Fixing and evaluating texts

Mixed text reconstruction method for data fusion environments

Antonio Juan Sánchez Martín
Department of Computer Science
University of Salamanca
Salamanca, Spain
anto@usal.es

Fernando de la Prieta Pintado
Department of Computer Science
University of Salamanca
Salamanca, Spain
fer@usal.es

Giovanni De Gasperis
Dip.di Ing. Elettrica e dell'Informazione
Università degli Studi dell'Aquila
L'Aquila. Italy
giovanni.degasperis@univaq.it

***Abstract*—This paper presents a method for the reconstruction of malformed texts into Spanish, these texts can be found regularly in social media environments. Also a method of sentiment analysis based on analysis of pairs will be presented. This sentiment analysis method includes intensifiers and applies the same procedures for analyse words, phrases and paragraphs. The method is centred on Spanish but the solutions proposed can be extended to any language.**

Keywords— text analysis, language recognition, fixing text algorithm, sentiment analysis, opinion mining

I. Introduction

In environments such as Twitter, Facebook, forums, news, or comments of web pages, the use of the lexicon and grammar is not quite correct, or rather it can be said that is not standard. Deformations of words and abbreviations that are not include in the normal use of language are produced continuously. These deformations can even mutate in a very specific way depending on the area of the country, according to the various peculiarities of the regional variations of the language.

This phenomenon is not closed to only one language, as expected, it occurs regularly in all the major languages used in social environments.

The study of Almansa, Fonseca and Castillo [1], warm of variability in the use of language both Spain and South America, and the study of González Alba [2] about lexical variations in Tuenti [1] environment describes how those variations change depending on the age (according Arroyo [3]) and localization [also in 1] . The process of evolution of language is a purely biological cycle, guided by natural selection, where the chances of success and acceptance determine the survival of the new terms within the community. It makes this model would be difficult to predict and the continuous adaptation seems to be the only alternative.

Throughout this article we will see how to identify the change of language and how "reconstruct" the texts with identifiable human symbols.

On the other hand one of the fundamental problems in analyzing texts where can be found mutated words, is the polysemy of terms, such as the term "bn", in common use, which can be translated as "good night", "good" or "well" (in spanish "buenas noches", "bueno" o "bien"). This provokes several important effects into the sentiment analysis ("good" would have a different value within a "good" analysis). In this article we will also see solutions to this conflict.

In the field of Information Fusion, these techniques of language standardization are very important when combining and evaluating texts from different sources (not just from social networks). Then, after this pretreatment of the infromación, is possible to apply the same techniques of analysis for all generated texts.

II. Variability studies in language into social media networks

The study of Alba González [2] gives us one piece of evidence about how great are the variations on the standard language can be found on social networks. The following list reflect which are the general cases of deformation, modification and new creation of words in this environment (most of them are collected in the Montiel Escandell study [5]):

- Loanword not recognized by the RAE[2], as "family", "party", "kiss", "statute", "belo", "muito". Most often, these casts come from English or Latin languages such as Italian, Portuguese or French, and they are often used in spoken language, but the RAE does not recognize them.

- Onomatopoeia with nominal, adverbial or adjectival sense such as "buuff" (means boredom), "grrrr" (means angry), "tic tac" (to designate delay), "muak" (to send a kiss).

[1] Tuenti is a Spain-based social networking service. It features many tools common to social-networking sites. It allows users to set up a profile, upload photos, link videos and connect with friends; recently a chat application has been added.

[2] The Real Academia Española, RAE, is the official royal institution responsible for overseeing the Spanish language. It is affiliated with national language academies in twenty-one other hispanophone (Spanish-speaking) nations through the Association of Spanish Language Academies.

- Emoticons. Emoticons are defined by Yus [4] as "combinations of punctuation whose union produces different abstract facial expressions and other iconic signs". Its meaning can be varied, just as onomatopoeia, but the most common use is the representation of moods.

- Can be found acronyms abbreviations commonly used words in English as "WTF" (What the fuck?) Or "LOL" (laughing out loud) .

- There is also a list of common expressions in spanish as "bn" ("buenas noches" - good evening), "tq" ("te quiero" - I love you) which are abbreviated with initials. Often there is a high possibility of polysemy, this is caused by a limited number of combinations of letters in groups of 2 or 3, and it's depending on the context of the sentence.

- Shortening of words such as "peli" instead of "película" (movie), "xtremo" instead of "extremo" (extrem). This is also very common in spoken language.

