Defence Science Journal, Vol. 60, No. 4, July 2010, pp. 415-419

SHORT COMMUNICATION

Index of Garbledness for Automatic Recognition of Plain English Texts

P.K. Saxena, Pratibha Yadav*, Girish Mishra

Scientific Analysis Group, Delhi-110 054
*E-mail: saxenapk@hotmail.com, *pratiyadav@yahoo.com, mishrag76@rediffmail.com*

ABSTRACT

In this paper, an Index of Garbledness (I_G) has been defined for automatic recognition of plain English texts based on linguistic characteristics of English language without using a dictionary. It also works for continuous text without word break-up (text without blank spaces between words). These characteristics, being vague in nature, are suitably represented through fuzzy sets. A fuzzy similarity relation and a fuzzy dissimilarity measure have been used to define this Index. Based on a threshold value of the Index, one can test whether the given text (continuous without word break-up) is a plain English text or not. In case the text under consideration is not a plain text, it also gives an indication to what extent it is garbled.

Keywords: Index of garbledness, automatic recognition, fuzzy sets, fuzzy similarity relation, fuzzy dissimilarity measure, fuzzy pattern recognition, cryptanalysis

1. INTRODUCTION

Every language has got some inherent characteristics in terms of occurrences of letters out of the alphabet set in words, which in turn constitutes sentences. Each language has got its own grammar, which defines the constraints on word spellings and syntax of sentences. In each language there are some “vowel” like letters whereas others are “consonant” like, which are basically characterised by the phonetic properties.

The language English has five vowels (‘Y’ also treated sometimes as 6th vowel) and 21 consonants. Some letters such as E, T, O, A, I occur more frequently, whereas others such as B, J, X, W, Z occur occasionally in any given English text[1]. Similarly, there are certain letter combinations such as WX, ZX, BX (diagraphs), which are very rare whereas combinations such as TH, ER, HE occur more frequently. Such characteristics and affinity of certain letters with some specific letters makes one language distinguishable from the other and also enables one to check if the given text is meaningful valid plain text of the language or it is a random text (sequence of letters occurring randomly).

While dealing with Cryptography[1,2] one normally processes the text message without word break-up and transforms it to make it unintelligible by destroying the characteristics of the language as described above. But when it comes to crack such coded messages without knowing the key, one needs to check if the text generated out of trials is really developing into meaningful plain text of the language or not. Since word break-up is not known, for such intermediate tests to look for plain texts (solutions), it was felt that an Index can be defined suitably which may differentiate plain texts from random texts of English language. This would be a useful tool for cryptanalysis[2].

An attempt has been made in this direction using fuzzy sets[3] and fuzzy similarity relations[4,5,6]–an approach normally followed for fuzzy pattern recognition[7,8,9] and classification problems.

One of the main issues related to fuzzy set theoretic approach is to select suitable fuzzy sets (characterising linguistic features), called feature fuzzy sets (FFS). For defining index of garbledness (I_G) to identify plain English texts, 10 Feature fuzzy sets on 26 character alphabet set of English language have been derived and used as in case of the classification of texts of different regional languages[10] as described below.

2. FEATURE EXTRACTION

2.1 Characteristics of English Language

Every language has got certain characteristics, which can be taken as its signature. Similarly, English language also has certain characteristics[12]. Some of these characteristics are as under:

1. All the letters of English are not used equally. Some letters are used very frequently whereas others are used rarely.
2. Most prominent crests in the frequency graph of letters appear at E, T, O, A, N, I, R, S and most prominent troughs are marked at the letters J, K, Q, X, Z.
3. Vowels constitute nearly 40 per cent of the texts with average gaps between these to be 2.5

Received 22 September 2009, Revised 1 March 2010

4. All vowels except U are high frequent letters.
5. Vowels have more affinity to contact consonants than vowels themselves.
6. Consonants have more affinity to contact vowels than consonants themselves.
7. Vowels contact with wide variety of letters.
8. Letters contacting low frequent letters are mostly vowels.
9. Most frequent reversals (XY.... YX) have one vowel and one consonant.
10. The contact of letters on right and left are different for different letters.
11. Some letters forms doublets whereas others never occur twice in a sequence.

2.2. Fuzzification of Characteristics of English Language

For 26-character normal texts, the feature fuzzy sets (FFS) are defined on the basic set $F = \{A, B, C, ... Z\}$ of 26-characters with membership values to be defined suitably which must capture the above mentioned characteristics of English language. The following fuzzy sets based on characteristics of a given text have been chosen as FFS.

