The Spanish Travel Subjective Lexicon (STSL)

Liliana Ibeth Barbosa Santillán
Information Technology
University of Guadalajara, México
ibarbosa@cucea.udg.mx

Inmaculada Alvarez de Mon y Rego
Mercedes Rodríguez Villareal
Lingüistica aplicada a la ciencia y la tecnología
Universidad Politécnica de Madrid
ialvarez@euitt.upm.es

Abstract

This paper presents a proposal for a recognition model for the appraisal value of sentences. It is based on splitting the text into independent sentences (full stops) and then analysing the appraisal elements contained in each sentence according to the previous value in the appraisal lexicon. In this lexicon, positive words are assigned a positive coefficient (+1) and negative words a negative coefficient (-1). We take into account word such as "too", "little" (when it is not "a bit"), "less", and "nothing" than can modify the polarity degree of lexical unit when appear in the nearby environment. If any of these elements are present, then the previous coefficient will be multiplied by (-1), that is, they will change their sign. Our results show a nearly theoretical effectiveness of 90%, despite not achieving the recognition (or misrecognition) of implicit elements. These elements represent approximately 4% of the total of sentences analysed for appraisal and include the errors in the recognition of coordinated sentences. On the one hand, we found that 3.6 % of the sentences could not be recognized because they use different connectors than those included in the model; on the other hand, we found that in 8.6% of the sentences despite using some of the described connectors could not be applied the rules we have developed. The percentage relative to the whole group of appraisal sentences in the corpus was approximately of 5%.

1 Introduction

The lack of studies relative to polarity analysis in the Spanish language is one of the motivations for this research because it is the fourth largest language spoken in the world with 6.9 million speakers. The automatic analysis of the appraisal in text corpora has significantly advanced in recent years. There are proposals of language codification that have been developed based on relatively reliable tools. Unfortunately, most developments are carried out only in English. Translation is not always a good solution and that is specially evident in the case of translating polarity. For example, it would be necessary to face a translation of every polarity word in English with the uncertainty that the cultural weight of each language does not suppose a modification of the polarity expression in the same way. Likewise, it would not be possible to translate some elements such as phrases, sayings, and popular expressions in a particular culture. (Gelbukh et al., 2002) said that analyzing a traditional corpus of texts has the disadvantage that few words occur many times thus using most of the processing time. In addition, the majority of interest words when appraisal or polarity are concerned have few or no representation in some of the existing corpora. This limits the study of word contexts. One solution to this problem is the use of the web which can be regarded as a large virtual corpus. It has sufficient information in order to study the properties of a large number of words. But this virtual corpus has some disadvantages compared to the traditional ones such as: the network response time, unstable results in time, etc. The corpus that has been examined in this research combines the advantages of both types, virtual and locally stored. Since it is created from the web where we find all kinds of material that can be stored in order to analyze the properties of a particular type of texts, in this case travel blogs. It is possible to gain access to stories narrated by many people simply by choosing the topic. For this research we chose to extract texts from travel blogs since they are an inexhaustible source of good material for the study of the language of appraisal. As by their nature they are personal experiences, these stories have a very strong subjectivity. But they also have the component that the authors want not only to convey their travel experiences, but also evaluate the places visited. This includes the things they have seen or the people they have met in order to advise future travellers about what to do or not do, what to see and not to see, where to eat or not to eat, where to sleep or not to sleep, etc. We see it as an ideal source to extract polarity because almost everything is based on opinion. These stories are very different compared to a travel guide where the narrative is much more objective, full of facts and with very few evaluations (González Rodríguez, 2011). The aim of this project is double; on the one hand, the creation of a corpus of travel blogs that can be by itself a useful tool for a linguistics corpus and on the other hand, a study of the value patterns in the sentences, using the corpus of text previously created for this purpose. In this way, we can propose a computational model for

Liliana Santillán Barbosa Ibeth,Inmaculada Álvarez de Mon Y Rego

to estimate the polarity of the sentence. The proposed model is based on the following phases: (1) To recognize the polarity elements (words or groups of words) in a sentence within the corpus, (2) To recognize certain grammatical structures also indexed in the lexicon, (3) To establish rules that switch the polarity elements or polarity structures, (4) To obtain a semantic sum of the polarity elements. The remainder of the paper is structured as follows: Section 2 describes the Spanish Travel Subjective Lexicon (STSL) approach. Section 4 presents details of our data sets, evaluation metrics and the result. Finally, Section 5 presents our conclusions and future research.

