A Metric for Automatic Word categorization

M. D. López De Luise - mlopez74@palermo.edu
Department of Informatics Engineering , Universidad de Palermo University
Av. Córdoba 3501, Capital Federal, C1188AAB, Argentina

***Abstract*—This paper presents a metric to be used by the working prototype WIH (Web Intelligent Handler). This metric (referred here as p_o) is designed to reflect main topic words and discriminate certain text profiles through word weightings. The actual version is designed only for Spanish web texts. Statistical analyses show that it is possible to differentiate text profiles upon p_o behavior. A poll is presented also, showing that it is a good main words discriminator. This paper is posted here as a new algorithm useful for Spanish text processing.**

Index Term — Text-Mining, Automatic summarization, morphosyntactic analysis**.**

I Introduction

Document handling is a part of index and retrieval process in the Web. It is a hard task due the complexity involved. Many approaches have been developed: noun phrases processing [1], morphemes processing [2], morphosyntactic analysis [3], etc. Sometimes the words searched in the documents are *expanded* with related words to improve recall. There are several alternatives to accomplish this work: expand upon morphology and syntax relations [4], morphosyntactic normalization of texts [5], conceptual and phonologic frames for word processing [6], etc.

Additional approaches have been developed for automatic paraphrasing ([7], [5], etc.), summarization of texts ([8], [9], [10], [11], etc.) and profiling of language usage of documents in the web ([12], [13], [14], [15], etc.). The contribution of this paper is twofold: to propose a metric to allow a kind of qualification of written texts and to provide an automatic word weighting for summarization activity.

Several mentioned algorithms use Natural Language Processing (NLP) which is a hard task due to the language expressions related to the writer's culture, education, geographical situation, etc. [16]. This paper presents a very simple proposal to process Web Spanish texts as part of the WIH (Web Intelligent Handler) prototype [17]. Its approach has many differences with NLP:

-In NLP it is required to process the entire document. WIH is not intended to process semantics but to extract certain features and words to represent it approximately.

-NLP works at a semantics level. WIH performs a simple morphosyntactic processing. It does just an analysis in the neighbor of each word.

-NLP considers five main levels of information [18]: lexical, morphologic, syntactic, semantic and pragmatic. WIH just makes use of some aspects from the first three levels.

-In NLP each level has to be robust enough to support upper abstraction levels. In WIH this is not necessarily true due to the simplicity of the process.

The remainder of this paper is organized as follows: section II presents the WIH prototype, which is the system that actually makes use of the proposed metric p_o; section III describes the metrics constraints; section IV defines p_o, section V shows statistical foundations; section VI, conclusions; and finally section VII states the main work remaining to be done.

II WIH Description

The Web Intelligent Handler (WIH) is a partially working prototype, first introduced in [17]. It has a three layered design (see Fig. 1):

-*Internal Structure:* gets data and metadata from the WWW and processes it to derive a set of Homogenized Basic Elements (HBE). These elements constitute a representation in an internal language.

-*Virtual Structure:* processes the actual stream of HBE and makes a structure named E_{ci} for the name in Spanish *Estructura de Composición Interna*, Internal Composition Structure in English. An E_{ci} is an oriented graph representing a statement in the original text. Sets of E_{ci} are then processed to make an E_{ce} (for the name in Spanish *Estructura de Composición Externa*, External Composition Structure in English, a supra-structure composed by a set of E_{ci} structures all of them related to the same text).

-*Visual Structure:* it works with a Virtual Network composed by the set of E_{ci} and E_{ce} structures. It can be considered as an interface between the Virtual Structure and any user.

The WIH prototype has implemented a first version of the Internal Structure and Virtual Structure. In the Virtual Structure there is a set of components to perform: the composition of E_{ci} and E_{ce} structures (Composition Engine, CE), insertion into the Virtual Net (Assimilation Engine, AE) through a set of functions (named effect functions, f_e) regulated by a set of dynamic parameters (metric functions, f_m). All the activity is controlled by a feedback system (composed by a general controller System Controller, a Manager for f_e named Metrics Manager, and a manager for f_m named Metrics Engine). The system performs all the activity through a set of f_e that can be changed dynamically. Some of such effect functions define the way a set of HBE is transformed to an E_{ci}, and how a set of E_{ci} is converted into an E_{ce}. For this activity, a metric p_o was defined to categorize each HBE and therefore determine how to process it.