2.2.1 Fuzzy Features

The fuzzy sets are defined below:

(1) Number of occurrences of letters.
(2) Very high, high, medium, low and very low categorisation of occurrences of letters.
(3) High occurrences of doublets of the letters.
(4) High occurring diagraphs starting with a letter.
(5) High occurring diagraphs ending with a letter.
(6) Wide variety of left contacts of a letter.
(7) Wide variety of right contacts of a letter.
(8) Wide variety of two-sided contacts of a letter.
(9) Occurrence of the letter with which highest diagraph starting with a given letter is formed.
(10) Occurrence of the letter with which highest diagraph ending with a given letter is formed.

For all these, fuzzy sets on the set F of 26 letters are suitably constructed[10] by defining characteristic values in [0, 1]. In case of (1) above, the normalisation of frequencies by dividing them by the highest frequency brings down the value in [0, 1]. Thus

$$\mu_1(x) = \frac{freq(x)}{\text{Highest freq of single letters}}$$

However, in case of (2) above, the characteristics have been defined intuitively by the following expressions:

$$\mu_2(x) = \left\{ \begin{array}{lll} 0.1 & if & \mu_1(x) < 0.05 \\ 0.2 & if & 0.05 \le \mu_1(x) < 0.1 \\ 0.4 & if & 0.10 \le \mu_1(x) < 0.2 \\ 0.6 & if & 0.20 \le \mu_1(x) < 0.4 \\ 0.8 & if & 0.40 \le \mu_1(x) < 0.9 \\ 1.0 & if & 0.90 \le \mu_1(x) \end{array} \right\}$$

For constructing fuzzy sets corresponding to (3), (4) and (5) above, the corresponding frequencies of the diagraphs have been divided by the highest single letter frequency, since no diagraph frequency can exceed this number. Similarly, the characteristic values corresponding to (6), (7) and (8) are obtained by dividing the number of different letters with which a specific letter is contacting (right, left or two-sided) by 26 as that is the highest value for variety of contacts. Lastly, for fuzzifying (9) and (10) as listed above, the highest single letter frequency is used as normaliser.

Thus, ten fuzzy sets characterising the given text from which these have been extracted are formed. It can be noted that though in the above fuzzy sets, the vowels have not been identified exclusively, the fuzzy sets expressing different kinds of diagraphic patterns implicitly take care of the affinity of some letters (vowels) with other letters. However, the difference and identification of vowels can further be exploited by working out the confidence factors (CFs) of the following occurrences using trapezoidal kind of fuzzy sets on the interval 0-100, as described below:

2.2.2 Intervals of Occurrences of Vowels and Consonants

Based on the statistics of English language, a general pattern of occurrence of vowels and consonants in plain English texts[12] is as follows:

Percentage of occurrence of	Interval per cent
(a) Vowels	
(i) A, E, I, O, U	35-45
(b) Consonants	
(ii) T, S, R, N, H	30-40
(iii) D, C, L, M, P, F, Y, G, B, V	20-30
(iv) J, K, Q, W, X, Z	0-5

The confidence factors for all these four percentages can be worked out using the general form of trapezoidal type of fuzzy set, as shown in Fig. 1.

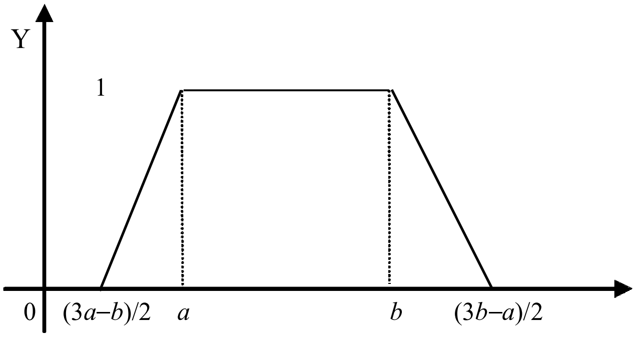

Figure 1. Trapezium type of fuzzy function.

Mathematically,

$$\mu_{ab}(x)=\left\{\begin{array}{ccc} 1 & if & a\le x\le b \\ \dfrac{3a-b-2x}{a-b} & if & \dfrac{3a-b}{2}\le x\le b \\ \dfrac{3b-a-2x}{b-a} & if & b\le x\le\dfrac{3b-a}{2} \\ 0 & elsewhere & \end{array}\right\}$$

For (i) to (iv) above, *a* and *b* are chosen as the given intervals and confidence factors (CF) can be calculated. These CF's will be combined with the index obtained through the 10 FFSs to work out the combined index, as described below.

