Idiom Polarity Identification using Contextual Information

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

[1] Universidad Autónoma Metropolitana, Ciudad de México,
Mexico

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

abps@azc.uam.mx, dpinto@cs.buap.mx

Abstract. Identifying the polarity of a given text is a complex task that usually requires an analysis of the contextual information. This task becomes to be much more complex when, in such analysis, we consider smaller textual components than paragraphs, such as sentences, phraseological units or single words. In this paper, we consider the automatic identification of polarity for linguistic units known as idioms based on their contextual information. Idioms are a phraseological unit made up of more than two words in which one of those words plays the role of the predicate. We employ three lexicons for determining the polarity of those words surrounding the idiom, i.e., in its context and using this information we infer the possible polarity of the target idiom. The lexicons we are using are: *ElhPolar* dictionary, *iSOL* and *ML-SentiCON* Sentiment Spanish Lexicon, all of them containg the polarity of different words. One of the aims of this research work is to identify the lexicon that provides the best results for the task proposed, which is to count the number of positive and negative words in the idiom context, so that we can infer the polarity of the idiom itself. The experiments carried out show that the best combination obtain results close to 57.31%, when the texts are lemmatized and 48.87%, when they are not lemmatized.

Keywords. Polarity, idiom, lexicon.

1 Introduction

The rise of social media such as blogs and social networks has increased the interest in sentiment analysis. Polarity identification is a basic task of sentiment analysis which aims to automatically identify whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.

There are some research works in literature presenting different approaches for detecting the polarity of phrases or sentences in general. Turney [17] and Pang [10], for example, are two early works in this research area who applied different methods (at document level), for detecting the polarity of product reviews and movie reviews, respectively. Other authors such as Hu et al. [4] have attempted to identify polarity for adjectives using some linguistic resources such as Wordnet. This very fast approach does not need of training data necessary for obtaining a good predictive accuracy (around 69%), but the main disadvantage is that it does not deal with multiple word sense, context issues, and it does not work for multiple word phrases (or non-adjective words).

One of the major advances obtained in the task of sentiment analysis has been done in the framework of the Semantic Evaluation (SemEval) competition, since 2013 this task is proposed in SemEval. In SemEval-2013 Task 2, [7] and SemEval-2014 Task 9, [15] had an expression-level and a message-level polarity subtasks. SemEval-2015 Task 10, [14, 9] further added subtasks for topic-based message polarity classification, detecting trends towards a topic, and determining the out-of-context (a priori), strength of association of terms with positive sentiment.

SemEval-2016 Task 4, [8] dropped the phrase-level subtask and the strength of association subtask, and focused on sentiment with respect to a topic. It further introduced a 5-point scale,

ISSN 1405-5546
doi: 10.13053/CyS-22-1-2791

which is used for human review ratings on popular websites such as Amazon, TripAdvisor, Yelp, etc.; from a research perspective, this mean moving from classification to ordinal regression. Moreover, it also focused on quantification, i.e., determining what proportion of a set of tweets on a given topic is positive/negative about it. It also featured a 5-point scale ordinal quantification subtask [3].

Most of the aforementioned works have contributed to the target task of sentiment analysis by proposing methods, techniques for representing and classifying documents towards the automatic classification of sentiments in Tweets. This phenomenon is due to the massification of this social network around the world and the easy manner we can access to the Tweets from API's provided by Twitter itself. Some of these works have focused on the contribution of some particular features, such as Part of Speech (PoS) tags, emoticons, etc. on the aforementioned task. In [1], for example, the a priori likelihood of each PoS is calculated. They use up to 100 additional features that include emoticons and a dictionary of positive and negative words. They have reported a 60% of accuracy in the task. On the other hand, in [6], a strategy based on discursive relations, such as connectives and conditionals, with a low number of lexical resources is proposed. These relations are integrated in classical models of representation like bag of words with the aim of improving the accuracy values obtained in the process of classification. The influence of semantic operators such as modals and negations are analyzed, in particular, the degree in which they affect the emotion present in a given paragraph or sentence.

