IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728,p-ISSN: 2319-765X, Volume 8, Issue 1 (Sep. - Oct. 2013), PP 05-10
www.iosrjournals.org

Analysis for the significance of statistical word-length features in genre discrimination of Hindi texts

Hemlata Pande[†] and H. S. Dhami[*]
Department of Mathematics, Kumaun University, S. S. J. Campus Almora, Uttarakhand, INDIA

Abstract: *In automatic text categorization procedure, quantifiable features' information is extracted from a text and on the basis of the information the text is sorted as a category. This information consists of values of set of one or more measurements, where the measurements can be considered as frequencies or function of frequencies of linguistic elements.*

In the process of text classification and genre discrimination, the role of the systematic study of word length and the analyses of word-length statistics of different texts has been established by researchers for various languages. In the present paper an attempt has been made to test the contribution of quantitative word length features in classification of written texts of Hindi Language by extracting quantitative measures with the help of word length profiles and frequencies.

Keywords: *Categorization, Feature, Text, Word-length.*

I. Introduction

Těšitelová (1992) has mentioned that "For deeper understanding of an object or phenomenon and thus also language it is necessary to know not only its qualitative but also its quantitative side (by counting and measuring)." In Quantitative Linguistics (QL), languages are studied by counting and by determination of the frequencies of various language elements such as words, parts of speech, word combinations, lexemes etc. QL mainly refers to statistical linguistics, which offers the methods of making conclusions in text processing on the basis of previously gathered statistics. Statistical techniques have got significant advances in language processing tasks due to wide availability of various linguistic resources and machine learning texts. An account of the contribution of statistics in linguistic studies can be found in the books of and Manning and Schütze (1999) and Gómez(2013).

Kelih et al (2005) have cited that "Word length is a central characteristic in the framework of quantitatively oriented linguistics". Diverse studies corresponding to the word length frequencies have been done by various researchers. In this context we can cite the works of Dittrich (1996), Frischen (1996), Alekseev (1998), Leopold (1998), Rottmann (2003), Lupsa and Lupsa (2005), Antić et al (2006) Grzybek (Ed.)(2006), and Pande and Dhami (2012) to mention only a few.

Classificatory tasks in computational linguistics are mainly performed on the basis of text genre detection and authorship attribution. Genre determination of text usually refers to identification of the kind of the text. Research articles, news articles, court decisions, home pages, poems, novels are some examples of genres of texts. Genre discrimination practices have applications in many natural language processing tasks. Kelih et al (2005) in their work have stated that individual style of texts can be quantitatively described in the field of Stylometry. Can and Patton (2004) have utilized the word lengths in texts and word lengths in vocabulary for the comparison of the old and new writing style of two authors. Renkui and Minghu (2012) have used word length as one of the quantitative stylistic features. The applications of word length in text classification on the basis of topic, genre and/or author can be found in the works of Mikros and Carayannis (2000) for texts of Modern Greek, Kelih et al (2005) for Russian texts, Grzybek et al (2005), Antić et al(2006) and Stadlober and Djuzelic (2006) in the case of Slovenian texts, and of Grieve(2005) for variety of language, to mention only a few.

The aim of the paper is to examine the contribution of statistical; word length features in classification of texts of Hindi language. With the help of information about the proportion of words of different length in a text various statistical measures have been calculated. The length of a word has been considered in terms of the number of graphemic components present in the word. Next section 2 gives a brief outline of the quantitative features selected and the texts taken for the study. In section 3 we have reported the methodology while the results and conclusion have been given in section 4.

