corresponding author is Belém Priego Sánchez.