- Substitution of syllables per digit, as for example in "to2" ("todos" - everyone). These numeral expressions have a phonic equivalence with the replaced syllables.

- Deformation of words such as "tronxo" instead of "tronco" (boy), "desfaxar" instead of "desfasar" (offset). These deformations can occur in words of the native language and in the language model that have been seen previously.

- It's too common to use ratings as a sign of exclamation and question as an intensifier ("¿¿??" o "!!!!!"). This use is closely linked to the now common use in comics.

- The syllables of the malformations of words can be also trimmed cut in its anterior, posterior and interior part. These cuts are very common in the language and are derived from the vowel relaxations of use spoken language, such as "andao" instead of "andado", "comio" instead of "comido", "istdo" instead of "instado", or "dstinto" instead of "distinto." Most of these malformations have common orthographic patterns and have been widely studied.

III. Proposed solutions to textual analysis

According to the cases seen in the previous section, the problems about the modifying the spelling has been bounded on the following assumptions for the problem of reconstruction of texts:

- There are words, phrases or graphemes that are completely new with respect to the standard use of language, such as onomatopoeia, emoticons, and abbreviations of expressions in different languages or loanword. In these cases, the solution is a direct replacement for a word or expression used. In other words, it's necessary to create a dictionary of terms of substitution. Examples:

;)	=> "Guiño"
☹	=> "Enfadado"
aagggg	=> "Queja"
party	=> "Fiesta"
tq	=> "te quiero"
bn	=> "Bien" / "Buenas noches"
WTF	=> "Sorprendido"

This is the most common solution.

- There is another case in which the analyzed word is a mutation of another word that already exists. This particularity case will be solved through the identification and replacement of prefixes, suffixes and infixes. Examples:

- ao	=> - dao
x -	=> ex -
n -	=> en-
ds-	=> dis-
k-	=> que
-q-	=> que

On a word can apply a multiple transformations, depending on the degree of degradation of this. Examples:

kdao	=> quedado
smrart	=> esmerarte

This method could solve most of the deformities commented before, including the use of numerical expressions which replace syllables.

The substitution of words can cause problems due to the inherent polysemy of terms used, as in the example of "bn". Why must consider the context when making substitutions?. In the aforementioned case of "bn", the replacement by "buenas noches" (good night) would be only made in the context in which can be found words as "night", "sleep", "night", "dinner", etc.

IV. Sentiment analysis solutions

The majority of approaches to sentiment analysis nowadays are aimed at calculating polarity of words that comprise the text.

Polarity is considered within this context to the intent of the review (positive or negative) that an issuer gives to a given issue.

The polarity of a term is calculated either through an automatic sorting process and system training analysis (see the work of Pang, Lee, and Vaithyanathan [8]), or through by a language expert (see the work of Turney [9]).

There are a great number of methods of analysis of polarity with variations in the weighted calculus.

In Spanish, among the most important works in this field is The Spanish OS Calculator Brooke Tofiloski and Taboada [7], which includes the use of different polarity assignments syntactic elements and includes use of intensifiers.

```
Dictionary
{
      myDictionary: HashMap
      infixs: HasMap
      getWordsPosibilities (searchedword: String, context:String)
      {
            returnvalue -> {}
            if searchedword is into myDictionary:
                  for wordincontext in context
                        if wordincontext is myDictionary[searchedword][‘context’]:
                              returnvalue -> returnvalue + myDictionary[searchedword][‘value’]
            return returnvalue
      }
      applySustitutions (searchedword: String)
      {
            returnvalue -> {}
            for (infix, infixvalue) in infixs
                  if searchedword contains infix:
                        returnvalue -> returnvalue + replaceInfix (searchedword, infix, infixvalue)
            return returnvalue
      }

}

searchedword: String;
context:String

posibilities = Dictionary->getWordsPosibilities (searchedword, context)

if (empty (posibilities) == true):
      posibilities -> Dictionary.applySustitutions ()
```

Fig. 1. Pseudocode of a method of reconstruction of texts

Within the spanish language has a denial treatment particularly different from other languages, such as that of the English, where the double denial has a different meaning. In studies of Taboada et al. [10] the morphological information of the sentence analysed is used to quantify the effect of negation, while the studies of Yang [11] and Fernández Anta [12] consider only if the denial is at the left of the denial expression (Yang believes that the scope of negation to the right is 2 terms, and Fernández Anta extends the range up to 3).

