Repositorio Institucional de la Universidad Autónoma de Madrid

https://repositorio.uam.es

Esta es la **versión de autor** de la comunicación de congreso publicada en:
This is an **author produced version** of a paper published in:

Information Theory Workshop, 2006. ITW '06 Punta del Este. IEEE, 2006. 337-341

DOI: http://dx.doi.org/10.1109/ITW.2006.322834

Copyright: © 2006 IEEE

El acceso a la versión del editor puede requerir la suscripción del recurso
Access to the published version may require subscription

Detecting the Same Text in Different Languages

Kostadin Koroutchev[1,2]
Depto. de Ingeniería Informática
Universidad Autónoma de Madrid
28049 Madrid, Spain

Manuel Cebrián[3]
Depto. de Ingeniería Informática
Universidad Autónoma de Madrid
28049 Madrid, Spain

Abstract — **Compression based similarity distances have the main drawback of needing the same coding scheme for the objects to be compared. When two texts are translated, there exists significant similarity with no literal coincidence. In this article, we present an algorithm that compares the redundancy structure of the data extracted by means of a Lempel-Ziv compression scheme. Each text is represented as a graph and two texts are considered similar with our measure if they have the same referential topology when compressed. We give empirical evidence that this measure detects similarity between data coded in different languages.**

I. Introduction

Recently the problem of finding similarities and dependencies in textual data was treated by several authors (see [1, 2]) using general purpose compression algorithms, having in mind that the information carried out in two dependent texts will essentially lead to better compression of the concatenation of the texts than of each text in isolation.

The following measure of the divergence $d(.,.)$ between two texts, t_1 and t_2, is proposed:

$$d(t_1, t_2) \equiv \frac{L(t_1 \circ t_2) - \min(L(t_1), L(t_2))}{\max(L(t_1), L(t_2))}, \qquad (1)$$

where $\circ$ means concatenation and $L(X)$ denotes the length of the text X compressed using some compression algorithm that asymptotically reaches the entropy of X, when the length of X tends to infinity [1].

The measure $d(.,.)$ varies from zero for identical texts to one for totally different texts. Even for very similar texts it is usually more than 0.8 [3].

Because the common text compression algorithms are based on the presence of the repetitions in the text, it is clear that the only situation detectable by the compression algorithms is when literal repetitions of the texts are present.

This essentially leads to a paradoxical situation. Namely, one and the same text, written using two different, non-intersecting alphabets, is classified as totally different, because $L(t_1 \circ t_2) = L(t_1) + L(t_2)$. This can be very easily tested, for example, using the Cyrillic and the Latin version of the Universal Declaration of the Human Rights in Serbian[1]. If we care to eliminate the structure, imposed by the formatting rules of the document (white space and enumeration), we actually find these texts very dissimilar. But the two texts, on the other side, are exactly the same.

Therefore, the interesting question arises: Can we find an algorithm that can detect translations of one and the same text?

In this article we present a Lempel-Ziv(LZ) [4], inspired algorithm, that can detect whether some text is a translation of another text.

Text data, produced by humans, usually represent some concept by using unidimensional character string. Such text, essentially includes cross references and therefore repetitions imposed by the nature of the concept, as well as structure imposed by the rules of description of the concept, as for example language syntax and morphology. Usually, this type of data is compressed well by LZ. The compression is due to the high degree of predictability of the future of the text, looking at its past.

The compression of every text is achieved due to the repetitions in the text. The repositions are coded in a similar way. Actually, there are two sources of the ability to achieve compression in the human produced texts. On the one side, the rules that ought to be imposed to transmit the information, for instance, the grammar and on the other side, the structure of the concept itself that ought to be transmitted. For example, when we compress this article, the particle "the " will be compressed using the grammar, and the word "text" can be attributed to the context.

Resuming, the reasons for the compression of some set of similar strings is the structure, imposed due to (1) *the coding rules*, the structure, imposed due to (2) *the intrinsic structure of the transmitted concept* and the structure imposed due to (3) *the common initial source of the data.* In all the cases, we can regard that one of the strings is in a way a translation of the other. In this translation, the coding rules can change, but the common source and the internal structure of the concept are preserved.

It is clear, that detecting and separating the effect of these sources is at least an interesting task.