Sentiment analysis algorithms reported in literature mostly use simple terms to express sentiment about a product or service. However, cultural factors, linguistic nuances and differing contexts make it extremely difficult to turn a string of written text into a simple pro or con sentiment. The fact that humans often disagree with the sentiment of a given text, illustrates how difficult this task should be for computers to get it right. The shorter the string of text, the harder it becomes. In particular, identifying polarity for idioms is assumed to be a very high complex task.

In this paper we present the results of analyzing the polarity of idioms based on its contextual information. In [12], we can see a machine learning approach for dealing for the same task, however, the difference with such research work is that in this case we are employing unsupervised algorithms based on lexicons containing polarity of Spanish words, whereas in the mentioned paper, they employ annotated data for constructing a supervised model that takes into consideration the class/polarity of the paragraph (positive, negative or neutral).

This paper aims to identify the usefulness of each lexicon employed together with all the possible combinations of them, so that we can be able to determine the combination that better performs in the task of unsupervised automatic identification of idioms polarity.

In order to execute the experiment, it is assumed that the text to be analyzed contains at least one idiom. The process of automatic identification of idioms is out of the scope of this paper. In [11], we can find an interesting work presenting a methodology for such a challenging task.

We estimate the polarity of each idiom by counting the positive and negative words in their context using the following three lexicons and the combination of them: ElhPolar dictionary, ISOL y ML-SentiCON Sentiment Spanish Lexicon[1]. The greater the number of contextual positive words the idiom is assumed to be positive, the greater the number of contextual negative words, the greater the likelihood of the idiom to be negative. In case the number of positive and negative contextual words is similar, we assume that the idiom is neutral.

The following two research questions arise in this paper:

1. *What lexicon better fulfills the requirement for unsupervised automatic calculation of idiom polarity?*
2. *How easy is to classify the polarity of a given idiom in a particular domain or genre?*

[1] See Section 3.1.1 for a description of the lexicons employed.

The rest of this paper is structured as follows: Section 2, presents a description of the target problem. Section 3, describes the experiments carried out in this paper towards the automatic identification of the idiom polarity. Section 4 shows and discusses the results obtained in this paper. Finally, in Section 5, we give the conclusions and findings of this research work.

2 Problem Description

In natural language, there are many elements that are part of it and allow that communication exist; words and groups of them are examples of these parts. When we oral communication is maintained between two or more persons, intonation and gesture movements can help to determine some factors as feelings, for example, if the speaker is happy, angry, etc., in other words, if the person is talking in a positive or negative way. There is, however, a problem in written communication, because certain communication tones are difficult to be identified because we cannot see or hear the expression mode. This problem can sometimes be alleviated by analyzing the context of the main words in the text. This is one of the issues that complicate the automatic computational analysis of the sentiment identification task. Polarity identification is a subtask of sentiment analysis which aims to identify if a person has wrote a given text in a positive or negative way. In this case, it is not aimed to identify the particular emotion (happiness, sadness, etc.) but just the polarity of the emotion intended in the text.

Let us consider the phrase *ser pan comido* (a piece of cake), which in a compositional sense means *easy to be done or performed*. This phrase is difficult to be understood by computers, because the literal meaning differs significantly from the compositional one. Here the importance of developing computational methods for the automatic understanding of natural language. Such methods have a great variety of applications in real world including, automatic detection of bullying, depression, among others. Analyzing natural language requires a deep understanding of the different components of the linguistic structures, because texts can be analyzed through small parts such as sentences, which at the same time can be split out into smaller sentences such as words or phraseological units. The latter are the textual structures aim of the study of Phraseology, a subfield of linguistics.