Text analysis in social networks presents special features, which involves not only the use of a specific vocabulary, but also a different grammar. There are, for example, some jobs that use a group of Twitter messages as a corpus. In the study of Petrovic et al. [13], the corpus has been generated by 97 million of real commentaries, while in the corpus of Pak and Paroubek [14] the comments are self-generated. This article will opt for a different perspective, trying to "translate" texts to a standard Spanish rather than trying to identify the new grammar, which facilitates the application of sentiment analysis.

V. A METHOD OF RECONSTRUCTION OF TEXTS

According to the linguistic variations on the second point, a reconstruction system text based on the combined use of lexical and semantic solutions is proposed, applying first a word substitution dictionary and, secondly, a dictionary replacement morphemes (distinguishing whether these are the beginning, middle or end of a word). In case of a direct replacement, the second case is not applied, because the direct replacement is assumed like correct.

To solve the problem of polysemy with limited boundaries is proposed a criterion for replacements based in a context factor, ie, propose a context (a region of text surrounding the word assessed) with certain properties. These properties are referred to the inclusion of terms in semantic field of the word replaced. In Figure 1 you can see the pseudocode to solve this problem.

In the pseudocode is not specified explicitly the behaviour of "replaceInfix" function, which would have a different behavior for infixes, suffixes and prefixes.

The reconstruction process of the text would be preceded to a pre-analysis to see if the reconstruction is "profitable" or not computationally required. This pre-analysis can be a statistic on how many words in text would be necessary rebuild, making the comparison with a Spanish corpus.

The optimum solution would also include all of the above in an iterative process that generates all possible situations and word substitution and selects a criterion to decide which of all solutions is the best.

The proposal, in this sense, include a grammatical analysis of the sentence generated that would be compared with the analysis of a corpus collected from a well-formed sentences in Spanish. The proximity (proximity in the sense of similarity between the grammatical structures) of the phrase to the well-formed structures corpus gives the "quality" of the substitutions.

VI. Proposal of sense analysis

This proposal of sense analysis is based on a word-by-word analysis, including the use of intensifiers and attenuators. The analysis scheme is extended from the words to the paragraphs.

Since pre-sentiment analysis process has made a "translation" to standard Spanish, it is not necessary to take into account grammatical and lexical problems.

Prior to the application of the algorithm, the first action is divide the text into sentences, never mind that these are ordinated or subordinated.

Then it will be done a pre-processing of the text in which the adverbial sentences are replaced by adverbs, to make a word by word analysis (e.g. "better" with "better") that facilitates a precise detection of polarity of the expressions (understanding by polarity a "positive" or "negative " expression).

Also every word that do not have a relevant value in the analysis (as adverbs, pronouns, articles and conjunctions) are removed from the text with the intention to increasing the processing speed.

Calculating the polarity of the sentence is carried out through a variation of the proposal by Brooke, Tofiloski formula and Taboada [6]. The variation of this system is on the one hand in implementing Yann considerations [7], which includes a change in the polarity by intensifiers, and the other side by the expansion of the analytical system pairs a set of sentences of a text until the paragraph level.

To make the analysis the following considerations will apply:

Each bigram (set of 2 words) will have a numeric value.

A phrase of x words contain x-1 bigrams, bigrams will be formed for each word of the sentence and its preceding.

First, the polarity of each word bigram which may vary between 1 and -1 according to a table of equivalence (for example: "bueno"-good- would have a value of 0.5, "mejor" -better- would have a value of 1.0, "malo"-bad- would have a value -0.5 or "mejorable" -improved- would have a value -0.2.

If a word with a negative value is followed by a word with a positive value (such as "good bad") both change its value to negative.

If a word is given preceded or succeeded by an intensifier then the intensifier increased the value of the preceded word multiplies it the value of intensifier increased in one unit.

Intensifiers values vary from -1 to 1. Intensifiers with value less than 0 will be considered attenuators, and those with value equal to 0 not offer any intensification.

The numerical value of a phrase is the sum of the values of bigrams that contains. Each phrase will associate additional value to the sum of the values that form the intensifiers.

If a sentence is considered as one of the monograms forming a bigram, with its values of polarity and intensification, a paragraph will be considered as sum of its bigrams (sets of 2 sentences).