[1] Also with Inst. for Communications and Computer Systems, Bulgarian Academy of Sciences, 1113, Sofia, Bulgaria `k.koroutchev@uam.es`

[2] This work was partially supported by grants TIN 2004-07676-G01. of the Spanish Ministry of Education and Culture.

[3] Partially support by TSI 2005-08255-C07-06 of the Spanish Ministry of Education and Culture.

[1] Serbian language has the peculiarity, that it can use both alphabets – Latin and Cyrillic usually not mixing them in one and the same text.

In this article we are trying to find these effects, using LZ code representation of the string and ignoring the characters from the alphabet (the member s in the triplet). We will process the LZ code of a concatenation of a set of strings $L(t_1 \circ t_2 \circ t_3 ... \circ t_n)$, where we suppose that t_i is a translation of t_j. We extract all the information from the LZ coding and try to separate the types of compressibility from this information. We attempt to diminish the influence of the coding rules of the string (e.g. the grammar and the morphology).

The article is organized as follows: Sect. II explains the concepts and the algorithms used. Sect. III describes the data and the computer experiments, and finally Sect. IV offers a summary of the results.

II. Algorithm

In this article we are using LZ as a basic compression algorithm. LZ parses the string in one direction. If the string is coded up to some position p, the next portion of the string is coded by finding the position q in the already coded portion such that (1) the substrings starting at the positions p and q coincide; (2) the coinciding substring has the maximal length of all such strings and (3) if there exist more than one positions q with these properties, the maximal q is chosen. Thus, the portion of the text that is coded is represented as a triplet $(p-q, l, s)$, consisting of the displacement from the current position $p-q$, the length l of the substring that coincides and the symbol s, that follows the coinciding part of the text.

We are going to measure the similarity between two texts t_1 and t_2. To begin with, we apply the LZ algorithm to each text obtaining the typical LZ-triple set $\{(p-q, l, s)\}$ We don't care about the alphabet but only about its compression structure, so we can leave apart the symbol s, obtaining an equivalent set $G_{\text{LZ}} = \{(q, p, l)\}$, which can be interpreted as a graph with the positions of the text p and q as vertices, and edges between them, with weight $l > 0$ which is the length of the identical string in the positions p and q. This graph is extremely sparse.

If the substrings with length l in the positions p and q are identical, then the substrings at position $p+1$ and $q+1$ are also identical with length $l-1$, the same happens with $p+2, q+2, l-2$ and so on. Therefore, we can increase the density of the graph, by defining:

$$\begin{aligned} G^0 \quad = \{ \quad & (q+i, p+i, l-i), (p+i, q+i, l-i) \\ & \mid (p, q, l) \in G_{\text{LZ}},\ 0 \le i < l\} \end{aligned}$$

Members of G^0 with small l are generally imposed by the grammar and also by random matches between short strings due to the limited alphabet. Although the grammar is an interesting aspect, we prefer to ignore it in the present article and we prune the graph by deleting the edges with small weights. We can simplify the consideration by regarding all edges of weight more than some limit L as equivalent and thus, instead of regarding the weighted graph G^0, we consider a normal graph G^1_L defined as:

$$G^1_L = \{(q, p) \mid (q, p, m) \in G^0,\ m \ge L\}.$$

Some edges remain to be added to complete our graph: if we have edges (p, q) and (q, r) in G^1_L, it is clear that we have to add edges (q, p), (p, r) and (r, p) because of two reasons: (i) p can be compressed with position r, (ii) p and r belong to substrings with lengths greater than L which LZ identified as useful for achieving compression. This process can continue iteratively, until finally each node will be connected to all its reachable nodes. In other words we define

$$G_L = (G^1_L)^*,$$

where G^* denotes the transitive closure of G.

In order to compare the structure of two texts t_1, t_2, strictly speaking, we must compare the structure of the graphs $G_L(t_1)$ and $G_L(t_2)$. However, as we will see, this level of detail is not necessary when dealing with human written texts. In these cases we will compare instead of $G_L(t_1)$ and $G_L(t_2)$ just the degrees of the nodes of the graphs. As the two texts t_1 and t_2 may have different length (number of positions) their number of nodes may be different and therefore their degree functions may have different ranges, which can be a drawback for a direct comparison. For example, a Russian text, coded with KOI8-r (one byte per letter), will use approximately the half of the bytes used by coding the same text using UTF-8 (2 bytes per letter).