III Metric Constraints and Working Hypotheses

As the metric p_o is used by the f_e to perform CE and AE activities, it has some constraints:

a-domain values must not exceed [-2.0; +2.0].

b-most of the values are 0.0.

c-must depict simple morphosyntactic considerations.

d-a 0.0 value must reflect an HBE with no influence in the global process.

e-an absolute value far from 0.0 must reflect an HBE with more influence than values nearby 0.0.

f- HBE's metric value must be able to be projected in some way to the E_{ci} level and E_{ci} value to an E_{ce} level.

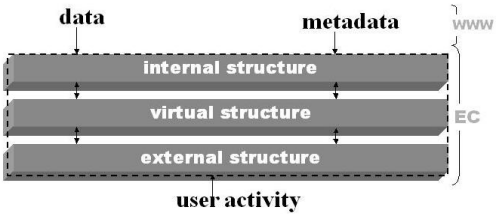

Fig. 1. Three layered structure of the EC. Data and metadata are extracted from the WWW.

g-must provide a qualification for HBE selection.
h-must provide a fuzzy manipulation of original texts.

There is a set of working hypotheses also:

a-There are many writer profiles.
b-Some sentences in a text are more representative of the main topic than others.
c-Sentences can have different quality.
d-Texts can have different styles.

IV The Metric Description

Set tables defining a numeric weighting of HBE whose original words match with certain prefix, define the quantification for each HBE. Table 1 presents a short list of some weightings and the Spanish word related to the HBE. The same procedure is followed for a list of special words (i.e. Some articles, conjunctions, modifiers, etc.).

The weightings range in [-1, +1]. When sentences are processed, this p_o values for HBE are combined with equation 1 to get a p_o^{Eci} for the sentence.

$$p_o^{Eci} = (p_i + p_{i-1})/2 \text{ , } \forall \text{ HBE}_i \qquad (1)$$

The metric is further used to evaluate the resulting E_{ce} with a p_o^{Ece} value (the more optimistic[1] p_o^{Eci} value is used).

V Statistical Analysis

Following the ***h*** constraint, the main steps to derive a fuzzy HBE model [19] are: frequency analysis, pattern and relationship extraction, model description and validation. The model description was performed in the previous section. The rest of the steps are detailed below.

A Frequency Analysis

To perform this testing, two samples were used:

1) Testing hypothesis a: To perform this testing, the profiles proposed are: forum, web index, document, and blog. A sample with 200 individuals was selected (50 of each one). The dot plot in Fig. 2 shows some potential outliers. There is not any confirmed outlier according to the information processed.

2) Testing hypothesis d: To perform this testing, three document styles are proposed: Literary, Technical and Messages. A sample with size 150 was selected (50 of each one). Texts were downloaded from sites specifically dedicated to each of these topics. The dot plot in Fig. 3 shows some potential outliers. There is not any confirmed outlier according to the information processed.

[1] The CS could change the optimistic concept.

TABLE 1
SPANISH WORD WEIGHTING.

muy	**0.7**
tan	**1**
pocos	**0.2**
mucha	**0.75**
muchos	**0.8**
bastantes	**0.7**
escaso	**0.15**
excesivamente	**1.5**
abundantemente	**1.3**
abundante	**1.1**
demasiada	**1.7**
exagerados	**1.9**
no	**-1**
sin	**-1**
desX	**-1**
inhX	**-1**
antiX	**-1**
disX	**-1**

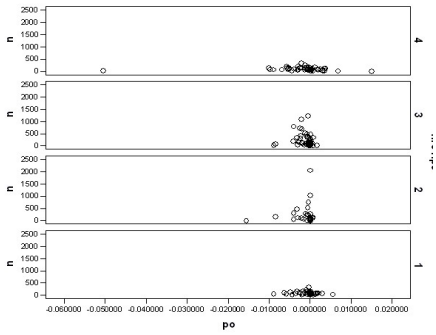

Fig. 2. Profiles dot plot. Set 1: forum. Set 2: web index. Set 3: documents. Set 4: blog. The p_o value corresponds to average of eq. (1).

