The Role of WordNet Similarity in the Affective Analysis Pipeline

Alejandra Segura Navarrete¹, Christian Vidal-Castro¹, Clemente Rubio-Manzano¹, Claudia Martínez-Araneda²

¹ Universidad del Bío-Bío, Facultad de Ciencias Empresariales, Concepción, Chile
² Universidad Católica de la Santísima Concepción, Facultad de Ingeniería, Concepción, Chile

{asegura,clrubio,cvidal}@ubiobio.cl, cmartinez@ucsc.cl

Abstract. Sentiment Analysis (SA) is a useful and important discipline in Computer Science, as it allows having a knowledge base about the opinions of people regarding a topic. This knowledge is used to improve decision-making processes. One approach to achieve this is based on the use of lexical knowledge structures. In particular, our aim is to enrich an affective lexicon by the analysis of the similarity relationship between words. The hypothesis of this work states that the similarities of the words belonging to an affective category, with respect to any other word, behave in a homogeneous way within each affective category. The experimental results show that words of a same affective category have a homogeneous similarity with an antonym, and that the similarities of these words with any of their antonyms have a low variability. The novelty of this paper is that it builds the bases of a mechanism that allows incorporating the intensity in an affective lexicon automatically.

Keywords. Natural language processing, computational linguistics, affective computing, sentiment analysis, knowledge representation.

1 Introduction

Nowadays, Sentiment Analysis is a useful and important discipline in Computer Science, which allows obtaining potentially valuable knowledge about user’s perceptions, expectations and attitudes in order to improve the decision-making process regarding products and marketing strategies, among other uses. Sentiment Analysis has not only been applied in business but also in very different areas such as recommender systems [1, 2], electoral analysis [3] management of virtual museums [4], multilingual processing [5, 6] among others.

In general terms, there are two approaches to perform this type of analysis according to [7]. The first one uses a corpus of tagged texts that allows the construction of a classifier trained to execute this task. This approach uses supervised learning techniques that come from machine learning and statistics [8]. The second approach uses lexical resources, such as dictionaries or lexicons which are defined as a previously tagged set of words [9], i.e., every word is tagged according to its orientation [10]. There are two main lines of work: The identification of both positive and negative opinions, emotions and evaluations, using computing tools to assign a polarity to the content [11].

The estimation of the affective aspect of a text [9] called Affective Analysis, where there is a lexicon containing a set of words classified according to the emotions they represent [12, 13]. The emotion expressed in a sentence or text is obtained considering the emotion of all the words contained in that text [14].

The classification process in Sentiment Analysis is simpler than in Affective Analysis.
The first one classifies in two or three categories, either (positive, negative) or (positive, negative or neutral) and the second one can do so in many, depending on the model of emotions used. In addition, a text can express more than one emotion and even two different texts can express the same emotion but with different intensities.

In Affective Analysis based on lexicon approach, the results depend on the quality and completeness of the lexicon used in the process. The affective lexicons include the words grouped by affective category. This fact only allows a words-bag analysis since the words of each affective category do not contain affective intensity information; therefore, it is not possible determining affective profiling of a document. In Sentiment Analysis, there are works having improved the quality of affective lexicons by adding information such as valence, arousal and dominance [15–17]. In the case of affective analysis based on lexicon, studies are mainly aimed at increasing the number of words of an affective lexicon [18].

The objective of this paper is to enrich a lexicon of affects by the analysis of the similarity relationship between words. The hypothesis of this work states that the similarities of the words belonging to an affective category, with respect to any other word, behave in a homogeneous way within each affective category. We found evidences that similarities of the words belonging to an affective category, with respect to any other word, behave in a homogeneous way within each affective class. This finding will allow us to determine intensities for the emotions of an affective category and to improve automatic enrichment process of affective lexicons.

The rest of the article is structured as follows: Chapter 2 presents a background and a brief state of the art about the use of lexicons in affective analysis and similarity measures between two words. Chapter 3 presents the hypothesis and experiments performed to prove it. Chapter 4 presents the results and discusses the word’s similarity behavior of the lexicon’s affective classes. Finally, conclusions and future work lines are presented in Chapter 5.

2 Background and Related Work

In lexicons commonly used in Affective Analysis, consider different classifications of basic emotions, assuming that all other emotions would depend on these subsets.

For example, author proposed 6 categories in [19]: anger, disgust, fear, joy, sadness and surprise. In [20] was proposed an affective lexicon called WordNet-Affect, which was built based on the WordNet knowledge base, through the selection and tagging of affective concepts. This initial base was extended using sentences and patterns extracted from Open Mind Commonsense [21]. WordNet-Affect classifies words into the six categories of Ekman.

Each word in the lexicon contains lexical and affective information, for example, the role of the word in speech (part-of-speech), classification according to emotion theory or representation, among others. Another affective lexicon was proposed in [22], which considered 8 affective categories: anger, anticipation, disgust, fear, joy, sadness, surprise and trust. This lexicon was generated from a list of affective words extracted from the Thesaurus WordNet-Affect and the most frequent words in Google n-gram Corpus [23].

In the case of Affective Analysis based on lexicon, the studies have as main objective to increase the number of words of an affective lexicon. For example, the authors in [18] present an approach for the Japanese language where a similarity metric was used to expand a small group of emotionally-charged words (containing 503 nouns) into an emotions dictionary (containing 15612 verbs). Other studies have improved the lexical resources through the integration [24] or the creation of lexicons for specific domains [25, 26].

One aspect to consider is that a lexicon can be used in different domains. This may imply that one word may represent different affect in different domains [27]. On the other hand, in order to incorporate the concept of semantic similarity among words in Affective Analysis, it is necessary to analyze both, the metrics based on the structure and the metrics based on the Information Content (IC).

The semantics’ similarity measures based on structure add variables such as lowest common