Automatic Emotional Personality Description using Linguistic Data*

Grigori Sidorov and Noé Alejandro Castro-Sánchez

Natural Language and Text Processing Laboratory,
Center for Research in Computer Science,
National Polytechnic Institute,
Av. Juan Dios Batiz, s/n, Zacatenco, 07738, Mexico City,
Mexico
sidorov@cic.ipn.mx

Abstract. In the paper, we present the system designed for a usage of a psychologist during analysis of a special type of texts – texts of emotional autoreflexive writing. On the basis of linguistic analysis, the psychologist can conclude about emotional state of a person or about a type of his personality. The system is aimed to assist the psychologist. The system has the following features: automatic morphological analysis, calculation of various statistical parameters (frequencies, lexical richness, etc.). The data about words with emotional connotations are given apart because these words represent person's current state. We implemented the mechanism for synchronization of measuring of temperature during text writing and the resulting text. Also, we describe the application of the system in the other field – analysis of political discourse in Mexico.

1 Introduction

One of the main tasks of computer linguistics is providing models for development of applied systems with various kinds of automatic linguistic analysis. Such systems can be applied in diverse areas for solving the problems specific for these areas.

One of the possible areas of application of linguistic data is psychology because it also treats human beings, as well as linguistics. Thus, it is possible to make conclusions about psychological state or about a personality type. For example, in [6] it is claimed that there is a relation between the frequency of usage of auxiliary words, like prepositions, pronouns, articles, etc., and several demographic parameters or personality features. Another example is correlation of suicide tendencies of poets, namely, those, who committed suicide, had more self references in the texts than those who did not [10].

* The work was done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute, Mexico (CGPI, COFAA, PIFI).

© *A. Gelbukh, S. Torres, I. López (Eds.)*
Advances in Artificial Intelligence
Research in Computing Science 20, 2006, pp. 89-94

One more interesting tendency is related with the detection of intentions to lie or hide information [8]. It is shown that less self-evaluating phrases are used; words with negative emotional evaluation are more frequent; cognitive complex markers are less used.

In the paper, we describe linguistic parameters that are used for automatic emotional personality description. The system that implements this evaluation is described for Spanish language. The proposed approach is rather universal and can be applied in other areas. We applied it for analysis of political discourse of presidential candidates in Mexico.

2 Linguistic Parameters

There are several linguistic parameters for various levels. We calculate standard statistics for text –number of types, number of tokens, number of sentences, medium length of sentence and of paragraph, percentage of vulgarisms, lexical richness (we used two different formulae to calculate it, see below); also, some features of morphological and syntactic structure; and a kind of semantic analysis related with the usage of words from previously prepared lists, like negative, positive, etc.

Let us explain a little bit a concept of lexical richness. Note that it is incorrect to simply calculate the number of lemmas in different texts because it depends non-linearly from the text length, according to the Heaps law [1].

We use two formulae for lexical richness calculation. The first one, index *Brunét* is calculated according to the formula:

$$W = N^{V^{(-0.165)}}$$

where N is text length taken in words, V is the number of different lexemes. Usually, the obtained values belong to the interval from ten to twelve. The lower is the value of this parameter, the greater is the lexical richness.

The other parameter is *Honoré* statistics. It is based on the idea that lexical richness in general is proportional to the number of lexemes used exactly once in the text. The following formula is used:

$$R = \frac{100 * \log N}{1 - (V_1 / V)}$$

where N is the length of the text calculated in words, V is the number of all lexemes used, and V_1 is the number of lexemes with frequency one. In this case, the greater is the resulting value, the greater is the lexical richness.

For calculation of these statistics or for any further analysis, it is important to perform morphological normalization. In our case, we used morphological analyzer for Spanish described in [2].

Apart from lemmatizing, this kind of morphological analyzers allows for calculation of frequencies of grammar forms, for example, verbs in first person, etc.

For resolving the homonymy of parts of speech, we used part of the SVMTool library that has such functionality. The package was trained for Spanish data. It uses the model of Support Vector Machines. The POS tagging accuracy of 96% is claimed.

As far as statistics related to syntactic structures is concerned, we calculate the number of various types of subordinate and coordinate constructions.

For semantic analysis, we take into account the scale related with positive and negative evaluations [9] Corresponding words were chosen in experiments by psychologists.