To overcome situations like this, we choose an arbitrary but fixed number of bins $B \ll N$ and unite several text positions of the text (vertices of G_L) in one vertex.

More exactly, we define the function $\deg_{L,B}(t; k)$ as the number of edges in $G_L(t)$ of the vertices p with $k = \lfloor Bp/(N+1) \rfloor$. We will omit the parameter k if we refer it as a vector index. Of course, we have to choose the same B for the objects being compared. In this way, the two binned degree functions of both Russian versions will be very similar.

The scale parameter B can be chosen in a way to achieve enough statistics for the estimation of the density of each bin k which in practical terms means to have some 10 edges per bin [8].

Once we have $\deg_{L,B}(t_1)$ and $\deg_{L,B}(t_2)$ in the same, unidimensional range, $[0, B-1]$, we can compute the distance between these two values in many different ways. However, for this study even a simple correlation ρ of the smooth averages of these these functions $\rho(\deg_{L,B}(t_1), \deg_{L,B}(t_2))$ serves in order to demonstrate the proof of concept. Thus we have a similarity measure:

$$M_1(t_1, t_2; L, B) \equiv \text{corr}(\deg_{L,B}(t_1), \deg_{L,B}(t_2)), \qquad (2)$$

where with $\rho(x, y)$ is by the definition the correlation between x and y:

$$\text{corr}(x, y) \equiv (\overline{xy} - \overline{x}\,\overline{y}) / \sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}.$$

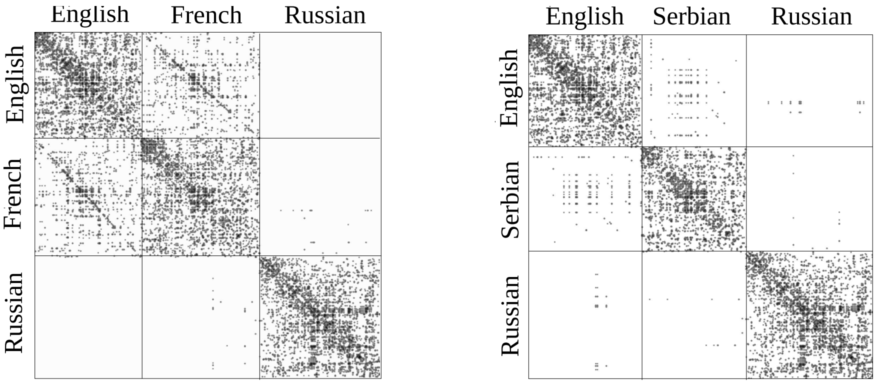

Figure 1: The connectivity of the graph G_L for two groups of three languages each, using UD as text. The left panel represents the concatenation of the English, French and Russian (KOI8-r coding) version of the UD, in that order, that gives significant cross compression between French and English. The right panel represents the connectivity matrix of English, Serbian (Latin) and Russian version. The cross-correlation between the English and the Serbian version is larger, but actually the Serb language is much closer to the Russian than to English.

III. Experimental Results

As a testbed, we use the "Universal Human Rights Declaration" (UD) [9] and the first two chapters of four translations of "Don Quixote" (DQ) [10, 11, 12, 13].

These data have a suitable size of about 10 Kbytes per text. We are trying to see whether we can detect the similarity in the structure and eliminate the influence of the grammar and the formatting of the document.

In all cases the parameter L was set to 5 letters, which seams to eliminate many speech particles that carry essentially grammatical and morphological information, such as particles that determine the gender, definitive and indefinite particles, etc.

Fig. 1(left) represents a typical result of compression of the concatenation of the texts of the UD in three languages – English, French and Russian. The grey levels represent the presence of a link between the nodes that are placed in the axes of the graph. The graph is blurred, converting each point in a Gaussian bump with radius of about 40 text positions. The total length of the concatenated text is about 30K, so each third is represented by a square-like area. The upper left quadrant represents the result of compression of UD English, compressed with UD English, the quadrant at the center represents the UD of French, compressed with itself and the lower right quadrant represents the Russian version, compressed with itself. Two of the concatenated texts use one and the same alphabet and belong to similar language groups. Therefore, we can observe significant cross-compression pattern in the middle quadrant of the first raw (and the second quadrant of the first column, which is the same). However, we can not see any cross-compression between the Russian version of the UD and the two other versions, due to the non-intersecting codes of the Cyrillic and Latin alphabets, with the exception of the formatting information (white spaces, capture enumeration and similar).