3. INDEX FORMULATION

3.1 Similarity Fuzzy Relation and Dissimilarity Measure

For the comparison of two fuzzy sets, μ_1 and μ_2, different kinds of fuzzy relations can be used. For our purpose, the following two fuzzy relations were considered:

$$\mu_R(x,y)=\frac{k}{k+|\mu_1(x)-\mu_2(y)|^l}$$

$$\mu_R(x,y)=e^{-k|\mu_1(x)-\mu_2(y)|^l} \qquad (1)$$

where *k* and *l* are parameters that can be suitably chosen[11]. $\mu_1(x)$ and $\mu_2(y)$ represent membership values corresponding to plain text and the given text, respectively. Both of these give almost similar kinds of results except scaling. Hence the first one has been used in designing the present system, the parameters *k* and *l* have been adjusted through learning by providing the system different plain English texts of varying lengths. The values finally fixed up are k = 0.7 and l = 2. This fuzzy similarity relation induces a 26 × 26 fuzzy similarity matrix *R* with entries from [0,1]. These have 10 such fuzzy similarity matrices corresponding to the 10 characteristics considered in Section 2.1. The following measure of dissimilarity[10] has been used:

$$D=\alpha(26-Trace\ \ R)+\beta\sum\sum(\mu_R(x,y)-\mu_R(y,x)) \qquad (2)$$

where the ΣΣ infact is nothing but the sum of symmetric differences of entries in the matrix R. Again, here α and β are some parameters, which can be learnt through known examples. In the present case, α and β were fixed up as 1 and 0.3, respectively. These values were adjusted through learning process using known plain texts of different lengths.

3.2 The Algorithm and Decision Process

Part I

The extraction and stabilisation of characteristic values of various fuzzy sets for clear English texts was the first problem. The values change with length and are never fixed in two texts. Therefore, for fairly faithful prototyping, quite a large number of plain English texts of various lengths (200 – 400 characters) were run and the fuzzy sets (all 10 say $P_1, P_2, \ldots, P_{10}$) were calculated. The variations in characteristic values were studied. Based on these, some standard prototype fuzzy sets were constructed which are used in the Part II of the algorithm.

Part II

This module is the main module for the calculation of Index of Garbledness (I_G). A given text (whatever it may be) as a continuous string of characters is passed to this module and all 10 fuzzy sets $C_1, C_2, \ldots.., C_{10}$ are constructed in the specified manner as already pointed out. Now the question comes to compare these with the corresponding fuzzy sets $P_1, P_2, \ldots.., P_{10}$ prototyped for plain English texts in first *Part I*.

For each i = 1,.....,10 ; $\mu_1 = P_i$ and $\mu_2 = C_i$ and the fuzzy similarity relation (1) is applied to calculate a 26 × 26 fuzzy similarity matrix $R^{(i)}$. Using the expression in (2), dissimilarity measures defined as

$$D_i=\alpha\left(26-Trace\ R^{(i)}\right)+\beta\sum\sum\left(R^{(i)}(j,k)-R^{(i)}(k,j)\right)$$

are calculated. These give the distances of individual characteristics present in the text under scrutiny. But these have to be combined to arrive at an index which can measure the garbledness present in the text. Just averaging was tried first but that does not give good results. Therefore, a weighted average was considered suitable. Again the weights were adjusted through the outcomes of known examples and were fixed up as follows:

w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
.142	.132	.031	.093	.093	.115	.115	.093	.093	.093

The total dissimilarity was thus calculated as

$$D=\sum_{i=1}^{10}w_iD_i$$

This *D* has been combined with the confidence factors of the text being plain in terms of various percentages of vowels and consonants as discussed earlier. For correct plain texts, it is expected that all the four confidence factors CF_1, CF_2, CF_3, and CF_4 will have highest value 1, since the four percentages (vowels, high, medium and low consonants) will lie in the admissible interval. Thus the dissimilarity due to these factors could be measured by

$$DP=4-\left(CF_1+CF_2+CF_3+CF_4\right)$$

and the final index of garbledness present in the text has been defined as

$$I_G=D+DP$$

The value of I_G gives an indication about the garbledness present in the text.

4. RESULTS

The algorithm developed has been tested on a number of English texts of various types namely plain text, partially-garbled text (garbledness between 10 per cent and 50 per cent and between 50 per cent and 90 per cent), and totally-garbled text. Partially-garbled texts are prepared by replacing some of the characters in the plain-text by other characters. Preparation of totally-garbled text has been done by encryption of the plain-text with simple substitution scheme. Working of this algorithm is explained by the following examples.