For example, the following words and their derivates are used:

Table 1. Fragment of the table of emotional words

Words with positive evaluation	Words with negative evaluation	
	Physical threat	Social threat
sincere	asphyxiate	shyness
honest	suffocate	failure
joy	faint away	rejection
kind	infarct	insult
inspired	assault	arrogance
pleasure	suicide	uselessness
calm	illness	awkwardness
contain...	heart...	shame...

3 Application in Psychology

Writing of the texts, where the stress events are modeled in positive perspective, helps to overcome the negative experience. This a special type of text that is called texts of emotional autoreflexive writing ([4], [5], [6], [7]).

The technique of this writing and its further analysis is developed at the psychological faculty of the National University of Mexico in collaboration with psychologists from the United States. The technique implies further thorough analysis of the text by a psychologist. The system allow for automatizing of this analysis, while before it had to be performed manually.

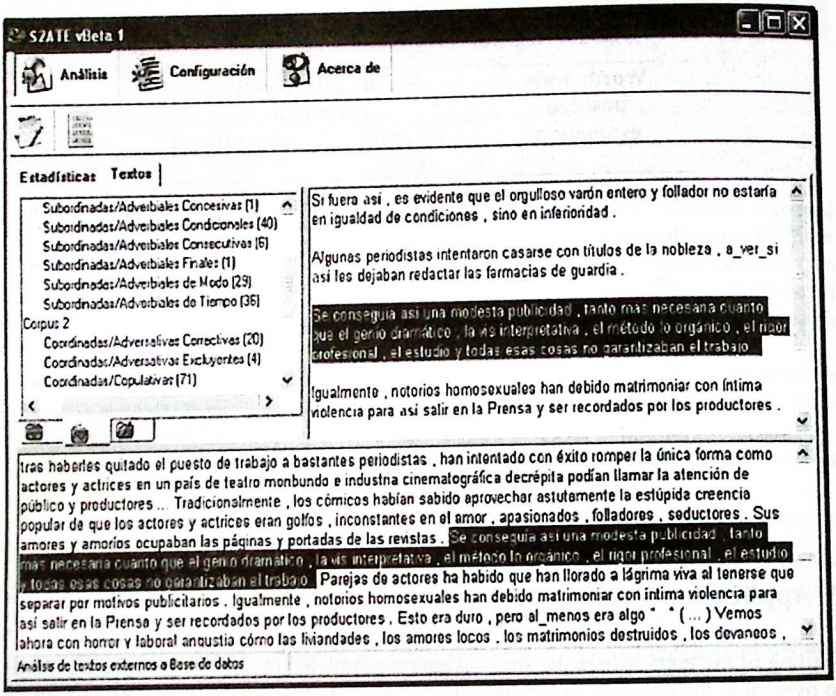
The system contains a database with the information about the persons (patients), their visits, and the corresponding texts. The texts can be grouped into the user-defined corpora or processed independently.

Results of automatic processing are presented for further manual analysis. Besides, there is a possibility to see contexts of usage of a given word, as it is shown in Fig. 1. One context is selected in the list of all contexts. The position of this context in the whole text is selected also, as it is shown at the bottom part of the figure.

There is a possibility of editing of the lists of vulgarisms, positive and negative words.

The system has the possibility of synchronizing with the text a special file that contains the measurements of the temperature taken during text writing for each phrase. Thus, a psychologist can choose a textual fragment and verify at the temperature diagram the corresponding values.

Fig. 1. Contexts of the word "*así*"



4 Another Application: Political Discourse

The developed system is rather universal instrument. We also applied it to analysis of the political discourse of the Mexican presidential candidates. Obviously, the data about the temperature were not available.

For the time being, there are three parties that have chances to win the elections. We can denote them as "party of the right" (PAN), "party of the left" (PRD) and centrist party (PRI). We had access to discourses of candidates of PAN and PRI during their campaign of the 2000, Fox and Labastida; and to the discourses of Lopez-Obrador, the candidate of PRD for the 2006.

Totally, we analyzed 73 texts (41, 16 and 16 correspondingly).

The resulting statistical data are given in Table 2.

