



S. V. Znamenskij

Stable assessment of the quality of similarity algorithms of character strings and their normalizations

ABSTRACT. The choice of search tools for hidden commonality in the data of a new nature requires stable and reproducible comparative assessments of the quality of abstract algorithms for the proximity of symbol strings. Conventional estimates based on artificially generated or manually labeled tests vary significantly, rather evaluating the method of this artificial generation with respect to similarity algorithms, and estimates based on user data cannot be accurately reproduced.

A simple, transparent, objective and reproducible numerical quality assessment of a string metric. Parallel texts of book translations in different languages are used. The quality of a measure is estimated by the percentage of errors in possible different tries of determining the translation of a given paragraph among two paragraphs of a book in another language, one of which is actually a translation. The stability of assessments is verified by independence from the choice of a book and a pair of languages.

The numerical experiment steadily ranked by quality algorithms for abstract character string comparisons and showed a strong dependence on the choice of normalization.

Key words and phrases: string similarity, data analysis, similarity metric, distance metric, numeric evaluation, quality assessment.

2010 *Mathematics Subject Classification:* 97P20; 91C05, 91C20

Introduction

The task of comparing character strings arises when processing large data of a new, uncharted nature. Methods that routinely use syntax and

semantics stop working. General algorithms for the similarity of symbolic sequences are tried and adapted based on new knowledge of the applied area. So it is important to understand the effectiveness of well-known general algorithms and techniques for their application in comparison with each other.

Comparison of models and algorithms used for highlighting requires arrays of similar strings of various origins [1], which are usually comes from either unpublished personal data arrays [2–5], or from hand-marked linguistic corps or thesauri, as in [6], or from artificially generated data [7]. The public unavailability of some excludes the reproducibility of experiments and an independent assessment of the quality of the initial data, while the high labor-consuming nature of others also limits their volume and availability. The inaccessibility, small volume and unclear origin of the initial data deprive the experiments of persuasiveness.

There exists remarkable ability to freely use parallel texts in different languages for the evaluation of the quality of proximity metrics that were kindly selected and provided to researchers on the site http://www.farkastranslations.com/bilingual_books.php by Hungarian programmer and translator Andras Farkas.

1. Purpose and rating scale, data sources

How does the model, algorithm and metric normalization affect the efficiency of an abstract (not using the specific alphabet, language and data) metrics (or similarity measures) of character strings? In searching for a transparent answer to this question, one can confine to well-known algorithms with widely used executable well-debugged executable code and with a clearly described model that does not require an empirical selection of parameters.

Usually for evaluations use (for example, in figures 3–6 in [8]) the completeness and quality of search results, monotonously connected through the organization of queries. However, the scalar characteristic is more convenient than the vector of two dependent characteristics. A simple and clear scalar measure of the (in) efficiency of the proximity metric is *percentage of mistakenly selected translations* defined as the average proportion of translation fragments that are closer to the metric under test than the correct translation fragment.

TABLE 1. Parallel texts used

Author	Title and Languages	Number of Paragraphs	Paragraph Size
Edgar Po	Escher House Fall (en, hu, es, it, fr, de, eo)	7×269	158 ± 211
Mark Twain	Tom Sawyer (en, de, hu, nl, ca)	$5 \times 414\ 0$	102 ± 135
Lewis Carroll	Alice in Wonderland (en, hu, es, it, pt, fr, de, eo, fi)	9×805	174 ± 245

For it, the inequality $0 \leq E_s(\mu) \leq 100$ is true, the ideal value is 0, and the value 50 means a result equivalent to random guessing, and $E_s(\mu) > 50$ indicates an inadequate metric.

For the study were taken three described in Table 1 books in English (en), Hungarian hu, Spanish (es), Italian (it), Catalan (ca), German (de), Portuguese (pt), Finnish (fi), French (fr) and Esperanto (eo).

2. Compared metrics

Well-known metrics included in the widely used R stringdist package participated in the tests. For clarity of discussion of the results, we briefly recall the compared metrics.

lcs(x, y) — the total number of deletions and inserts at the shortest transition from one substring to another. Is the metric normalization of the length of the $\text{LCS}(x, y)$ of the longest common subsequence using the formula $\text{lcs}(x, y) = l(x) + l(y) - 2\text{LCS}(x, y)$, where l is the length of the string.

lv (x, y) is the classical Levenshtein metric that counts the total number of replacements, deletions, and inserts when moving from one substring to another,

dl (x, y) is the Levenshtein–Damero metric, additionally counting unit permutations.

osa (x, y) (*Optimal string alignment*) is a variation of the Levenshtein–Damero metric that allows multiple permutations.

jw (x, y) (Jaro metric) is not a metric in the strict mathematical sense of the distance between lines, more sophisticated taking into account the transposition, coincidence and position of characters.

jwp (x, y) (Jaro-Winkler metric) — Winkler's Jaro metric correction with the deforming correction parameter $p = 0.1$.

qgram1 (x, y) is the number of different characters including repetitions, that is, the sum of all the letters $s_i \in \{s_1, \dots, s_n\}$ of the expression alphabet $|X_i - Y_i|$ where \vec{X} and \vec{Y} are the vector of the numbers of occurrences of all characters of the alphabet in each of the compared lines.

cosine1 (x, y) is calculated using the formula $1 - \frac{(\vec{X}, \vec{Y})}{\|\vec{X}\| \|\vec{Y}\|}$.

qgram2 (x, y) is the number of different diagrams (pairwise combinations) of characters, taking into account repetitions.

cosine2 (x, y) is calculated by the same **cosine1** formula for digrams.

qgram3 (x, y) is the number of different trigrams (triple combinations) of characters, taking into account repetitions.

cosine3 (x, y) is calculated using a similar formula for trigrams.