Using the similarity measure (1), we observe a small distance between the English and French versions, but large distances between Russian and both of them that seems a reasonable result.

However, on the right panel of the same figure, Fig. 1, we see the UD text compression of the concatenation of English, Serbian (Latin alphabet) and Russian. We see that the dissimilarity between the English and the Serbian version, using the measure (1) is smaller (0.9466) than the dissimilarity between the Serbian and the Russian version (0.9944), which can be observed also by the density of the middle quadrant of the first row. But in reality, the Russian and the Serbian belong to one and the same language group and ought to group closer.

The compression in either cases (English - French - Russian) and (English - Serbian - Russian) is dominated by the compression within the same text, e.g. the structure and the vocabulary of each language are predominant factors in the compression. The connectivity matrices of $G_{L,B}$ have a typical block-diagonal structure. Therefore, we can try to compare the similarity of the texts, using the unidimensional measure M_1, Eq.(2). Fig.2 (the two bottom panels) represents the smoothed degrees of each graph e.g. $\deg_{L,B}(\mathrm{UD}_{\mathrm{French}})$ and $\deg_{L,B}(\mathrm{UD}_{\mathrm{Russian}})$. The correlation coefficient is rather large, 52.3%. In all cases of the human rights declaration, written in different languages, we can observe similar values (with the exception of very similar languages).

The results for several languages and two different texts are represented in Table 1. For the UD the correlation vary from 51.6% to 66.8%.

In order to see that what is captured is the structure of the text, and not the particular language coding, we can contrast the results of UD with another text.

		DQ				UD					
		eng	fra	rus	spa	eng	fra	rus	spa	sre	sry
DQ	eng	100	31	18	34	5	2	-10	4	-6	-3
	fra	31	100	24	51	-1	-6	-10	-17	-9	-5
	rus	18	25	100	17	-16	4	-6	5	6	-2
	spa	34	51	17	100	6	-4	-5	-13	-10	-9
UD	eng	5	-1	-16	6	100	58	52	34	56	56
	fra	2	-6	4	-4	58	100	52	56	64	50
	rus	-10	-10	-6	-5	52	52	100	27	58	42
	spa	4	-17	5	-13	34	56	27	100	50	34
	sre	-6	-9	6	-10	56	64	58	50	100	67
	sry	-3	-5	-2	-9	56	50	42	34	67	100

Table 1: The similarity measure M_1 in percents, between the texts of the universal declaration of human rights and "Don Quixote" in different languages (eng – English, fra – French, rus – Russian, sre – Serbian/Latin, sry – Serbian/Cyrillic/UNICODE). The difference between the Serbian versions is due to the two or one byte coding of the Cyrillic alphabet in "sry".

We choose as a different text the translations of the first chapter of "Don Quixote", by Miguel de Cervantes and we calculate the same similarity measure M_1 for English, French, Russian and Spanish versions. The results are represented in Table 1

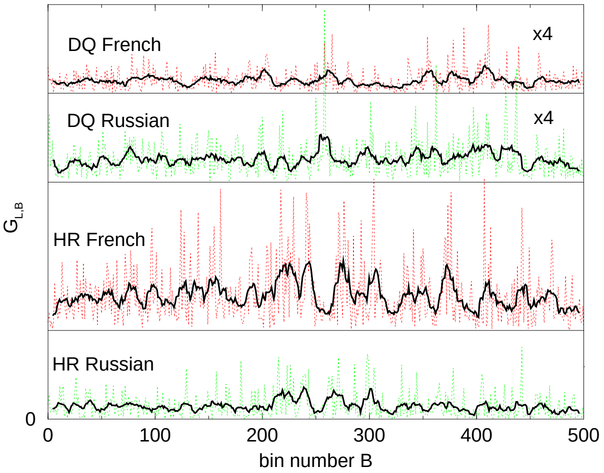

Figure 2: The degrees of the nodes of the graphs $G_{L,B}$ of UD.