[†] Corresponding author(Email: hlpande@rediffmail.com)
[*] Email: drhsdhami@gmail.com

II. Texts And Word-Length Features

For the present work, various texts have been selected from '*Navbharat Times*' for the period from 30 Oct. 2008 to 5 Dec. 2008 and from the corpus '**ELRA-W0037**' . The texts taken for the analysis have been mentioned in the following Table 1:

Table 1. Texts considered for the present study

S. No.	Texts	S. No.	Texts
1.	62 texts from '*Navbharat Times*' (नवभारत टाइम्स) under '*Sampadakeey*' (संपादकीय) section.[‡]	2.	20 essay (tagged as essay under literature in the corpus ELRA-W0037)[§]
3.	37 articles from '*Navbharat Times*' under '*Nazaria*' (नज़रिया) section[‡]	4.	45 texts from novel (tagged as novel under literature in the corpus ELRA-W0037)[§]
5.	94 media articles (tagged as media in the corpus ELRA-W0037)[§]	6.	29 stories from story section (tagged as story under literature in the corpus ELRA-W0037)[§]
7.	50 texts from news (have been taken from 'ranchiexpress' section of the corpus ELRA-W0037, each text is collection of news articles)[§]		

Texts mentioned in number 1, 3 and 5 in the above table (Table 1) have been considered as media texts and similarly texts of number 4 and 6 have been assumed as creative writings. Thus the categories: media, news, essay and creative writing have been considered and total 337 texts have been chosen (74 creative writings, 20 essays, 193 media texts and 50 news texts).

The frequencies of words of different lengths have been determined in all these 337 texts, by considering a word as a combination of graphemic units, where the length of a word has been determined in terms of the number of vowels, consonants, vowel marks "*Matra*" (मात्रा) and the symbols corresponding to the half letter, contained in the word. Words joined by '-' and '~' have been considered as separate words and the numerical and non-Hindi alphabetic words have not been considered. Before counting the length of words of the texts, first the occurrences of ':' in place of '○:' have been replaced by '○:' manually. Following 15 features have been determined from the word length frequency data:

Table 2. List of features calculated from word length frequency and profile data and corresponding used formulae

S. No.	Feature	
1	Mean word length	$m = \frac{\sum l f_l}{\sum f_l}$
2	Average deviation from arithmetic mean *m*	$\delta_{m=\frac{\sum f_l \lvert l-m \rvert}{\sum f_l}}$
3	II central moment	$\sigma^2 = \frac{\sum f_l (l-m)^2}{\sum f_l} = \mu_2$
4	III central moment	$\frac{\sum f_l (l-m)^3}{\sum f_l} = \mu_3$
5	IV central moment	$\frac{\sum f_l (l-m)^4}{\sum f_l} = \mu_4$
6	Standard coefficient of dispersion or Coefficient of variation	$\frac{\sigma}{m}$
7	Frequencies of words of length 3 relative to words of length 2	$\frac{f_3}{f_2}$

[‡] These Corpora have also been analysed for letters' frequencies and pattern of words in our earlier studies (Pande and Dhami (2010, 2013)).
[§] Written Hindi monolingual corpus section of the corpus **'ELRA-W0037'.** The corpus is acquired from **ELDA** (Evaluations and Language resources Distribution Agency <http://www.elda.org/>). "ELRA catalogue (http://catalog.elra.info), The EMILLE/CIIL Corpus, catalogue reference: ELRA-W0037"

8	Frequencies of words of length 4 relative to words of length 3	$\frac{f_4}{f_3}$
9	Pearson's coefficient β_1	$\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}}$
10	Pearson's coefficient β_2	$\beta_2 = \frac{\mu_4}{\mu_2^{\,2}}$
11	Normalized frequency of words of length 2	$\frac{f_2}{\sum f_l}$
12	Normalized frequency of words of length 3	$\frac{f_3}{\sum f_l}$
13	Normalized frequency of words of length 4	$\frac{f_4}{\sum f_l}$
14	Normalized frequency of all words of length greater than 4 (longer words)	$\frac{\sum_{l>4} f_l}{\sum f_l}$
15	Coefficient of dispersion mentioned by Juilland(1970)(cited in the book of Těšitelová (1992))	$D = \frac{\sigma}{m\sqrt{n-1}}, \quad n = \sum f_l$

III. Method

After determining the above mentioned 15 word length features for all 337 texts, under study, we have applied the Discriminant analysis (with the help of StatistiXL[**], where the value of minimum allowed tolerance has been set to 0.0001) process to categorize a text as a category out of the selected two categories and the results of categorization process between different categories have been given in the following tables (Table 3 to Table 8). In these tables, news, essay, media and creative writing have acronyms nw, es, md and cr respectively and predicted group represents the number of texts classified as a category by the process of determination of the group membership from the linear classification functions. In the Holdout method, to classify a particular case, first it is removed from the data set used for classification analysis, which (data set) is then used to classify the omitted case by considering it as a new case.