A detailed description of these metrics is provided in [9] with links to sources.

Additionally, the experimentally selected normalization of NCS/OCS similarity metrics, promoted by the author as a more effective alternative to LCS, proposed and investigated in [10–12], were considered. Briefly repeating, NCS is the maximum possible number of different common substrings in a common subsequence of symbols, which is bounded by a value $\psi(n) = \frac{n(n+1)}{2}$ for a string and its substring of length n , and $\text{OCS}(x, y) = \psi^{-1}(\text{NCS}(x, y)) = \frac{\sqrt{8 \text{NCS}(x, y) - 1} + 1}{2}$ is LCS-like normalization of NCS. The similarity metrics are directed opposite to the distance metrics [13, 14] and use differently defined normalization of distance metrics as distance metrics. During the experiments, simple and efficient functions were distinguished for using these similarity metrics as distance metrics to determine the order of the pairs:

$$\text{NCS1}(x, y) = \frac{l(x) + l(y) - 3 \text{NCS}(x, y)}{l(x)l(y)},$$

$$\text{NCS2}(x, y) = \frac{1 - \text{NCS}(x, y)}{l(x) + l(y)},$$

$$\begin{aligned} \text{OCS1}(x, y) &= l(x) + l(y) - 2 \text{OCS}(x, y), \\ \text{OCS2}(x, y) &= \frac{l(x) + l(y) - 2 \text{OCS}(x, y)}{\sqrt{l(x)l(y)}}. \end{aligned}$$

Prepared for comparison graphs also present the lengths difference $\text{LENGTH}(x, y) = |l(x) - l(y)|$ as a simple distance function and the average of all metrics **AVERAGE**. Like the stringdist packet metrics, all of these functions except **OCS1** are not metrics in the strict sense of the word, but with a little complication (the construction from the clause Basic definitions in [15]) can be replaced by metrics in the strict sense defining the same order relation on pairs.

For calculations, in addition to the stringdist metrics in question, we used C code, published in [16] and launched from Perl XS. For basic processing, a Perl script was used. Archive with scripts and main results of processing is attached to the article.

3. Setting and the result of the first experiment

Since not all metric calculation procedures support utf8, transliteration of the diacritics was required. For this purpose, the packages `Text :: Unaccent` and `Text :: Unidecode` were used in the procedure `sub{unac_string('utf8', unidecode(1c $_[0]))}` after which all non-ascii characters were removed from the lines.

Script to get information about languages on behalf of the user. The calculated values are recorded in a separate file with labels and languages. Immediate archiving of Bzip2 is about three times (up to 14 GB) reduced the amount of recorded information about metrics. Used books have less than 3% of available texts. Processing more is suppressed by the quadratic computational complexity of the problem. In particular, a distance matrix can not be calculated at all on a 64-bit computer for the “Three Musketeers” book.

In the event of a computer freezing or an unintended power outage (calculating metrics on a PC with a four core processor and 16 Gb of RAM required several days), such an organization allowed the calculations to continue from the time the archive was last recorded. Reuse of calculated

TABLE 2. Values of errors of metrics in the group (1) ($\{de, en\}$, $\{es, fr\}$, $\{es, it\}$, $\{fr, it\}$)