Even if there are different studies associated with the analysis of Phraseological Units (PU), one of the major interest in this paper is to analyze the polarity of such units. There are some PU having positive charge, such as *estar en forma* (to be in a good shape), whereas other PUs have a negative charge, for example *meter la pata* (to screw up). Finally, other PUs are considered neutral since they cannot be defined in terms of positive or negative polarity. In [12], we have previously presented a method for the automatic identification of polarity in idioms, a particular type of phraseological units, which are defined in [13], as “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”. We have shown that different machine learning methods can be employed for such task with interesting results. The polarity of the idiom, in such paper, is identified by means of the contextual word polarity of the idiom.

Once verified the performance when machine learning techniques are employed, we are now interested in verifying the performance that can be achieved when knowledge based methods are employed, in particular, when different lexicons available in literature are used. In the next section we describe the methodology employed for the task of lexicon-based idiom polarity identification.

3 Methodological Framework

We have considered different lexicons freely available in literature which consider the manual annotation of word polarities. Having such lexicons we can infer the polarity of a given idiom present in a news histories by means of the contextual polarity. In short, we can calculate the idiom polarity as shown in Eq.(1):

$$Polarity(Idiom) = \left\{ \begin{array}{l} \text{Positive iff } |w_i^+| > |w_j^-| \\ \text{Negative iff } |w_i^+| < |w_j^-| \\ \text{Neutral iff } |w_i^+ - w_j^-| < \epsilon \end{array} \right\}. \quad (1)$$

Let V^+ be the types with positive charge and V^- the types with negative charge. If w_i is a token belonging to the idiom context, then we can define w_i^+ as the i-th contextual word of the idiom with positive charge, i.e., $w_i^+ \in V^+$. We can also define $w_j^- \in V^-$ as the j-th contextual word of the idiom with negative charge. If $|w_i^+| > |w_j^-|$ then we can say that the linguistic phrase has a positive polarity, thus, assuming that the core of the phrase, the idiom, has also a positive polarity. In case, $|w_j^-| > |w_i^+|$ the outcome is that the idiom has a negative charge. Finally, when the difference is very small, there exist a balance in the both polarities, assuming that the idiom charge would be neutral.

A description of the lexical resources employed in the experiments follows.

3.1 Lexical Resources

The high interest in studying natural languages has motivated the construction of lexical resources available in different languages. In this paper, we have used such resources for the solution of problems associated with the identification of idiom polarity. Three lexicons have been selected from literature for such purpose, each one containing a polarity charge (positive, negative) for different Spanish words. The dataset consist of a news corpus containing phrases with idioms, each one manually annotated with a particular polarity charge.

3.1.1 Lexicons

The selection of the following lexicons was based on the availability in the literature and their use in other research works. We would like to emphasize that resources of this kind are difficult to be find in Spanish, since they are usually being constructed for English language.

a) *ElhPolar* [16]. The ElhPolar polarity lexicon for Spanish was created from different sources, and includes both negative and positive words. The first source was an English polarity lexicon [18] which was automatically translated to Spanish language by using an English-Spanish bilingual dictionary. Ambiguous translations were solved manually by two annotators. Polarity was also checked and corrected during this manual annotation. The second source were words automatically extracted from a particular training corpus. And the third source was a list of colloquial polarity vocabulary. This lexicon contains 1,897 positive words and 3,302 negative words, and was created for the task of sentiment analysis in the TASS 2013 evaluation campaign[2].

b) *iSOL* [5]. iSOL is a list of Spanish words indicating domain independent polarity. The types come from a word list of the Bing Liu's Opinion Lexicon[3]. In this case, the words has been automatically translated to Spanish using the Reverso translator, with a posterior manual correction process. The lexicon contains 2,509 positive words and 5,626 negative words.

c) *ML-SentiCON* [2]. Multilingual, layered sentiment lexicons at lemma level. This resource contains lemma-level sentiment lexicons for English, Spanish, Catalan, Basque and Galician. In this paper, we have used only the Spanish lexicon. For each lemma, it provides an estimation of polarity (from very negative -1.0 to very positive +1.0), and a standard deviation (related with ambiguity of the polarity estimation). This lexicon has 5,568 positive words and 5,974 negative words.