All pairings between the same text and different languages has $M_1 > 0.17$. In contrast, all pairs between different texts, including different texts in the same language, give $M_1 \in [-0.2, 0.06]$, which corresponds to random match. The smoothed functions of the French and the Russian version of DQ are represented in the two upper panels of Fig.2. There exists a clear distinction between DQ and UD texts and a similarity between the Russian and the French version of one and the same text (either DQ or UD).

Because the UD and DQ texts are very different as a style, one can argue that what is actually captured is the specific dictionary of each text and not the structure of the text itself. In order to discard such a possibility, we also compare two chapters of DQ with similar lengths. The results are represented in Table 2, where it can be clearly seen that the distinction is between different texts and not between different writing styles, dictionaries or authors.

		C1				C2			
		eng	fra	rus	spa	eng	fra	rus	spa
C1	eng	100	14	10	13	-5	-5	-1	-3
	fra	14	100	16	34	-4	4	2	-4
	rus	10	16	100	10	-13	0	-2	3
	spa	13	34	10	100	-4	-1	-2	-2
C2	eng	-5	-4	-13	-4	100	24	15	24
	fra	-5	4	0	-1	24	100	25	32
	rus	-1	2	-2	-2	15	25	100	20
	spa	-3	-4	3	-2	24	32	20	100

Table 2: The similarity measure M_1 in percents, between two chapters of DQ. The notation of the languages is the same as in the previous table.

As a conclusion we can see that the structure of the text is captured by the similarity measure M_1, Eq. (2) and the measure can effectively detect human-like text translations.

IV. Summary

As a conclusion, the experimental data confirm the hypothesis that by using LZ inspired structures, we can detect similarities in the texts even if the language is different. The results are good for any languages and texts larger than 10 KB letters. The measure is insensitive to the language and the writing style, even if the author and the style are the same. The only important characteristic is the structure of the content.

Acknowledgments

We would like to thank Manuel Alfonseca and Alberto Suárez for the thoughtful discussion and comments on the draft of the article.

References

[1] M. Li, et al, "The similarity metric", *IEEE Trans. Information Theory*, 50:12(2004), 3250-3264.

[2] R. Cilibrasi, P.M.B. Vitanyi, "Clustering by compression", *IEEE Trans. Information Theory*, 51:4(2005), 1523-1545.

[3] M. Cebrian, M. Alfonseca, A. Ortega, "Common pitfalls using the normalized compression distance: what to watch out for in a compressor", submitted to *Journal of Communications in Information and Systems*, 2005.

[4] J. Ziv, A. Lempel, "A universal algorithm for sequential data compression", *IEEE Trans. Information Theory*, vol. IT-24, pp. 337-343, May 1977.

[5] N. Abramson, "Information Theory and Coding", McGraw-Hill, New York, 1963.

[6] M. Burrows, D. J. Wheeler, "A block sorting lossless data compression algorithm", Tech. Rep. 124. *Digital Equipment Corporation*, Palo Alto, Calif., 1994.

[7] J. Cleary, I. Witten, A. C. Calgary, "Data compression using adaptive coding and partial matching", *IEEE Trans. Communications*, 32, 396-402, 1984.

[8] L. Paninski, "Estimation of entropy and mutual information", *Neural Computation*, 15 (6), 1191-1253, 2003.

[9] "The Universal Declaration of Human Rights" in different languages http://www.unhchr.ch/udhr/index.htm

[10] Miguel de Cervantes, "Don Quijote", electronic version: http://cvc.cervantes.es/obref/quijote/edicion/parte1/parte01/cap01/default.htm

[11] Miguel de Cervantes, "Don Quixote", translated by J. Ormsby, electronic version: http://www.gutenberg.org/etext/996

[12] Miguel de Cervantes, "Don Quichotte", translated by L. Viardot, electronic version: http://www.gutenberg.org/files/16066/16066-8.txt

[13] Miguel de Cervantes, "Don Kihot", translated by N. Ljubimov, Hudozhestvenaya literatura, Moskva, 1988, electronic version: http://lib.ttknet.ru/koi/INOOLD/SERVANTES/donkihot1.txt