metric	Fall	Tom	Alice	total
OCS2	$1.6\% \pm 1.7\%$	$4.4\% \pm 0.9\%$	$4.1\% \pm 0.7\%$	$3.1\% \pm 1.8\%$
NCS2	$2.2\% \pm 2.6\%$	$4.6\% \pm 0.2\%$	$4.1\% \pm 0.6\%$	$3.3\% \pm 2.0\%$
NCS1	$4.5\% \pm 5.6\%$	$8.1\% \pm 0.9\%$	$7.0\% \pm 1.5\%$	$6.0\% \pm 4.1\%$
qgram1	$4.9\% \pm 2.9\%$	$9.8\% \pm 2.5\%$	$8.9\% \pm 2.1\%$	$7.2\% \pm 3.3\%$
jwp	$5.6\% \pm 3.4\%$	$7.4\% \pm 0.3\%$	$9.3\% \pm 1.3\%$	$7.4\% \pm 3.0\%$
jw	$5.6\% \pm 3.7\%$	$8.4\% \pm 0.9\%$	$9.1\% \pm 1.4\%$	$7.5\% \pm 3.2\%$
LENGTH	$6.8\% \pm 1.2\%$	$11.9\% \pm 0.9\%$	$11.2\% \pm 1.4\%$	$9.3\% \pm 2.6\%$
dl	$6.4\% \pm 8.1\%$	$17.1\% \pm 7.9\%$	$13.3\% \pm 6.5\%$	$10.7\% \pm 8.4\%$
osa	$6.5\% \pm 8.1\%$	$17.2\% \pm 7.9\%$	$13.3\% \pm 6.6\%$	$10.7\% \pm 8.4\%$
lv	$6.5\% \pm 8.2\%$	$17.3\% \pm 8.0\%$	$13.5\% \pm 6.6\%$	$10.8\% \pm 8.5\%$
cosine3	$10.3\% \pm 10.2\%$	$17.3\% \pm 0.7\%$	$17.3\% \pm 4.4\%$	$14.2\% \pm 8.2\%$
AVERAGE	$13.8\% \pm 6.7\%$	$21.8\% \pm 1.5\%$	$19.8\% \pm 2.7\%$	$17.4\% \pm 5.9\%$
cosine2	$16.6\% \pm 10.4\%$	$21.3\% \pm 1.1\%$	$24.2\% \pm 4.7\%$	$20.5\% \pm 8.4\%$
cosine1	$25.7\% \pm 7.3\%$	$29.5\% \pm 2.0\%$	$33.1\% \pm 5.5\%$	$29.4\% \pm 7.0\%$
qgram2	$20.0\% \pm 15.6\%$	$44.1\% \pm 1.3\%$	$31.1\% \pm 10.0\%$	$27.6\% \pm 14.6\%$
lcs	$18.8\% \pm 16.1\%$	$41.8\% \pm 2.2\%$	$36.4\% \pm 5.8\%$	$29.2\% \pm 14.8\%$
qgram3	$38.7\% \pm 6.0\%$	$49.4\% \pm 0.2\%$	$44.6\% \pm 2.4\%$	$42.5\% \pm 5.7\%$
OCS1	$47.2\% \pm 1.8\%$	$51.3\% \pm 0.1\%$	$48.0\% \pm 0.8\%$	$48.0\% \pm 1.8\%$

metric values saved time for experiments on the selection of suitable normalization of NCS and OCS metrics.

The processing of each translation consisted in calculating the error of the metric

$$(1) \quad E(m) = \frac{\sum_{x \in X} |\{y \in Y : m(x, y) < m(x, y_x)\}|}{|X| \cdot |Y|} \cdot 100\%,$$

where X and Y are the set of parallel text paragraphs in two different languages, $|X|$ and $|Y|$ are the powers of these sets, m — the metric under test, and y_x — the translation of the paragraph x in the set Y .

The pairs of common languages of books were divided into four groups according to the proximity of transliterated paragraphs:

- (1) most close $\{de, en\}$, $\{es, fr\}$, $\{es, it\}$, $\{fr, it\}$;
- (2) relatively close $\{en, eo\}$, $\{en, es\}$, $\{en, fr\}$, $\{en, it\}$, $\{eo, es\}$, $\{eo, it\}$;
- (3) relatively far $\{de, es\}$, $\{de, eo\}$, $\{de, fr\}$, $\{de, it\}$, $\{es, hu\}$, $\{hu, it\}$;
- (4) most far $\{de, hu\}$, $\{en, hu\}$, $\{eo, hu\}$, $\{fr, hu\}$.

TABLE 3. Values of errors of metrics in the group (2) ($\{en, eo\}$, $\{en, es\}$, $\{en, fr\}$, $\{en, it\}$, $\{eo, es\}$, $\{eo, it\}$)

metric	Fall	Alice	total
OCS2	$1.6\% \pm 0.8\%$	$5.8\% \pm 1.1\%$	$3.7\% \pm 2.3\%$
NCS2	$2.4\% \pm 0.8\%$	$6.7\% \pm 0.9\%$	$4.6\% \pm 2.4\%$
LENGTH	$7.3\% \pm 1.4\%$	$11.1\% \pm 1.4\%$	$9.2\% \pm 2.4\%$
NCS1	$5.2\% \pm 1.7\%$	$12.5\% \pm 2.3\%$	$8.8\% \pm 4.2\%$
qgram1	$7.4\% \pm 1.8\%$	$11.9\% \pm 2.5\%$	$9.6\% \pm 3.1\%$
jw	$8.7\% \pm 2.0\%$	$12.4\% \pm 1.1\%$	$10.5\% \pm 2.5\%$
jwp	$9.0\% \pm 2.0\%$	$12.4\% \pm 1.2\%$	$10.7\% \pm 2.4\%$
dl	$11.1\% \pm 6.1\%$	$19.7\% \pm 6.3\%$	$15.4\% \pm 7.6\%$
osa	$11.1\% \pm 6.1\%$	$19.8\% \pm 6.3\%$	$15.5\% \pm 7.6\%$
lv	$11.3\% \pm 6.1\%$	$20.0\% \pm 6.3\%$	$15.6\% \pm 7.6\%$
cosine3	$12.3\% \pm 2.9\%$	$21.9\% \pm 2.2\%$	$17.1\% \pm 5.4\%$
AVERAGE	$19.0\% \pm 1.7\%$	$24.8\% \pm 1.5\%$	$21.9\% \pm 3.3\%$
cosine2	$22.4\% \pm 2.8\%$	$31.7\% \pm 1.5\%$	$27.0\% \pm 5.2\%$
cosine1	$32.8\% \pm 1.9\%$	$39.8\% \pm 1.8\%$	$36.3\% \pm 3.9\%$
qgram2	$36.5\% \pm 5.8\%$	$42.0\% \pm 4.5\%$	$39.2\% \pm 5.9\%$
lcs	$39.4\% \pm 2.4\%$	$44.6\% \pm 1.3\%$	$42.0\% \pm 3.3\%$
qgram3	$44.3\% \pm 1.6\%$	$47.0\% \pm 0.8\%$	$45.7\% \pm 1.8\%$
OCS1	$48.5\% \pm 0.6\%$	$48.9\% \pm 0.5\%$	$48.7\% \pm 0.6\%$