Table 3. Classification Table for news texts and essays:

Actual Group	Predicted Group (Std)		Correct Classification rate	Predicted Group (Holdout)		Correct Classification rate
	nw	es		nw	es	
nw	50	0	1.000	50	0	1.000
es	0	20	1.000	0	20	1.000
Overall Correct Class. Rate			1.000			1.000

Table 4. Classification Table for news texts and creative writings:

Actual Group	Predicted Group (Std)		Correct Classification rate	Predicted Group (Holdout)		Correct Classification rate
	nw	cr		nw	cr	
nw	50	0	1.000	50	0	1.000
cr	0	74	1.000	2	72	0.973
Overall Correct Class. Rate			1.000			0.984

[**] http//www.statistixl.com

Table 5. Classification Table for news texts and media texts:

Actual Group	Predicted Group (Std)		Correct Classification rate	Predicted Group (Holdout)		Correct Classification rate
	nw	md		nw	md	
nw	50	0	1.000	50	0	1.000
md	24	169	0.876	27	166	0.860
Overall Correct Class. Rate			0.901			0.889

Table 6. Classification Table for essays and creative writings:

Actual Group	Predicted Group (Std)		Correct Classification rate	Predicted Group (Holdout)		Correct Classification rate
	es	cr		es	cr	
es	17	3	0.850	16	4	0.800
cr	2	72	0.973	3	71	0.959
Overall Correct Class. Rate			0.947			0.926

Table 7. Classification Table for essays and media texts:

Actual Group	Predicted Group (Std)		Correct Classification rate	Predicted Group (Holdout)		Correct Classification rate
	es	md		es	md	
es	15	5	0.750	10	10	0.500
md	23	170	0.881	29	164	0.850
Overall Correct Class. Rate			0.869			0.817

Table 8. Classification Table for creative writings and media texts:

Actual Group	Predicted Group (Std)		Correct Classification rate	Predicted Group (Holdout)		Correct Classification rate
	md	cr		md	cr	
md	174	19	0.902	170	23	0.881
cr	9	65	0.878	10	64	0.865
Overall Correct Class. Rate			0.895			0.876

We have also classified the texts by considering two types of texts: training and testing. The training texts have been used for extracting the statistical information for each category while the testing texts are the texts used to check the results of the classification process and were not used for the statistical information extraction of the categories. 25 news texts, 10 essays, 37 creative writings (15 stories and 22 novel texts) and 97 media texts (31 texts from no. 1, 19 from no. 2 and 47 from no. 3 mentioned in Table 1 of the previous saction) have been taken as the training texts while the remaining texts have been treated for the testing purpose, for classification as a category out of the two categories. The results of classification between different categories have been depicted in the following table:

Table 9. Classification results between different categories of testing articles

1. News vs. essay

Actual Group	Predicted Group		Correctly Classified
	nw	es	
nw	25	0	1.000
es	1	9	0.900
Overall Correct Class. Rate			0.971

2. News vs. creative writing

Actual Group	Predicted Group		Correctly Classified
	nw	cr	
nw	25	0	1.000
cr	1	36	0.973
Overall Correct Class. Rate			0.984

3. News vs. media

Actual Group	Predicted Group		Correctly Classified
	nw	md	
nw	25	0	1.000
md	10	86	0.896
Overall Correct Class. Rate			0.917

4. Essay vs. creative writing

Actual Group	Predicted Group		Correctly Classified
	es	cr	
es	8	2	0.800
cr	1	36	0.973
Overall Correct Class. Rate			0.936

5. Essay vs. media				6. Media vs. creative writing			
Actual Group	Predicted Group		Correctly Classified	Actual Group	Predicted Group		Correctly Classified
	es	md			md	cr	
es	7	3	0.700	md	81	15	0.844
md	18	78	0.812	cr	4	33	0.892
Overall Correct Class. Rate			0.802	Overall Correct Class. Rate			0.857