2 Background

We understand appraisal theory as the discursive construction of attitude and intersubjective posture (Pérez Nieto and Redondo Delgado, 1997). This approach is a term of wide scope, that includes all evaluative uses of the language through which speakers and writers not only bring particular value attitudes but also negotiate these positions with their actual and potential interlocutors (Kaplan, 2004).

Following the work of (Kaplan, 2004) and using her examples and terminology, we will discuss the way in which appraisal theory is divided into three semantic domains: attitude, compromise, and gradation.

2.1 Attitude

The attitude domain includes the meanings by which the texts or speakers attribute a value or an intersubjective appraisal to the participants and processes. These can be related to both emotional answers and with value systems that are culturally determined. All statements are classified as attitudinal if they convey an evaluation both positive and negative, or can be interpreted as an invitation to the reader to provide their own negative or positive evaluations

This category is divided into affect, judgment and appreciation subsystems.

Affect Affection is the evaluation of how the writer indicates his or her emotional disposition towards persons, things, situations or events. The emotions are concentrated in three major groups that deal with happiness or unhappiness; safety and insecurity; and satisfaction or dissatisfaction. The linguistic indicators of affection can be verbs of emotion that refer to mental processes (eg.: to love / hate); adverbs that indicate circumstances of mood (e.g.: happily and sadly); adjectives that express emotion (e.g. : happy/sad), and nominalizations, i.e. transformations of verbs and adjectives into nouns (e.g.: happiness / desperation).

Judgment Judgment can be understood as the institutionalization of emotions in the context of rules on how people should and should not behave. The social norms that act into these appraisal judgements take the form of regulations or social expectations.

Judgments of social esteem are subdivided into: (a) relative to normality, (b) the capacity or the determination demonstrated in the conduct; all are evaluated in order to know how normal is a person, how competent or how decisive and determined he or she is, and (c) judgments of social sanction related to the veracity and moral integrity.

Appreciation

Appreciation can be considered as the system where human assessment is expressed toward products, processes and entities that are valued positively or negatively. Artefacts, texts, abstract ideas, plans and policies, and objects are evaluated according to polarity. Individuals can also be evaluated through appreciation, but only when they are perceived as entities and not as humans.

3 The approach

The STSL approach has seven stages: (1) search for texts, (2) text selection, (3) building of textual documents, (4) tagging of travel blogs on the web, (5) classification of data, (6) indexation, and (7) analysis of appraisal patterns. Then a database is created in order to save or delete lexical elements. As a result, we have a corpus of text, appraisal sentences, and an appraisal lexicon that will be later used for the analysis of appraisal patterns.

Search for Texts Our text is a travel forum, with a section for blogs. There are many entries, although not all meet the characteristics of inclusion in our corpus, either due to incomplete or general ideas or any other type of feature. STSL allows access to blogs and the possibility to search by title, user or by continent. Next we found that the blog is structured in two parts: (a) the most recently published diaries, and (b) most popular daily entries depending on the number of visits. The method of search was random, since texts did not fulfill all the selection features of the blogs. It was necessary to perform a tracing process, examining many blogs and performing a preliminary inspection of all of them one by one. Then STSL selected and discarded irrelevant blogs.

Text Selection The quality of the corpus is measured by the degree of compliance of the documents that meet the purpose for which the corpus is compiled. Thus it was necessary to take special care in the selection of documents attempting to maintain homogeneity. Therefore, it was necessary to establish the following criteria that govern the selection and inclusion of documents. (a) Quantity: It was decided to include 24 blogs of different dimensions. The total number of words collected was 201.678. (b) Quality of text: Given that the selection was manual, special care was taken in that the texts were written in the correct language, without spelling mistakes, in clear writing. (c) Published in travel blogs: Due to the nature of the project, we only included published blogs, discarding blogs in restricted

personal pages. (d) Type of travel: the trip must have been carried out as a tour, following a route or path, i.e. through a single country or region as a whole (for example we recognize Scandinavia as a region although it includes three countries, or the westerm United States given its extensive area). (e) Text form: The texts must be written in the form of logs or diaries, discarding the texts of general impressions of a journey for its lack of detail. (f) Style: The texts must be comprehensive, describing the journey from beginning to end, discarding free or incomplete texts introduced in unfinished or abandoned blogs. (g) Additional information: Each sample must be marked with a series of additional data, which gives extra information and allows for identification. These marks are the: web page from which it has been extracted, country or area where the trip has been realized, language, date of the trip, date of creation of the blog, and name of the author or nickname.