The results of the experiment showed in the tables 2, 3, 4 and 5 high stability of the ranking of metrics by quality, almost independent either of the book, or of a particular pair of languages in the group. The results are graphically presented in Figure 1; the percentage of error is plotted vertically, pairs of languages are ordered to the right in descending order of the average error.

The graphs show that the sharply increased spread of metrics *dl*, *lv*, *osa* is closely related to the significant influence of the order of languages in a pair and the difference in paragraph lengths.

Surprising that the ranking of metrics by quality looks almost unrelated to the complexity of the algorithms: The simplest algorithm that calculates the difference in paragraph lengths turned out to be one of the best. This confirms the hypothesis of the exceptional importance of the correct choice of the normalization of the metric.

TABLE 4. Values of errors of metrics in the group (3) ({de, es}, {de, eo}, {de, fr}, {de, it}, {es, hu}, {hu, it})

metric	Fall	Alice	total
OCS2	7.1% ± 1.1%	9.4% ± 2.2%	8.2% ± 2.1%
LENGTH	9.2% ± 1.0%	12.1% ± 2.4%	10.6% ± 2.4%
NCS2	12.2% ± 1.5%	12.0% ± 1.8%	12.1% ± 1.7%
qgram1	12.6% ± 3.6%	14.9% ± 3.9%	13.7% ± 3.9%
jw	16.0% ± 2.6%	16.5% ± 2.8%	16.2% ± 2.7%
jwp	16.4% ± 2.5%	16.4% ± 2.8%	16.4% ± 2.7%
NCS1	22.7% ± 3.0%	21.0% ± 2.7%	21.9% ± 3.0%
dl	24.6% ± 7.1%	25.5% ± 6.7%	25.1% ± 6.9%
osa	24.7% ± 7.1%	25.6% ± 6.7%	25.2% ± 6.9%
lv	24.9% ± 7.1%	25.8% ± 6.7%	25.3% ± 6.9%
AVERAGE	29.6% ± 1.4%	29.6% ± 1.8%	29.6% ± 1.6%
cosine3	35.1% ± 1.0%	32.6% ± 1.8%	33.9% ± 1.9%
cosine2	39.6% ± 1.0%	38.4% ± 2.2%	39.0% ± 1.8%
cosine1	41.7% ± 1.4%	43.7% ± 2.4%	42.7% ± 2.2%
qgram2	48.2% ± 0.6%	46.9% ± 0.9%	47.5% ± 1.0%
lcs	48.4% ± 0.7%	47.5% ± 0.6%	47.9% ± 0.8%
qgram3	49.8% ± 0.4%	48.7% ± 0.5%	49.2% ± 0.7%
OCS1	50.6% ± 0.3%	49.3% ± 0.4%	50.0% ± 0.8%

TABLE 5. Values of errors of metrics in the group (4) ({de, hu}, {en, hu}, {eo, hu}, {fr, hu})