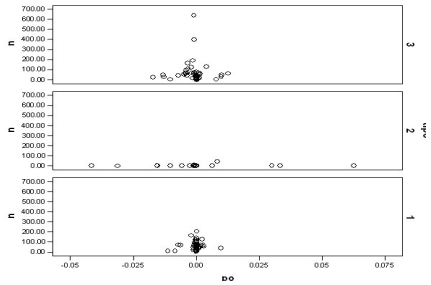

Fig. 3. Styles dot plot. Set 1: literary. Set 2: technical. Set 3: messages. The p_o value corresponds to average of eq. (1).

3) Testing hypothesis c: To perform this testing, the p_o mean value for each statement was calculated (let it be p_o^m). Afterwards, the GREATER_ZERO, LESS_ZERO and EQUAL_ZERO variables were defined as the number of p_o^m greater, less and equal 0.0 respectively. This value is intended to represent quality according to the profile/style (hypotheses c). The distribution was tested as a Binomial (eq. 2).

$$F(x_i) = p(X \le x_i) = \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \ldots + \binom{n}{k} p^k q^{n-k} \qquad (2)$$

Table 2 shows results: a Chi-Squared value greater than 0.05 indicates a Binomial distribution in all the subsets. The parameters are ***n***=50 cases and a specific ***p*** value found out for each subset in the table.

Therefore, each subset has a specific Binomial distribution of p_o^m. Parameter p indicates predominant EQUAL_ZERO for messages, decreasing for technical and literary.

GREATER_ZERO is higher for literary, decreasing for technical and message. All this can be thought as a valuation scale: literary, technical, and message; where literary seems the opposite of message and technical is in between. But a similar analysis for profiles can't give a definite scale for: blog, doc, forum and web index.

TABLE 2
BINOMIAL TEST FOR STYLES (A) AND PROFILES (B).

	subset	GREATER ZERO	LESS ZERO	EQUAL ZERO
		p (Chi²)	p (Chi²)	p (Chi²)
Style	literary	0.00720 (0.9434)	0.01000 (0.6795)	0.00280 (0.9998)
	message	0.00320 (0.9996)	0.00640 (0.9718)	0.01040 (0.6196)
	technical	0.00600 (0.9809)	0.01000 (0.6795)	0.00400 (0.9984)
Profile	doc	0.00200 (>0.9999)	0.01520 (0.0577)	0.00280 (0.9998)
	forum	0.00400 (0.9984)	0.00960 (0.7347)	0.00640 (0.9718)
	webindex	0.00280 (0.9998)	0.00640 (0.9718)	0.01080 (0.5563)
	blog	0.00720 (0.9434)	0.01200 (0.3608)	0.00080 (>0.9999)

4) Testing hypothesis b: To perform this testing, the maximum and minimum p_o^m values from each document were selected for processing. The related sentences were checked against main topic and secondary topic of the original document. Table 3 shows this analysis performed for one subset from styles (messages) and one from profiles (documents). These subsets were selected because the average number of sentences is the fewest for messages and the highest for documents.

It can be seen that the main topic is represented in most cases. Sometimes the secondary topic is represented. Considering the main topic as being represented by the max. and/or min. p_o, the main topic in the first sentence was studied. Table 4 shows the percentage of cases where the max and min values appear in the first sentence.

B Pattern and Relationship Extraction

To find out the behavior of p_o, the Shapiro Wilks test was performed for profiles and styles. In all cases the estimator is $p<<0.05$, therefore it can be said the populations have not a normal distribution (see Table 5).