4.1 Working of the Tool Developed for I_G

Example I:

Plain Text:

```
elandmatungasionthehistoryofthisgroupofislandslocatedon
thewestcoastofindiagoeswaybacktothestoneagesstoneagei
mplementshavebeenfoundatseveralsitesintheseislandsthec
oastalregionsandpresumablytheislandswerethehomeofthek
olifisherfolkinthedistantpastinthirdcenturybctheseislandsw
erepartofthemagadhanempireunderemperorashokaafterhis
deaththeislandschangedhandsfromonekingtoanotherfortw
othousandyearstheybelon
```

Values of the corresponding 10 D_i's are:
$D_1 = 1.01$, $D_2 = 1.28$, $D_3 = 0.00$, $D_4 = 0.33$,
$D_5 = 0.35$, $D_6 = 0.68$, $D_7 = 0.76$, $D_8 = 0.73$,
$D_9 = 2.02$, $D_{10} = 2.59$

Value of $D = \sum_{i=1}^{10} w_i D_i = 9.75$

Values of 4 confidence factors (CF_i's, i =1,2,3,4) are:
$CF_1 = 1.00$, $CF_2 = 1.00$,
$CF_3 = 1.00$, $CF_4 = 1.00$
Value of DP (= 4–(CF_1+CF_2+CF_3+CF_4)) = 0.00
Hence value of I_G (=D + DP) = 9.75

Example II:

Partially-garbled Text:

```
ehandmattndasipntreristpxopatrisdxptppaishandshpqate
dpntrenestqpastpaindiadpesnaouaqktptrestpneadesstpne
adeimphementsraueueenaptndatseuexahsitesintreseisha
ndstreqpastahxedipnsandpxestmauhotreishandsnexetrer
pmepatrekphiaisrexaphkintredistantpastintrixdqenttxou
qtreseishandsnexepaxtpatremadadranempixetndexempe
xpxasrpkaaatexrisdeatrtreishandsqrandedrandsaxpmpn
ekindtpanptrexapxtnptrptsandoeaxstreouehpn
```

Values of the corresponding 10 D_i's are:
$D_1 = 3.17$, $D_2 = 3.18$, $D_3 = 0.00$, $D_4 = 0.45$,
$D_5 = 0.48$, $D_6 = 1.94$, $D_7 = 1.83$, $D_8 = 2.10$,
$D_9 = 4.19$, $D_{10} = 3.83$

Value of $D = \sum_{i=1}^{10} w_i D_i = 21.17$

Values of 4 confidence factors (CF_i's, i =1, 2, 3, 4) are:
$CF_1 = 0.80$, $CF_2 = 1.00$,
$CF_3 = 0.70$, $CF_4 = 0.00$
Value of DP (= 4-(CF_1+CF_2+CF_3+CF_4)) = 1.50
Hence value of I_G (=D + DP) = 22.67

Example III:

Totally-garbled Text (a cryptogram):

```
kjwptzwroplwsqgprdkdqsrgucgnrdqslugoygnqsjwptsjgxw
rktgprdkhksrxgwsrgnqptqwlgkwcewxargrdksrgpkwlkssrg
pkwlkqzyjkzkprsdwmkekkpngoptwrskmkuwjsqrksqprdks
kqsjwptsdkxgwsrwjuklqgpswptyuksozwejcrdkqsjwptshku
krdkdgzkgnrdkagjqnqsdkungjaqprdktqsrwrywsrqprdqutxk
proucexrdkskqsjwptshkukywurgnrdkzwlwtdwpkzyqukopt
kukzykuguwsdgnrkudqstkwrdrdkqsjwptsxdwplktdwptsnu
gzgpkaqplrgwpgrdkungurhgrdgoswptckwusrdkcekp
```

Values of the corresponding 10 D_i's are:
$D_1 = 5.19$, $D_2 = 4.74$, $D_3 = 0.00$, $D_4 = 0.52$,
$D_5 = 0.51$, $D_6 = 2.39$, $D_7 = 2.26$, $D_8 = 2.59$,
$D_9 = 3.66$, $D_{10} = 3.37$

Value of $D = \sum_{i=1}^{10} w_i D_i = 25.25$

Values of 4 confidence factors (CF_i's, i =1,2,3,4) are:
$CF_1 = 0.00$, $CF_2 = 0.40$,
$CF_3 = 0.00$, $CF_4 = 0.00$
Value of DP (= 4-(CF_1+CF_2+CF_3+CF_4)) = 3.60
Hence value of I_G (=D + DP) = 28.85

It may be observed that out of four CF_i's, three are zero in the present example. This is due to the fact that the text here is a cryptogram based on a simple encryption scheme and hence is not as random as it would be in case of strong encryption schemes. In that case even CF_i's are all non-zero and index I_G goes up further.