metric	Fall	Tom	Alice	total
OCS2	7.2% ± 1.8%	11.5% ± 0.8%	12.8% ± 1.7%	10.3% ± 3.0%
LENGTH	8.7% ± 2.2%	14.7% ± 1.2%	15.2% ± 2.0%	12.5% ± 3.6%
NCS2	13.6% ± 1.7%	17.5% ± 0.7%	15.6% ± 1.2%	15.2% ± 2.0%
qgram1	14.0% ± 5.2%	19.8% ± 2.8%	18.9% ± 5.1%	17.1% ± 5.4%
jw	18.5% ± 3.2%	21.0% ± 0.5%	20.8% ± 1.8%	19.9% ± 2.6%
jwp	19.3% ± 3.1%	21.2% ± 0.3%	20.9% ± 1.7%	20.3% ± 2.4%
NCS1	25.9% ± 3.5%	26.1% ± 0.9%	26.2% ± 2.4%	26.0% ± 2.7%
dl	26.0% ± 9.3%	29.5% ± 3.9%	28.4% ± 8.4%	27.6% ± 8.2%
osa	26.0% ± 9.3%	29.6% ± 3.9%	28.4% ± 8.4%	27.7% ± 8.2%
lv	26.2% ± 9.3%	29.7% ± 3.9%	28.5% ± 8.4%	27.8% ± 8.2%
AVERAGE	30.9% ± 1.6%	32.5% ± 0.7%	32.3% ± 1.7%	31.8% ± 1.7%
cosine3	35.9% ± 1.0%	31.0% ± 0.5%	35.5% ± 1.3%	34.8% ± 2.2%
cosine2	40.0% ± 1.3%	36.5% ± 0.6%	41.3% ± 0.6%	39.8% ± 2.0%
cosine1	42.1% ± 1.6%	40.9% ± 0.7%	45.8% ± 0.9%	43.4% ± 2.4%
qgram2	48.6% ± 0.7%	50.3% ± 0.5%	47.7% ± 0.7%	48.6% ± 1.2%
lcs	49.2% ± 0.5%	50.1% ± 0.5%	48.2% ± 0.6%	49.0% ± 0.9%
qgram3	50.3% ± 0.4%	51.7% ± 0.4%	48.8% ± 0.5%	50.0% ± 1.2%
OCS1	51.0% ± 0.2%	52.5% ± 0.4%	49.2% ± 0.5%	50.6% ± 1.3%

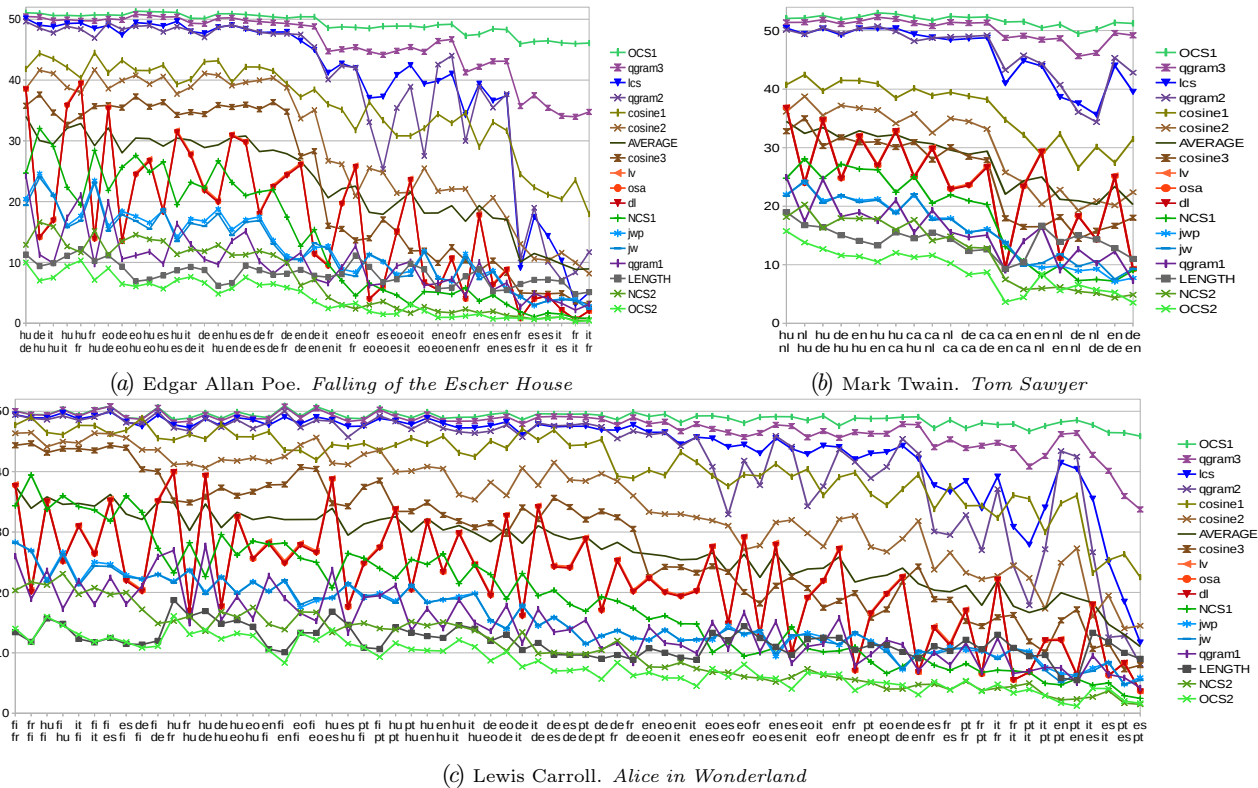


FIGURE 1. Percentage of a binary choice of correct paragraph translation in a multilingual book