TABLE 3
MAIN TOPIC FOR STYLES (A) AND PROFILES (B).

	%messages	%documents
Min. p_o		
-main topic	92	77
-secondary topic	8	15
-other phase	0	8
Max. p_o		
-main topic	83	38
-secondary topic	17	54
-other phase	0	8

TABLE 4
PERCENTAGE FOR STYLES (A) AND PROFILES (B).

	%messages	%documents
Max. p_o	29	0
Min. p_o	33	0
other	38	26

Pearson correlation between n and p_o denotes an increasing correlation according to the specific profile and style (see Table 6). The technical and blog subsets have a significant correlation (.94 and .97 respectively).

TABLE 5
SHAPIRO-WILKS TEST FOR STYLES (A) AND PROFILES (B).

class	Var.	n	Media	SD.	W*	p (1 tail)
literary	po	50	-3.5E-04	2.9E-03	0.73	<0.0001
	n	50	61.74	41.38	0.89	0.0005
message	po	50	1.1E-03	0.02	0.66	<0.0001
	n	50	4.14	5.94	0.34	<0.0001
technical	po	50	-1.0E-03	0.01	0.86	<0.0001
	n	50	65.46	105.24	0.56	<0.0001

a)

class	Var.	n	Media	SD.	W*	p (1 tail)
blog	po	50	-2.0E-03	0.01	0.64	<0.0001
	n	50	97.74	66.09	0.92	<0.0108
doc	po	50	-1.2E-03	2.0E-03	0.77	<0.0001
	n	50	246.08	273.52	0.78	<0.0001
forum	po	50	-6.8E-04	2.3E-03	0.85	<0.0001
	n	50	80.00	63.03	0.86	<0.0001
webindex	po	50	-9.1E-04	2.6E-03	0.47	<0.0001
	n	50	167.54	336.36	0.51	<0.0001

b)

As the distribution for subsets in both samples is not normal but a different correlation was detected for each one, a Krustal-Wallis test for p_o variability was performed. Therefore it can be stated if the subsets are statistically from different populations (that is, if Literary, Technical and Message belong to a main sample Style, and if Forum, Doc, Web index and Blog belong to a main sample Profile, as proposed here). As can be seen from Table 7, Chi-squared value p is 0.056 and 0.602 for styles and profiles. Both values are greater than the cut value 0.05. As a consequence, it cannot be said they are different populations.

The p_o median study for Styles and Profiles are depicted in Table 8. The Styles gives $p=0.071 > 0.05$, so it cannot be stated the median values are different. But for Profiles $p=0.00 < 0.05$ instead.

In a similar way, Krustal-Wallis test and median studies have been performed for n (number of non-zero p_o values). Table 9 shows a Sig. Value less than 0.05. As a consequence it can be said the n value is a good subset discriminator.

TABLE 6
CORRELATION TEST FOR STYLES (A) AND PROFILES (B).

style	correlation
literary	.43
message	.69
technical	.94

a)

profile	correlation
doc	.44
foro	.49
webindex	.84
blog	.97

b)

TABLE 7
KRUSTAL-WALLIS TEST FOR P_o WITH STYLES AND PROFILES.

Style	SubSet	parameter	value
	literary message technical	Chi-squared	1.014
		Degrees Freedom	2
		Sig. (p)	0.602
Profile	doc forum webindex blog	Chi-squared	7.555
		Degrees Freedom	3
		Sig. (p)	0.056

TABLE 8
MEDIAN FOR P_o WITH STYLES AND PROFILES.

Style	subset	parameter	value
	literary message technical	Chi-Squared	5.303
		df	2
		Sig.	0.071
Profile	doc forum webindex blog	Chi-Squared	18.720
		df	3
		Sig.	0.00

TABLE 9
KRUSTAL-WALLIS AND MEDIAN FOR N WITH STYLES AND PROFILES.

set	parameter	Median analysis	Variability analysis
Style	Chi-Squared	74.880	90.848
	df	2	2
	Sig.	0.000	0.000
Profile	Chi-Squared	15.607	11.680
	df	3	3
	Sig.	0.001	0.009

C *Validation*

A web page in Spanish was downloaded and text processed. A set of E_{ci} was derived. The p_o value for each HBE was calculated also. In order to perform the validation of p_o, HBE´s original words in Spanish were used. The resulting sets of words were classified according to the associated p_o value:

-The set of words whose $|p_o|< 0.110$ was selected and composed into two sentences representing the main topic. Let them be the Type I words.