The s/w based on this algorithm has been tested on 125 text files from each category of plaintext, partially garbled text and totally garbled text. Each text file has 400 characters. It is observed that in case of plain English texts, the value of this index is below 19. But in case of crypts and totally garbled texts, it is between 23 and 32. For texts having some plain text and some garbage (i.e. partially garbled text), the value of I_G in most of the cases is in between 19 and 23 and measures the plain text characteristics relatively present in the text i.e. the value of I_G gives an indication about the garbledness present in the text.

5. CONCLUSIONS

The index of garbledness (I_G) developed works very well for identification of plain English text. Therefore, it can be used for identification of the solution while going for exhaustive trial, especially in those systems where either the correct text is formed during a key trial or it is totally garbled. Even in case of cryptosystems, where partial texts could be formed by a key trial, the value of I_G indicates the degree to which the correct solution has been formed and thus may give a hint to the next key for trial in certain cases. Thus this I_G could be used for directed search of the reduced key space instead of sequential search. The value of I_G is also a powerful tool for fitness functions in case of genetic algorithms for cryptanalysis, which would help in selection of new generations from the old ones.

REFERENCES

1. Stinson, D.R. Cryptography: Theory and practice. CRC Press, 1995.
2. Schneir, B. Applied cryptography. Ed. 2, John Wiley and Sons, 1996.
3. Zadeh, L.A. Fuzzy sets, *Information and Control*. 1965, **8**, 338-53.
4. Kaufmann, A. Introduction to the theory of fuzzy subsets. Vol. **1**. New York, Santa Barbara, London, 1975.
5. Klir, G.J.& Folger, T.A. Fuzzy sets, uncertainty and information. Prentice Hall, Engle Wood Cliffs, New Jersey, 1988.
6. Zadeh, L.A. Similarity relations and fuzzy orderings. *Information Science,* 1971, **3**, 177-206.
7. Klir, G.J. & Yuan, B. Fuzzy sets and fuzzy logic: Theory and applications. Prentice Hall, 1997.
8. Rosenfeld, A. Fuzzy graphs. *In* Fuzzy sets and their applications, *edited by* KS Fu Zadeh L.A. and Shimura M. Academic Press, New York, 1975. pp. 77-95.
9. Zimmermann, H.J. Fuzzy set theory and its applications, Ed. 4. Kluwer Academic Publisher, Bonston / ordrecht/ London, 2001.
10. Yadav, Pratibha; Mishra, Girish & Saxena, P. K. Identification of Indian regional languages in romanized form. *In* Proceedings of the 6th International Conference, Advances in Pattern Recognition, ICAPR-2007, pp. 112-16.
11. Yadav, Pratibha; Pal, S.K. & Saxena, P.K. Identification of CVSD-modulated speech signals. *In* Proceedings of the 5th International Conference. Advances in Pattern Recognition, ICAPR-2003. pp.151-55.
12. Baker, H. & Piper F. Cipher systems. Northwood Books, London, 1982.

Contributors

Dr P.K. Saxena obtained his MSc (Mathematics) from Kanpur University and PhD (Algebra) from Indian Institute of Technology, Kanpur. He joined Scientific Analysis Group (SAG), DRDO in 1981 Presently working as Scientist H & Director, SAG. His areas of interest include: Algebra, cryptology, fuzzy logic, artificial neural networks and speech technology. He has published about 40 research papers in national, international journals and conferences.

Ms Pratibha Yadav obtained her MSc (Mathematics) from Delhi University in 1985. Presently she is working as Scientist 'F' at SAG and is heading Traffic Analysis Group. She has reviewed many research papers for well-known international conferences and journals. Her areas of interest include: Fuzzy sets, abstract algebra and cryptology. She has applied fuzzy sets to various problems of cryptology and linguistics.

Mr Girish Mishra obtained his MSc (Mathematics) and MPhil (Special Functions) from University of Rajasthan, Jaipur in 1998 and 2003 respectively. Presently, he is working as Scientist 'C' at SAG. His areas of interest are: Cryptology, soft computing and algebra. He has worked on areas of aanguage identification in written text, cryptosystem identification in ciphertext etc.