Table 2. Statistics of pre-election discourses

	Lopez-Obrador	Labastida	Fox
Totally words	29,720	65,000	53,571
Tokens	9,956	16,434	20,926
Types	7,956	12,213	17,478
Lexical richness (*Honoré*)	447.3	481.3	472.9
Lexical richness (*Brunét*)	9.535	9.334	8.24

As can be seen, according to both statistics of lexical richness, it has the lowest values for the candidate of the "party of the left" (note that for the lexical richness *Brunét*, the lower is the value, the greater is the richness). It can be explained by his orientation to the poorest strata of population. Statistics of lexical richness have controversial values for the candidates of the "party of the right" and centrist party, though the differences are not large. We explain it by the fact that one of candidates touched more themes in his discourses, and, thus, had a chance to use more words with frequency one, which is the crucial factor in one of the statistics (*Honoré*).

Table 3. Usage of some emotional words

Evaluation	Lopez-Obrador	Labastida	Fox
Positive	thank (3) trust (3) security (3)	security (9) honesty (8) calm (4)	sure (6) security (4) thank (1)
Social threat	reject (6) criticism (1) failure (1)	shame (1) fury (1) criticism (1)	insult (4) criticism (1) despite (1)
Physical threat	attack (1)	attack (6) illness (3) wound (3)	accident (1) attack (1) illness (1)

As can be seen, the candidate of the "party of the left" uses less words with positive evaluation, more words with negative social threat and avoids words with negative physical threat.

On the other hand, the candidate of the centrist party uses more words with positive evaluation and with negative physical threat.

This data is preliminary and deserves more detailed processing and analysis. We give it here as an example of the application of the system and its possibilities.

5 Conclusions

We described the system that is designed for helping the psychologist during analysis of a special type of texts – texts of emotional autoreflexive writing. These texts represent stressing situations of life of a person and allow him to blow off; usually they are written by victims of crimes. On the basis of the linguistic analysis of the texts, the psychologist can make conclusions about his emotional state. The system is aimed to help the psychologist, because before the system was developed, this kind of analysis was done manually and took a lot of time.

The system is implemented for Spanish language. It performs automatic morphological analysis, calculates various statistics (frequencies, lexical richness, etc.), calculates apart data for words with emotional connotations, because these words are especially important for psychological analysis. The system has patients' database and a mechanism of synchronization of temperature measurements during text writing with text writing process.

The system is rather universal instrument for linguistic analysis oriented to psychological or social tasks. We applied it without any changes to analysis of political discourse, namely, to the pre-election discourses of Mexican presidential candidates.

References

1. Gelbukh, A. and G. Sidorov. Zipf and Heaps Laws' Coefficients Depend on Language. Lecture Notes in Computer Science N 2004, 2001, Springer-Verlag, pp. 330–333.
2. Gelbukh, A. and G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. Lecture Notes in Computer Science, N 2588, Springer-Verlag, 2003, pp. 215–220.
3. Gelbukh, A, G. Sidorov, SangYong Han. On Some Optimization Heuristics for Lesk-Like WSD Algorithms. Lecture Notes in Computer Science, N 3513, Springer-Verlag, 2005, pp. 402–405.
4. Domínguez, B., J. Pennebaker, y Y. Olvera. Procedimientos no invasivos para la revelación emocional. Diseño, ejecución y evaluación de la escritura emocional autorreflexiva. 2003.
5. Baños, R. M., S. Quero y C. Botella. Sesgos atencionales en la fobia social medidos mediante dos formatos de la tarea de Stroop emocional (de tarjetas y computarizado) y papel mediador de distintas variables clínicas. International Journal of Clinical and Health Psychology. ISSN 1697-2600. Vol. 5, no. 1, pp. 23-42. 2005.
6. Pennebaker, J., M. Mehl and K. Niederhoffer. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Reviews Psychology, 2003.
7. González, L., *et al.* El impacto psicofisiológico y cognoscitivo de la expresión emocional autorreflexiva sobre la salud. UNAM, México, 2004
8. Newman, M., *et al.* Lying Words: Predicting Deception From Linguistic Styles. Society for Personality and Social Psychology, vol. 29, No. 5, 2003, pp. 665-675.
9. Turney, P. D. and M. L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems, vol. 21, no. 4, pp. 315-346. 2003.
10. Stirman, W. and J. Pennebaker. Word use in the poetry of suicidal and non-suicidal poets. Psychosomatic Medicine, No. 63, 2001, pp. 517-522