-The set of words whose $|p_o|> 0.856$ was selected and composed into five sentences representing secondary topics. Let them be the Type II words.

A poll with 35 volunteers aged between 22 and 56 years old. All of them spend less than 10 hours a week in the Internet. They were asked to read the original text and to extract 2 to 10 most representative words from it. Fig. 4 shows a chart with the frequency of types of words extracted by volunteers. From the chart, it can be said that type I words are preferred mostly. There are a number of invalid words, because they weren't in the text (denoted by label *out of text*). A set of words extracted from text, describing heterogeneous details were represented and classified as *other in text*. *T*ogether, t*ype* I and II represent a bigger portion of the graph than *other in text* words.

As a part of the poll, volunteers also wrote 2 to 4 short sentences describing the topics in the text. These were compared with the set of 7 sentences (main and secondary topics) derived previously. Fig. 5 shows that derived sentences represent a high percentage of the topics answered.

Some additional characteristics of the metric results are:

-the p_o values have not a bias other than the opposite and ambiguity HBEs mentioned in the metric description section.

-the lexical categories of the HBEs detected as Type I and II from the text can be mainly classified as noun, verb and other (Fig. 6).

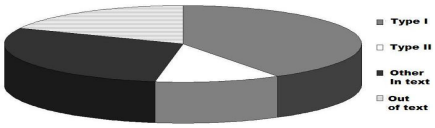

Fig. 4. Types of words 1: $|p_o|< 0.110$, 2: $|p_o|> 0.856$, other from text other not in text.

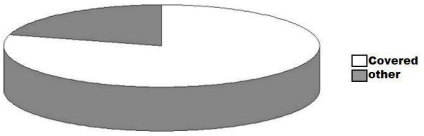

Fig. 5. Matching between topics declared and the 7 sentences derived.

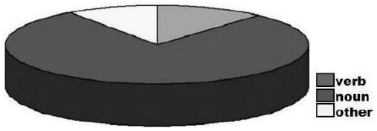

Fig. 6. Lexical category of words in the processed text.

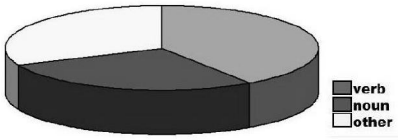

Fig. 7. Average p_o for each lexical category.

-the average p_o value is almost the same for each lexical category depicted previously (Fig. 7).

VI CONCLUSIONS

A p_o metric was presented and used here. From the results follows that it is not statistically possible to differentiate the *Text Styles* proposed here. But there is a clear distinction for the following *Writing Profiles*: doc, forum, webindex and blog. Therefore, this metric is invariant to document size and mentioned Text Styles but it is useful to detect certain writing profiles.

The p_o value was evaluated in a poll to assess its ability to detect main topics and relevant words. Statistics show that there is a pretty good relation between them and p_o values close to zero.

VII FUTURE WORK

It remains to do a better tuning of the weighting for different HBEs for texts in Spanish. It also has to be applied in other languages as well.

ACKNOWLEDGMENT

The author gratefully acknowledges the contributions of Dr. J. Ale and Prof. M. Bosch for their work on the original version of this document.

REFERENCES

[1] C. Berrut, P. Palmer. "Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing". ACM Conference on Research and Development in Information Retrieval. 1986.

[2] E. Wehrli. "Design and Implementation of a Lexical Data Base". European Chapter Meeting of the ACL. Proceedings of the second conference on European chapter of the Association for Computational Linguistics. Genova. Suiza. pp. 146 - 153. 1985.