TABLE 6. Error of metrics with equal lengths of arguments in a group of language pairs({de, en}, {es, fr}, {es, it}, {fr, it})

metric	Fall	Tom	Alice	total
lcs	7.4% ± 10.0%	8.9% ± 0.2%	8.6% ± 2.3%	8.1% ± 6.9%
NCS1, NCS2, OCS1, OCS2	8.4% ± 10.3%	8.6% ± 0.1%	8.0% ± 2.2%	8.2% ± 7.0%
dl,lv,osa	8.7% ± 9.4%	9.3% ± 0.2%	9.7% ± 2.4%	9.2% ± 6.5%
qgram3	8.5% ± 9.5%	13.1% ± 0.3%	11.3% ± 4.2%	10.3% ± 7.1%
AVERAGE	11.7% ± 10.0%	13.6% ± 0.3%	13.9% ± 3.0%	12.9% ± 7.0%
qgram2	11.5% ± 12.0%	14.4% ± 0.4%	13.3% ± 4.1%	12.6% ± 8.5%
cosine3	10.4% ± 10.5%	17.0% ± 0.4%	14.9% ± 4.7%	13.2% ± 8.1%
jwp	15.7% ± 8.2%	15.0% ± 0.4%	20.8% ± 3.0%	17.9% ± 6.4%
cosine2	14.5% ± 11.6%	20.0% ± 0.4%	19.2% ± 4.6%	17.2% ± 8.7%
jw	16.6% ± 9.0%	17.5% ± 0.4%	20.7% ± 3.4%	18.5% ± 6.7%
qgram1	17.1% ± 10.4%	19.0% ± 0.8%	21.1% ± 3.8%	19.1% ± 7.6%
cosine1	24.6% ± 9.1%	29.1% ± 1.1%	30.5% ± 4.6%	27.8% ± 7.4%

4. Experiment Eliminating the Effect of Normalization

To eliminate the effect of normalization, modify the formula (1) as follows:

(2)
$$E_{=}(m) = \frac{\sum_{x \in X} |\{y \in Y : m(x, y) < m(x, y_x) \& l(y) = l(y_x)\}|}{|X| \cdot |Y|} \cdot 100\%,$$

Rigid selection of arguments of metrics by equality of lengths naturally aligns the scatter of results and dramatically changes the rating. Under these conditions, simple formulae for a normalization are turned off and the quality of complex calculations comes to the fore, see Table 6.

Now also in the graphs on Figure 2, a sharply different order of metrics is clearly visible. In particular, two metrics with the largest errors **lcs** and **qgram3** turn out to be the best after NCS/OCS.

The emergence of a hypothesis about the possibilities of a better selection of normal norms. It is natural to expect that with the optimal choice of rating for the overall situation will be quite long. For example, normalization NLCS [17] of the LCS metric sets close to OCS2 order and can join the leaders.