Building of Textual Documents Once the 24 blogs were selected, they were copied and included in documents for accommodation in the database. We found additional difficulty with some personal blogs where each day of the trip was on different web pages and we needed to browse all the links by using an index.

Tagging of Travel Blogs on the Web Sentences with appraisal value we marked with colours, blue for positive and pink for negative.

Clasification of data Each sentence was extracted from the word or group of words that function as an appraisal element, classifying it into existing categories using the drop-down categories. Grammatical categories were defined in advance but are subject to changes.The classification included the concept of "gradation" where our approach allows to select a word and those associated with it that can modify its value either intensificating or weaken.

Indexation To accommodate the corpus and have the versatility and functionality required, it was decided to create a relational database containing: (a) The texts of each of the blogs. (b) All appraisal sentences classified as positive or negative. (c) All the appraisal elements of each sentence and its grammatical classification. (d) The established object, person or situation from the emitted evaluations.

Analysis of Appraisal Patterns First we studied the lexical items that indicate the ability to infer rules that permit automated recognition of the evaluation. This allowed us to recognize polarity, i.e. positive or negative elements. We can divide the lexical elements of the corpus in two large groups. On the one hand, words or groups of words that have a fixed structure and can be easily recognized. In this group are adjectives, verbs, adverbs, nouns and phrases. We call these elements "explicit appraisal elements". On the other hand are the elements, they are included in sentences with a clear polarity but that are subjective or complex to recognize, because sentence structure is variable or the nature of the assessment is not evident in any of its elements. We call these elements "implicit appraisal elements" are items that we cannot identify due to its semantic complexity . Irony is that property of speech by which speakers understand the opposite of what is said; this only is identified by using context, which makes it impossible to know the elements that indicate irony. We study the explicit appraisal elements based on their frequency of occurrence in sentences. We see that the greater weight of the appraisal is in adjectives, which is logical since the adjective is an element specifically designed to assess. Next are the nouns, verbs and adverbs.

4 Experiment and Results

Our approach evaluated 24 blogs consisting of 345 pages that contain a total of 201,678 words. We extracted from them 4,183 sentences and 6,295 lexical elements as shown in Fig. 1.

Place	*Words*	*Appraisal*	*Lexical*
Scotland	2488	33	43
Senegal	5296	109	129
Italy	9013	198	312
Poland	3105	59	76
Morroco	8414	165	256
Italy	3058	49	58
Japan	14747	290	427
Vietnam	10743	245	364
Cuba	9550	188	273
USA-Canada	15110	391	604
USA	5030	109	180
China	11251	194	290
Austria-Germany	12798	255	364
India	8827	201	332
Lapland	10976	234	351
Morroco	5802	147	274
Egypt-Lebanon	19512	336	506
Syria	5670	154	240
Japan	6452	92	142
Iceland	14152	301	443
Scotland	6169	125	167
Morroco	4179	70	104
Thailand	3597	128	213
Egypt	5745	110	138
TOTAL	201678	4183	6295

Figure 1: Number of appraisal and lexical elements by country

The explicit appraisal elements are shown in Table I, they constitute a total of 95% of all items that contain polarity, in contrast to 5% implicit appraisal elements.

explicit appraisal elements	implicit appraisal elements
Adjectives	Rhetoric figures
Nouns	Ironies
Verbs	Exclamations
Adverbs	Theoretical interrogation
Phrases	Quotes
Suffixes	Suspension points
Interjecciones	Change of record

Table 1: The explicit and implicit appraisal elements.