[3] H. Assadi. ”Knowledge Acquisition from Texts: Using an Automatic Clustering Method Based on Noun-Modifier RelationShip”. European Chapter Meeting of the ACL. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Madrid, Spain. pp. 504 - 506. 1997.

[4] C. Jacquemin, J. L. Klavans, E. Tzoukerman. ”Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax”. Conf: Meeting of the Association for Computational Linguistics. ACL: Annual Meeting of the ACL. 1997.

[5] C. Brun, C. Hagège. ”Normalization and Paraphasing Using Symbolic Methods“.ACL: Second International workshop on Paraphrasing, Paraphrase Acquisition and Applications, Sapporo, Japan. 2003

[6] A. A. Gulla, S. N. Moshagen. ”A sign Expansion Approach to Dynamic, Multi-purpose Lexicons”. International Conference on Computational Linguistics. Proceedings of the 16th conference on Computational linguistics - Volume 1. Copenhagen, Denmark. pp. 478 - 483. 1996.

[7] C. Fabre, C. Jacquemin.”Boosting Variant Recognition with Light Semantics”. International Conference on Computational Linguistics. Proceedings of the 18th conference on Computational linguistics - Volume 1. Saarbrücken, Germany. pp. 264 - 270. 2000.

[8] H. Assadi. ”Knowledge Acquisition from Texts: Using an Automatic Clustering Method Based on Noun-Modifier RelationShip”. European Chapter Meeting of the ACL. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Madrid, Spain. pp. 504 - 506. 1997.

[9] C. Berrut, P. Palmer. ”Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing”. ACM Conference on Research and Development in Information Retrieval. 1986.

[10] F. Ibekwe-Sanjuan. ”Terminological variation, a means of identifying research topics from texts”. International Conference on Computational Linguistics. Proceedings of the 17th international conference on Computational linguistics - Volume 1. Montreal, Quebec, Canada. pp. 564 - 570. 1998.

[11] L. Vangehuchten. ”La identificación y la selección de léxico a partir de un corpus de discurso económico empresarial en Español como lengua Extranjera con fines específicos”. Proc. TALC 06. España. 2004.

[12] O. Santana Suárez, Z. Hernández Figueroa, G. Rodríguez Rodríguez. ”Morphoanalysis of Spanish Texts: Two Applications for Web Pages”. Lecture Notes in Computer Science, (2722). pp. 511-514. ISSN: 03029743. 2003.

[13] T. Vosse. “Detecting and Correcting Morpho-Syntactic Errors in Real Texts”. The Third Conference on Applied Natural Language Processing, pp. 111-118. 1992.

[14] K. Kukich. ”Techniques for Automatically Correcting Words in Text”. ACM Computing Surveys, Vol. 24, No. 4. 1992.

[15] J. M. Pazos Bretaña, A. Pamies Bertrán. “Detección automatizada de colocaciones y otras unidades fraseológicas en un corpus electrónico”. Letras de Hoje. Porto Alegre. v. 41, nro 2. pp. 23-36. 2006.

[16] M. Bargalló, E. Forgas, C. Garriga, A. Rubio, “Las lenguas de especialidad y su didáctica”, J. Schnitzer Ed. Universitat Rovira I Virgili. Tarragona. Chapter 1 (P. Schifko Wirtschaftsuniversität Wien). pp. 21-29. 2001.

[17] M. D. López De Luise, “A Morphosyntactical Complementary Structure for Searching and Browsing”. In Advances in Systems, Computing Sciences andSoftware Engineering. Proc. Of SCSS 2005. Springer. pp. 283 - 290. 2005.

[18] G. Leech. ”Introducing corpus annotation”. Corpus Annotation: Linguistic Information from Computer Text Corpora. R. Garside, G. Leech, A.M. McEnery Eds. London: Longman. 1997.

[19] A. Morillas Raya, “Introducción al análisis de datos difusos”. Málaga University. Econometría y Estadística. España. www.eumed.net/libros/2006b/amr. 2006.