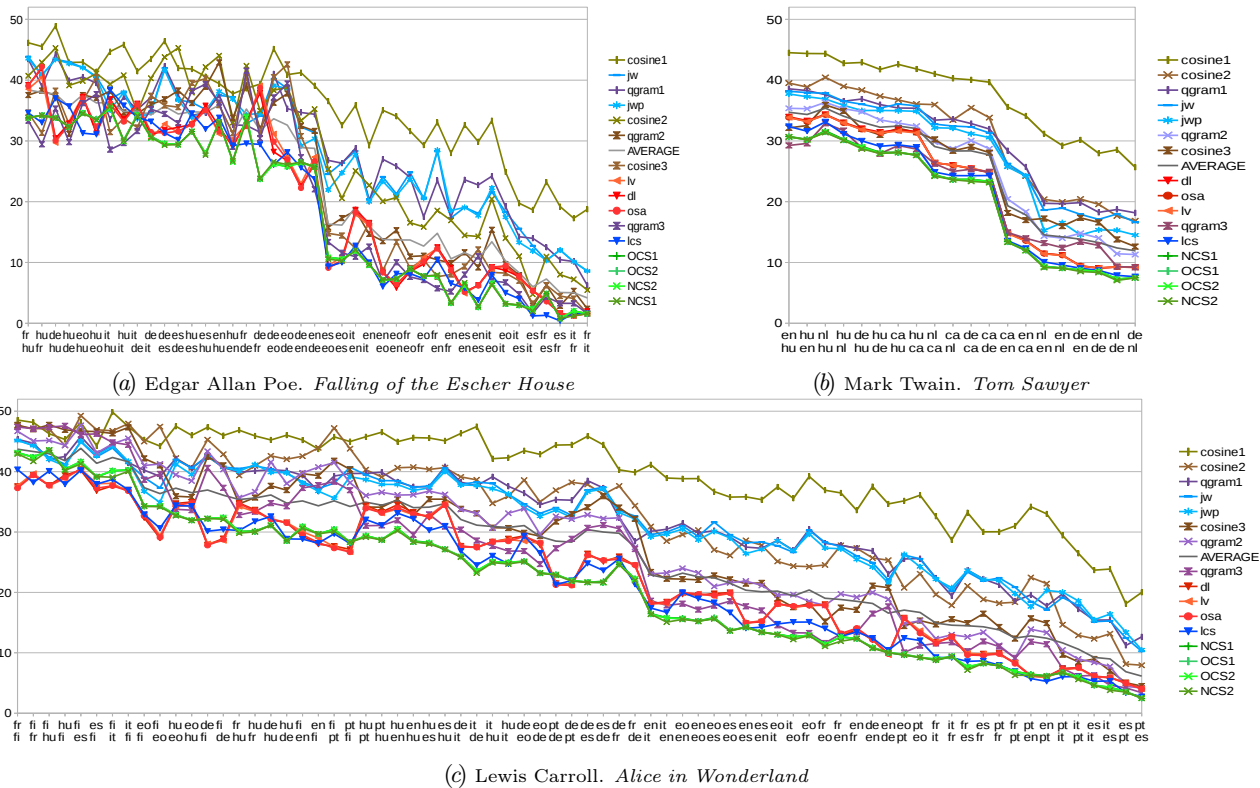


FIGURE 2. Errors of metrics with equal long arguments

TABLE 7. Errors of metrics with equal long arguments in the group (2) of language pairs({en, eo}, {en, es}, {en, fr}, {en, it}, {eo, es}, {eo, it})

metric	Fall	Alice	total
NCS1, NCS2, OCS1, OCS2	7.3% ± 3.0%	14.2% ± 1.4%	10.7% ± 4.2%
lcs	8.0% ± 2.5%	16.0% ± 2.2%	12.0% ± 4.7%
qgram3	9.1% ± 2.7%	16.4% ± 2.1%	12.8% ± 4.4%
dl,lv,osa	10.0% ± 3.9%	17.3% ± 2.4%	13.7% ± 4.9%
cosine3	11.2% ± 3.0%	20.8% ± 2.2%	16.0% ± 5.5%
AVERAGE	13.9% ± 2.6%	21.0% ± 1.6%	17.4% ± 4.2%
qgram2	13.5% ± 3.1%	21.4% ± 1.6%	17.5% ± 4.7%
cosine2	19.4% ± 3.8%	27.8% ± 1.9%	23.6% ± 5.1%
jwp	21.9% ± 3.5%	28.2% ± 1.5%	25.0% ± 4.2%
jw	22.3% ± 3.8%	28.8% ± 1.6%	25.6% ± 4.3%
qgram1	23.8% ± 3.3%	28.9% ± 1.4%	26.3% ± 3.6%
cosine1	32.1% ± 3.3%	37.0% ± 2.0%	34.6% ± 3.7%

TABLE 8. Errors of metrics with equality of long arguments for language pairs (3):({de, es}, {de, eo}, {de, fr}, {de, it}, {es, hu}, {hu, it})

metric	Fall	Alice	total
NCS1, NCS2, OCS1, OCS2	29.8% ± 3.4%	24.3% ± 2.0%	27.1% ± 3.9%
lcs	31.9% ± 3.3%	26.2% ± 2.7%	29.0% ± 4.1%
dl,lv,osa	32.9% ± 3.0%	28.2% ± 2.8%	30.6% ± 3.8%
qgram3	34.9% ± 4.1%	29.0% ± 2.3%	32.0% ± 4.4%
AVERAGE	34.4% ± 1.2%	31.0% ± 1.9%	32.7% ± 2.3%
qgram2	36.6% ± 2.0%	32.7% ± 2.5%	34.6% ± 3.0%
cosine3	36.7% ± 3.8%	32.9% ± 2.2%	34.8% ± 3.7%
jwp	36.3% ± 2.4%	36.0% ± 2.3%	36.1% ± 2.4%
qgram1	36.0% ± 3.1%	36.5% ± 3.2%	36.3% ± 3.2%
jw	36.7% ± 2.6%	36.4% ± 2.3%	36.5% ± 2.5%
cosine2	39.9% ± 3.1%	37.4% ± 2.1%	38.7% ± 2.9%
cosine1	42.5% ± 2.4%	43.8% ± 2.3%	43.2% ± 2.5%

5. Other Comparison Situations

For this purpose, results with other restrictions on the lengths of the metrics attached to the article file (Table 7,Table 8) may be useful. The choice of a suitable metric and normalization obviously should focus on the features of a specific task.

For example, Figure 3 presents graphs for 10% - restrictions on the difference of lengths.