3.1.2 News Histories Dataset

For the experiments carried out in this paper, we have used 2,263 news histories containing each at least one occurrence of an idiom. We have considered 112 different idioms with three possible polarity charge: positive, negative or neutral. In this dataset we have 1,492 positive idioms, 578 negative idioms and 193 neutral idioms.

[2] http://www.sepln.org/workshops/tass/2013/about.php

[3] https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

4 Experimental Results

The use of different lexicons for determining the polarity of a given idiom is the aim of this paper. However, it is very important to determine which one contributes better to the target task. In order to investigate that research question, we have constructed different scenarios of execution considering the different combinations of the lexicons.

In Table 1, we present the results of the experiment when the terminology has not been lemmatized. The best result is obtained when the *ML-SentiCON* lexicon is used. This is an expected outcome because the number of words in this lexicon is significantly greater than the ones in the other two lexicons. The best result is obtained when the two lexicons with greater number of types are combined, in this case, *iSOL* and *ML-SentiCON*. In summary, the number of types in the lexicons played an important role in the performance obtained.

Table 1. Results obtained when the terminology has not been lemmatized

Lexicon	Accuracy
iSOL	39.68
ElhPolar	40.30
ML-SentiCON	46.57
ElhPolar ∪ iSOL	43.65
ElhPolar ∪ ML-SentiCON	47.95
iSOL ∪ ML-SentiCON	48.87
ElhPolar ∪ iSOL ∪ ML-Senticon	48.38

In Table 2, we can observe the performance of the task when all different combinations of lexicons are also employed. The *ElhPolar* lexicon outperforms the other two ones when each lexicon is used alone. This result differs from the previous ones, not only the accuracy is significantly greater but the lexicon obtaining the best score is different. The lemmatization process clearly helps to improve the obtained results. In the same Table we can see that the best combination is precisely the union of types of *ElhPolar* and *ML-SentiCON* with an accuracy of 57.31%. Actually, when the *iSOL* lexicon is added to this combination, the performance decreases. This behavior indicate that the *iSOL* lexicon needs to be improved before to be employed in the task of sentiment analysis. At least with the purpose of identifying the polarity of idioms in news histories.

Table 2. Results obtained when the terminology has been lemmatized

Lexicon	Accuracy
iSOL	39.68
ML-SentiCON	49.27
ElhPolar	53.77
ElhPolar ∪ iSOL	53.33
iSOL ∪ ML-SentiCON	49.00
ElhPolar ∪ MLSentiCON	57.31
ElhPolar ∪ iSOL ∪ ML-Senticon	54.75

In Figure 1, we can see a comparison of all the results obtained in the experiment. It clearly can be observed that in this case the use of the *iSOL* lexicon lessen the accuracy of the intended task.