Thus the above mentioned tables (Table 3 to Table 9), show that the classification rate for all categories for the testing texts as well as for all considered texts is greater than or equal to 0.800 except in the case of essays for classification as essay or media (classification rate is 0.700 for the testing texts and 0.750 for all considered texts). The classification tables for all the texts and testing texts for classification among all considered four categories have been presented below (Table 10 and Table 11):

Table 10. Classification Table for all the considered articles for classification in four categories

Actual Group	Pred. Group (Std)				Correct Classification rate	Pred. Group (Holdout)				Correct Classification rate
	es	md	nw	cr		es	md	nw	cr	
es	15	3	0	2	0.750	11	6	0	3	0.550
md	23	131	20	19	0.679	23	124	23	23	0.642
nw	1	0	49	0	0.980	3	0	47	0	0.940
cr	1	7	1	65	0.878	1	8	1	64	0.865
Overall Correct Class. Rate					0.772					0.730

Table 11. Classification Table for 168 testing articles among four considered categories

Actual Group	Predicted Group				Correct Classification rate
	es	md	nw	cr	
es	8	0	1	1	0.800
md	17	58	6	15	0.604
nw	0	0	25	0	1.000
cr	0	4	0	33	0.892
Overall Correct Class. Rate					0.738

On the basis of above two tables also it can be said that more cases of misclassification arise between the categories essay and media texts. Therefore it can be assumed that in the case for word length features, media text are written similar to essay. Next we have classified the texts by considering three categories only: essay/ media (md), news (nw) and creative writing (cr). The classification tables for all the texts and the testing texts have been depicted in the form of following tables:

Table 12. Classification table for all the considered texts for classification among three categories essay/ media, news and creative writing

Actual Group	Predicted Group (Std)			Correct Classification rate	Predicted Group (Holdout)			Correct classification rate
	md	nw	cr		md	nw	cr	
md	165	25	23	0.775	160	28	25	0.751
nw	0	50	0	1.000	0	50	0	1.000
cr	8	1	65	0.878	9	1	64	0.865
Overall Correct Class. Rate				0.831				0.813

Table 13. Classification Table for 168 testing texts among three categories

Actual Group	Predicted Group			Correct Classification rate
	md	nw	cr	
md	80	10	16	0.755
nw	0	25	0	1.000
cr	4	0	33	0.892
Overall Correct Class. Rate				0.821

Thus the Table 12 and Table 13 depict that by considering the media text similar to essay (or by assuming one category for media texts and essays) the classification rate for all the three categories is more than 0.75 in all cases.

IV. Results And Conclusion

The previous section illustrates that, in the case of Hindi, for the presence of words of different lengths in various texts; the writing style of media texts is similar to the essays. Only on the basis of simple measures, which can be easily determined with the help of frequencies of words of different lengths in various texts and word length profiles of the texts, the classification rate upto 75% has been obtained for classification among the categories media/essay, news and creative writing. The results of classification process between different categories can be pointed as that the word length features plays signification role in classification of different written texts of Hindi language on the basis of genre of texts or in other words word-length features are stylistic characteristics of Hindi language as in the cases of other languages (Mikros and Carayannis (2000), Can and Patton (2004), Kelih et al (2005), Grzybek et al (2005), Grieve(2005), Antić et al(2006), Stadlober and Djuzelic (2006) and Renkui and Minghu (2012)). Therefore for the discrimination of genre the word length features should be included in the data set for the statistical measures selected for the process.

Acknowledgment

Authors are grateful to the University Grants Commission (UGC), New Delhi, INDIA for providing financial assistance in the form of Post doctoral fellowship to the first author. The research has been sponsored by the UGC under the 'UGC Dr. D. S. Kothari Post Doctoral fellowship scheme'.