The polarity values of a group of bigrams will be calculated as follows:

$$t(a,b) = \begin{cases} 1, & a \geq 0, b \geq 0 \\ -1 & i.o.c., \end{cases}$$

$$u(a,b) = \begin{cases} (a+1)\cdot(b+1), & if\ a \neq 1\ or\ b\ \neq 1 \\ 1, & if\ a = 1, b = 1 \end{cases}$$

$$V_F = \sum_{j=2}^{N}\Big((M_{j-1}^{P} + M_{j}^{P}) \cdot t(M_{j-1}^{P}, M_{j}^{P}) \cdot u((M_{j-1}^{I}), (M_{j}^{I}))\Big)$$

$$V_I = \sum_{j=1}^{N}\left(M_{j}^{I}\right)$$

Where:

t:	Function to calculate the intensifier value
u:	Function to calculate the polarity value of 2 bigrams
V_F :	Function to calculate the polarity value of a set of bigrams
V_I :	Function to calculate the intensifier value of a set of bigrams
M^I_j:	value of intensifier monogram in the position j
M^P_j:	value of polarity of monogram in the position j
N:	Number of monograms

The same procedure can be used to analyse any establishment, be it a list of words, a phrase or a whole group of paragraphs.

I. Results

Both algorithms, the reconstruction of the text and sentiment analysis, have been implemented in Python. The information relating to the intensity values and polarities are JSON files encoded into the following structure:

```
{
        "muy":  {"multiplicador": 0.5}
        "buenas":  {"multiplicador": 0.5}
        "bien": {"sentificador": 0.5},
        "alegra": {"sentificador": 1},
                “feliz”: {"sentificador": 1},
                “muchísimo”: {"multiplicador": 1}
                “poco”: {“multiplicador”: -0.2}
        }
```

In this structure the word and its values intensification ("multiplicador") and polarity ("sentificador") have specified.

For the implementation of dictionaries substituting words and morphemes, have also generated JSON files where substitutions and the semantic field of the context in which these should be performed are indicated.

```
 { “bn":          [
                  {"acepcion":"bueno"},
                  {"acepcion":"bien"},
                  {"acepcion":"buenas noches",
                      "contexto": {
                          "rango": 30,
                          "valores": ["noche", "cama"]
                      }
                  }
                ],
        "m":    [
                      {"acepcion":"me"},
                      {"acepcion":"mio"}
                ]
        "q":    [
                      {"acepcion":"que"}
                ]
        }
```

As we can see in the group context ("contexto") has two labels: range ("rango") refers to the maximum number before or after the analysed word which can contain the semantic field, and values ("values") words that are included the values that can be found within the semantic field.