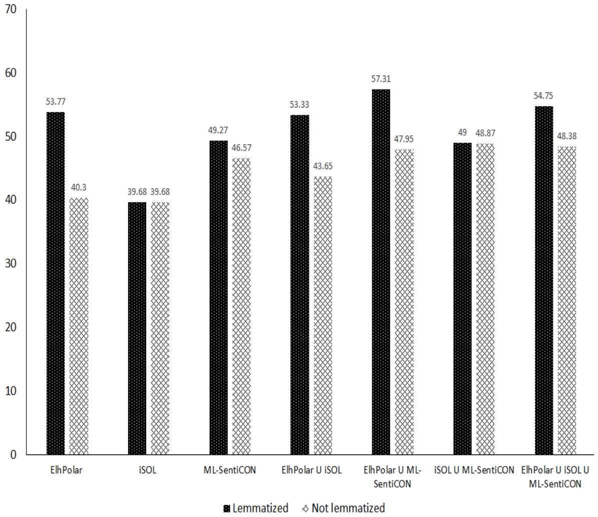

Fig. 1. Summary of the results obtained

5 Conclusions

In this paper we have analyzed three different lexicons for their usefulness in the particular task of identification of idioms polarity. These lexicons has been widely employed in literature with different tasks associated with sentiment analysis for the Spanish language. The obtained results of the experiments carried out in this paper show that two of them are competitive in the aforementioned task, whereas one lexicon obtained results lower than expected, since, as far as we know, the *iSOL* lexicon has presumably been constructed specifically for the Spanish language, whereas the other two are partially coming from a translation process from English to Spanish.

In any case, we have observed that the combination of two lexicons, *ElhPolar* and *ML-SentiCON*, in the task of idiom polarity detection obtained a performance of 57.31%, which is considered to be a good result because we are just employing a knowledge based methodology, in particular, using lexicons with terminology manually annotated with polarity.

In summary, with respect to the two research questions presented at the beginning of this paper, we can conclude that the *ML-SentiCON* lexicon is the one that better fulfill the requirement for unsupervised automatic calculation of idiom polarity. The second research question is not easy to be answered because the results are considered to be good enough for a simple technique based on lexicons, but they are not competitive yet with other methods, for example such ones that are based on machine learning which usually obtain performances significantly greater, but requiring high amount of annotated datasets.

References

1. **Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011).** Sentiment analysis of twitter data. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, Portland, Oregon, pp. 30–38.
2. **Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014).** Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Syst. Appl.*, Vol. 41, No. 13, pp. 5984–5994.
3. **Gao, W. & Sebastiani, F. (2016).** From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining*, Vol. 6, No. 1, pp. 19.
4. **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.
5. **Martínez-Cámara, E., Martín-Valdivia, M. T., Molina-González, M. D., & López, L. A. U. (2013).** Bilingual experiments on an opinion comparable corpus. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2013, 14 June 2013, Atlanta, Georgia, USA*, pp. 87–93.
6. **Mukherjee, S. & Bhattacharyya, P. (2012).** Sentiment analysis in Twitter with lightweight discourse analysis. *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 1847–1864.
7. **Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013).** *SemEval-2013 Task 2: Sentiment Analysis in Twitter*.
8. **Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016).** Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pp. 1–18.
9. **Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Stoyanov, V., & Zhu, X. (2016).** Developing a successful semeval task in sentiment analysis of twitter and other social media texts. *Language Resources and Evaluation*, Vol. 50, No. 1, pp. 35–65.
10. **Pang, B., Lee, L., & Vaithyanathan, S. (2002).** Thumbs up? sentiment classification using machine learning techniques. *Proceedings of EMNLP*, pp. 79–86.
11. **Priego-Sánchez, B. & Pinto, D. (2015).** Identification of verbal phraseological units in Mexican news stories. *Computación y Sistemas*, Vol. 19, No. 4.
12. **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.
13. **Priego-Sánchez, B., Pinto, D., & Mejri, S. (2015).** Towards the automatic identification of spanish

verbal phraseological units. *Research in Computing Science*, Vol. 96, pp. 65–73.

14. **Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015).** Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, Denver, Colorado, pp. 451–463.

15. **Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014).** Semeval-2014 task 9: Sentiment analysis in twitter. **Nakov, P. & Zesch, T.**, editors, *SemEval@COLING*, The Association for Computer Linguistics, pp. 73–80.

16. **Saralegi-Urizar, X. & San-Vicente-Roncal, I. (2013).** Elhuyar at tass 2013. *XXIX Congreso de la Sociedad Espańola de Procesamiento de lenguaje natural. Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pp. 143–150.

17. **Turney, P. D. (2002).** Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 417–424.

18. **Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005).** Opinionfinder: A system for subjectivity analysis. *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 34–35.

Article received on 01/08/2016; accepted on 12/10/2016.
Corresponding author is Belém Priego Sánchez.