References

[1]. Alekseev, P.M. (1998): Graphemic and Syllabic length of words in text and vocabulary. *Journal of Quantitative Linguistics, 5:1-2*, 5-12.
[2]. Antić, G. , Kelih, E. and Grzybek, P. (2006), Zero syllable words in determining word length, in P. Grzybek (Ed.): *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*, 117-156, Springer, Netherlands.
[3]. Antić, G., Stadlober, E., Grzybek, P. and Kelih, E.(2006). Word Length and Frequency Distributions in Different Text Genres. *From Data and Information Analysis to Knowledge Engineering*, 310-317, Springer, Berlin Heidelberg.
[4]. Can, F. and Patton, J. M. (2004). Change of Writing Style with Time. *Computers and the Humanities 38*, 61-82.
[5]. Dittrich, H. (1996): Word length frequency in the letters of G.E. Lessing. *Journal of Quantitative Linguistics, 3:3*, 260-264
[6]. Frischen, J. (1996): Word length analysis of Jane Austen's letters. *Journal of Quantitative Linguistics, 3:1*, 80-84
[7]. Gómez, P. C. (2013). *Statistical methods in language and linguistic research*. Equinox Publishing Ltd.
[8]. Grieve, J. W. (2005). Quantitative Authorship Attribution: A History and Evaluation of Techniques. *Master's thesis Simon Fraser University*, 2005. URL http://summit.sfu.ca/system/files/iritems1/8840/etd1721.pdf Retrieved on 17-05-2013.
[9]. Grzybek, P. (Ed.)(2006). *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*, Springer, Netherlands.
[10]. Grzybek, P., Stadlober, E., Kelih, E. and Antić, G. (2005). Quantitative text typology : The impact of word length. In: C. Weihs and W. Gaul (Eds.), *Classification- The Ubiquitous challenge,* 53-64, Springer, Heidelberg;.
[11]. Juilland, A., Brodin, D. and Davidoviteh, c. (1970). *Frequency dictionary of French words*. The Hague.
[12]. Kelih, E., Antić, G., Grzybek, P. and Stadlober, E. (2005). Classification of Author and / or Genre? The Impact of Word Length . In C. Weihs, W. Gaul (Eds.). *Classification, the Ubiquitous Challenge,* 498-505, Springer Berlin-Heidelberg,.
[13]. Leopold, E. (1998): Frequency spectra within word-length classes. *Journal of Quantitative Linguistics*, 5:3, 224-231
[14]. Lupsa, D. A. and Lupsa, R. (2005). The law of word length in a vocabulary. *Studia Univ. Babes-Bolyal, Informatica*, Vol. L, No. 2, 69-80.
[15]. Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing,* MIT Press, Cambridge.
[16]. Mikros, G. and Carayannis, G. (2000): Modern Greek Corpus Taxonomy. *2nd International Conference on Language Resources & Evaluation, LREC 2000* . Available at < http://www.lrec-conf.org/proceedings/lrec2000/pdf/351.pdf> Retrieved on 17-05-13
[17]. Pande, H. and Dhami, H. S. (2010). Mathematical Modelling of Occurrence of Letters and Word's Initials in Texts of Hindi Language. *SKASE Journal of Theoretical Linguistics* vol. 7, no. 2, 19-38.
[18]. Pande, H. and Dhami, H. S. (2012): Model generation for word length frequencies in texts with the application of Zipf's order approach. *Journal of Quantitative Linguistics, 19:4*, 249-261
[19]. Pande, H. and Dhami, H. S. (2013). Mathematical Modelling of the Pattern of Occurrence of Words in Different Corpora of the Hindi Language. *Journal of Quantitative Linguistics, 20:1*, 1-12
[20]. Renkui, H. and Minghu, J. (2012).Discrimination of Chinese Quantitative Style Features Based on Text Clustering. *11th International Conference on Signal Processing (ICSP)*, 2012 IEEE, 21-25 Oct. 2012, Beijing.
[21]. Rottmann, O. (2003). Word length in the Baltic languages-are they of the same type as the word lengths in the Slavic languages? *Glottometrics 6*, 52-60.
[22]. Stadlober, E. and Djuzelic, M. (2006). Multivariate Statistic Methods of Quantitative Text Analysis. In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. 259-276, Dordrecht, NL: Springer.
[23]. Těšitelová, M.(1992). *Quantitative Linguistics*. John Benjamins Publishing Company Amsterdam/ Philadelphia.