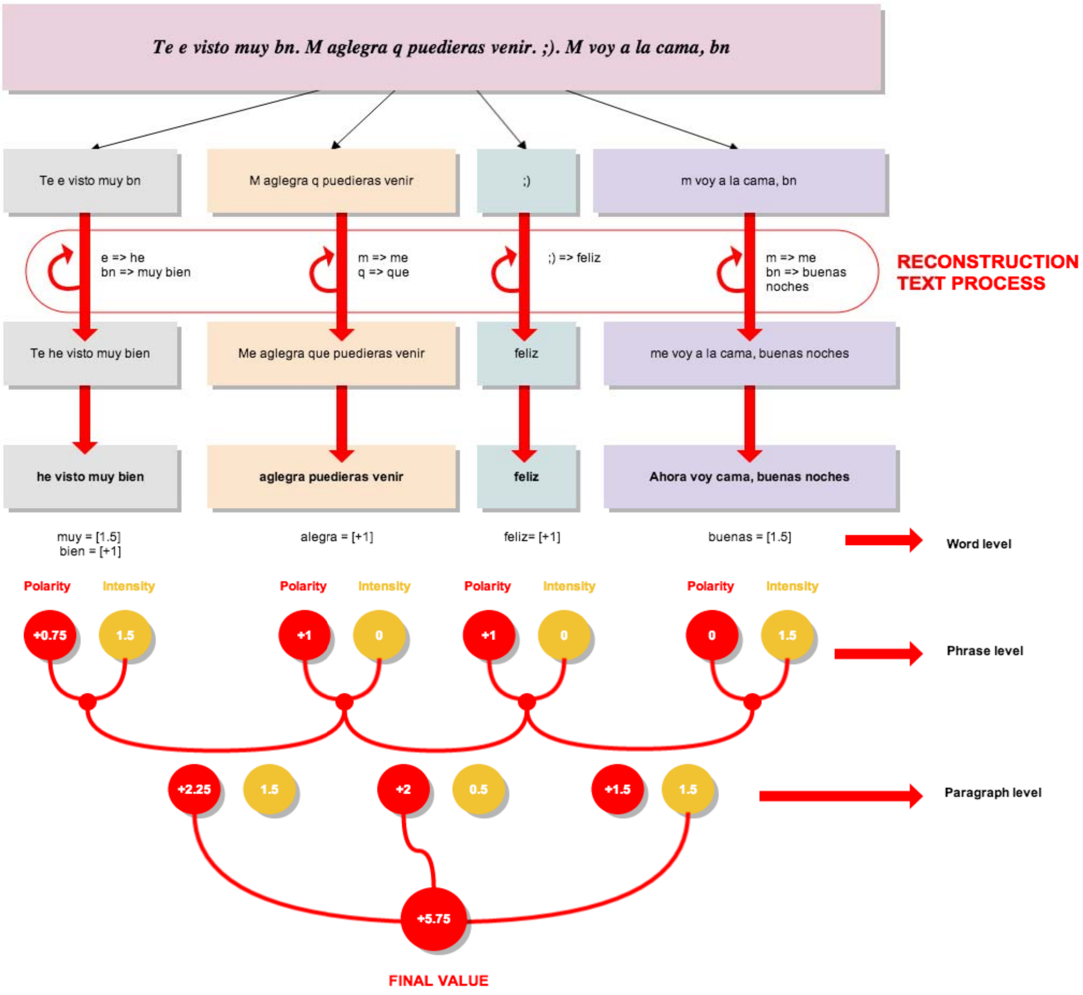

Fig. 2. Complete process diagram

Python classes generated supported the inclusion of one or more dictionaries of various types and include options for stricter checks or context.

Also the aim is identify the extent to which a sentence has been well rebuilt a lexical analysis of the components of this, to see if that structure is commonly used in the Spanish language.

Everything explained is reflected in example of "" (*"Te e visto muy bn. M aglegra q puedieras venir. Ahora me voy a cama, bn"*), which is very simple and easy to understand (the dictionaries described above these lines are used in the example).

Currently it has begun working with texts from Twitter, where the character limit makes the language used in it be enough schematic.

The objective is collected of at least 10,000 Twitter comments to evaluate algorithms with them. To evaluate the effectiveness two factors will be revised: on the one hand, the "quality" of the substitutions made, and the other hand, categorizing sentences according to their intensity of sentiment.

AGRADECIMIENTOS

Acknowledgement. This work has been partially supported by the MINETUR project PIAR, TSI-100201-2013-20.

REFERENCES

[1] A. ALMANSA, A., FONSECA, O., y CASTILLO REDES, A. sociales y jóvenes. Uso de Facebook en la juventud colombiana y española Social Networks and Young People. Comparative Study of Facebook between Colombia and Spain. Comunicar, nº 40, v. XX, 2013, Revista Científica de Educomunicación; ISSN: 1134-3478; p. 127-135DOI: http://dx.doi.org/10.3916/C40-2013-03-03

[2] TORREGO GONZÁLEZ, Alba. Algunas observaciones acerca del léxico en la red social Tuenti. Tonos Digital; nº 21, July 2011. ISSN 1577-6921

[3] BLAS ARROYO, José Luis. La variación léxica. En: De Miguel Aparicio, E. (ed.).

[4] Panorama de lexicología. Barcelona: Ariel, 2009, p. 189- 219.

[5] YUS, Francisco. Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet. Barcelona: Ariel, 2010.

[6] ESCANDELL MONTIEL, Daniel. Ciberpragmática en ELE. Aspectos fundamentales para una comunicación. FIAPE. IV Congreso internacional: La enseñanza del español en un mundo intercultural. Jornadas pedagógicas. Santiago de Compostela, 17-20/04-2011.

[7] BROOKE, J., M. TOFILOSKI, y M. TABOADA. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. En Proceedings of the International Conference RANLP-2009, p. 50–54, Borovets, Bulgaria. ACL.

[8] Pang, B., L. Lee, y S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. En: Proceedings of EMNLP, p. 79–86.

[9] TURNEY, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL'02, p. 417–424, Stroudsburg, PA, USA. ACL.

[10] TABOADA, M., J. BROOKE, M. TOFILOSKI, K. VOLL, y M. STEDE. 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2):267–307.

[11] YANG, K.. 2008. WIDIT in TREC 2008 blog track: Leveraging multiple sources of opinion evidence. En E. M. Voorhees y Lori P. Buckland, editores, NIST Special Pu- blication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008).

[12] Petrovic, S., Osborne, M., Lavrenko, V. (2010). The Edinburgh Twitter corpus. SocialMedia Workshop: Computational Linguistics in a World of Social Media, p. 25–26.

[13] Pak, A., P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proc. of the Seventh Conf. on International Language Resources and Evaluation (LREC'10), (ELRA), Valletta, Malta, p. 19–21.