The adjectives represent 48% of the total explicit appraisal elements. Some examples of adjectives in superlative degree which appear in the corpus are adjectives that already appear in the list of appraisal adjectives and include the prefix "super". Considering gra-

dation when it accompanies explicit adjectives it does not change polarity since it only reinforces the positive or negative value of the specific adjective. The most common graders are: ”Más, mas o menos, mucho, mucho más, muy o bien, tan, bastante, algo”, etc. The nouns represent 48% of the total explicit appraisal elements. The verbs represent 14% of the total explicit appraisal elements. The adverbs represent 7% of the total explicit appraisal elements. The most used adverbs or greater number of repetitions are shown in Table II.

adverb	frequency
bine	175
mal	59
perfectamente	13
tranquilamente	24
others	206

Table 2: The most used adverbs

They represent 4% of the total and include the identify margin of error. In the same way, we found in the corpus, evaluative nouns that are created by adding a suffix to the corresponding adjectives like ”-idad” or ”-ez”. In this way, the appraisal noun ”majestuosidad” would be the evaluative adjective ”majestuoso” or of the appraisal noun ”rapidez” would be the adjective ”rápido” by adding the corresponding suffixes. It would be possible to extend the list of nouns, by utilizing the appraisal adjectives, with their corresponding endings of nouns, identifying all the possible endings. However, given the variety of sufixes and not all adjectives can be transformed into nouns. For this study, we have extended the list of nouns only for those adjectives that appear in the corpus sample. The appraisal of the entire corpus by country with its polarity is shown in Fig. 2.

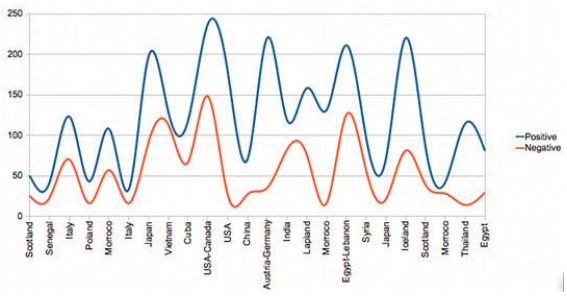

Figure 2: The appraisal by country with its polarity

In this study, the blog writers value the following the most:

$$IV :\doteq \begin{cases} amplitud, apariencia, aprovechamiento, \\ cantidad, comodidad, distancia, \\ eficacia, historia, importancia, \\ limpieza, mobiliario, olor, organización \\ precio, proximidad, puntualidad, sabor, \\ situacion, sonido, tamano, tolerancia. \\ valor\ añadido, velocidad, visitas \end{cases} \quad (1)$$

5 Conclusions

Appraisal theory can be considered as related to sentiment analysis. The application of this type of analysis has many possibilities, but all of them are based on sentence polarity recognition. Our model also accounts for the possibility of modifying the original value by means of a negative word. In this case, the possible negations must be located within the context of the sentence: ”no,” ”nor,” ”neither,” ”without,” ”none,” and ”never,” as well as any double negations that are dealt with as one: ”no/nor neither,” ”no/nor none,” ”no/nor never,” ”without nothing,” and ”without any.” If there are negations then the coefficient will change its sign by being multiplied by menus one (1). Then, the links with coordinated sentences should be sought: ”but,” ”despite” and ”although.” If there are more than one of these, then the following rules are applied: positive (P) ”but” negative (N) = negative; N ”but” P = P; P ”despite” N = P; N ”despite” P = N; P ”although” N = P; N ”although” P = N; ”although” P, N = N; ”although” N, P = P; ”although” P, N = N; ”although” N, P = P. It is possible to extend the number of appraisal words in our lexicon through the transformation of lexical categories into others by applying the relevant suffixes. We have observed that in most cases it is the root that has the appraisal meaning; thus, we can conclude that if ”glad” is an appraisal adjective then ”gladder” is an appraisal adjective too. Extension is also possible using techniques that were based on expanding lists of basic words to full lexical terms with the recursive consultation of synonymous words using electronic dictionaries.

Acknowledgments

We are grateful to the Sciences Research Council (CONACYT) and Multilingüismo en ontologías y linked data (BabelData), TIN2010-17550, funded by the Ministry of Science and Technology, 2011-2013.

References

Gelbukh, A., Sidorov, G., and Chanona-Hernández, L. (2002). Computational linguistics and intelligent text processing. 2276:285–288.

González Rodríguez, M. J. (2011). La expresión lingüística de la actitud en el género de opinión: el modelo de la valoración. *Revista de lingüística teórica y aplicada*, 49:109 – 141.

Kaplan, N. (2004). Nuevos desarrollos en el estudio de la evaluación en el lenguaje: La teoría de la valoración. *Revista Boletin de Lingüística*, 22:52–78.

Pérez Nieto, M. A. and Redondo Delgado, M. M. (1997). Procesos de valoración y emoción: Características, desarrollo, clasificación y estado actual. *Revista Electrónica de Motivación y Emoción (R.E.M.E.)*, IX(22).