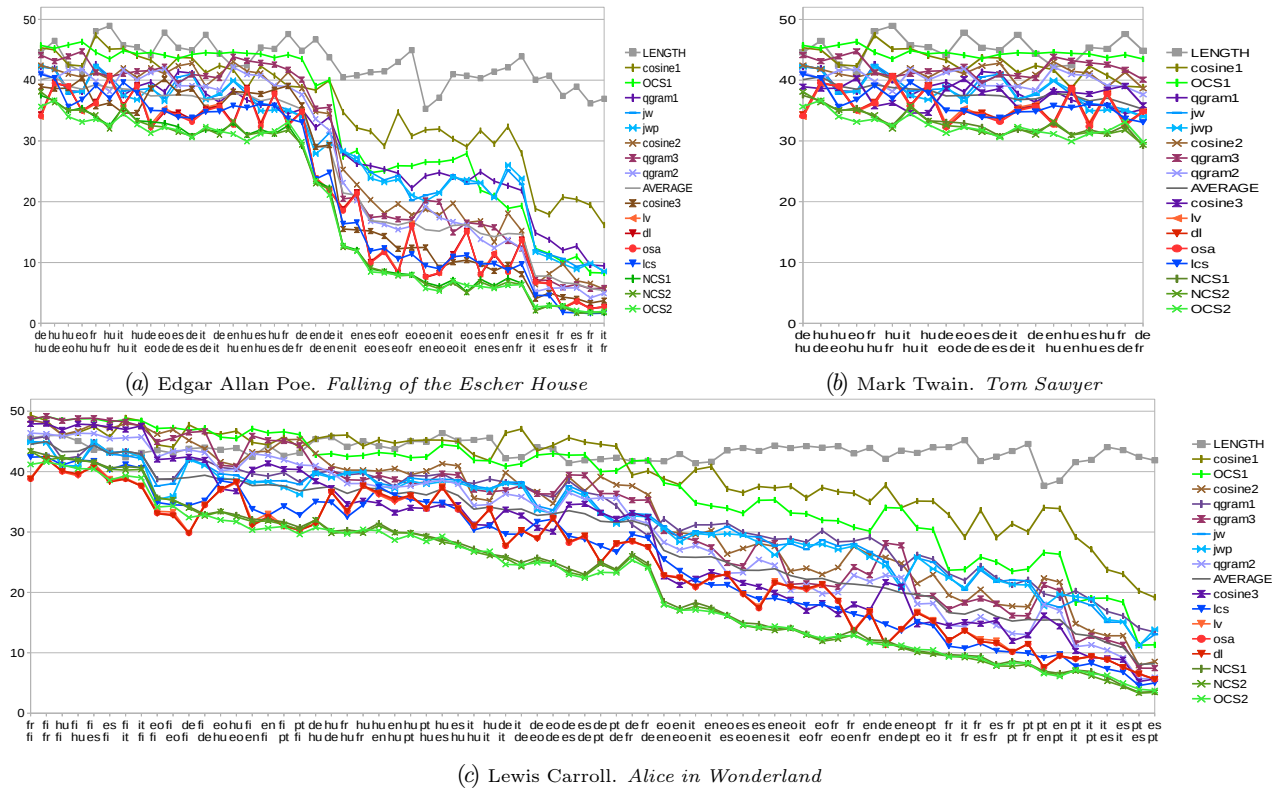
FIGURE 3. Errors of metrics with arguments lengths differ by $\leq 10\%$

TABLE 9. Values of errors of metrics in the group (4) of the languages pairs ($\{de, hu\}$, $\{en, hu\}$, $\{eo, hu\}$, $\{fr, hu\}$)

metric	Fall	Tom	Alice	total
NCS1, NCS2, OCS1, OCS2	32.8% \pm 2.4%	29.5% \pm 1.0%	30.4% \pm 1.4%	31.2% \pm 2.3%
lcs	33.2% \pm 2.4%	30.8% \pm 1.3%	32.2% \pm 1.8%	32.3% \pm 2.2%
qgram3	33.4% \pm 3.3%	28.9% \pm 0.6%	33.0% \pm 1.5%	32.3% \pm 2.9%
dl,lv,osa	34.3% \pm 4.3%	32.6% \pm 1.0%	33.5% \pm 1.1%	33.6% \pm 2.9%
cosine3	35.1% \pm 3.1%	31.9% \pm 0.6%	35.6% \pm 1.4%	34.7% \pm 2.6%
AVERAGE	36.3% \pm 2.2%	33.6% \pm 1.0%	35.6% \pm 0.9%	35.5% \pm 1.9%
qgram2	37.0% \pm 3.2%	34.7% \pm 0.8%	37.8% \pm 1.9%	36.8% \pm 2.6%
qgram1	39.6% \pm 3.4%	37.4% \pm 1.0%	39.9% \pm 1.3%	39.3% \pm 2.5%
jwp	41.0% \pm 2.3%	36.3% \pm 1.3%	39.6% \pm 1.4%	39.5% \pm 2.5%
jw	41.1% \pm 2.4%	36.9% \pm 1.2%	40.0% \pm 1.4%	39.8% \pm 2.4%
cosine2	40.8% \pm 3.5%	38.5% \pm 0.8%	41.2% \pm 1.0%	40.5% \pm 2.5%
cosine1	43.1% \pm 3.4%	43.4% \pm 1.1%	46.0% \pm 0.8%	44.3% \pm 2.7%

Figure 4 shows graphs with a restriction on the length $l(y) \leq l(y_x)$, and on Figure 5 graphs with the opposite restriction $l(y) \geq l(y_x)$. We see a sharply manifested difference in practical problems, by the nature of which the correct choice usually has close to the shortest or close to the greatest length.

Conclusion

Experiments have shown that the effectiveness of the strings similarity metrics critically depends on the matching of the normalization choice of the algorithm to the distribution of the lengths in the data.

Difficult questions became opened:

- How to calculate the most effective formula for the normalization of a given metric from specific data?
- Will the calculated formulas give a significant gain for the metrics considered?
- How to calculate the appropriate normalization of a given metric from data statistics?
- How to estimate the adequacy of the normalization of a given metric by data statistics?

It seems reasonable to continue research in search of answers to these questions.

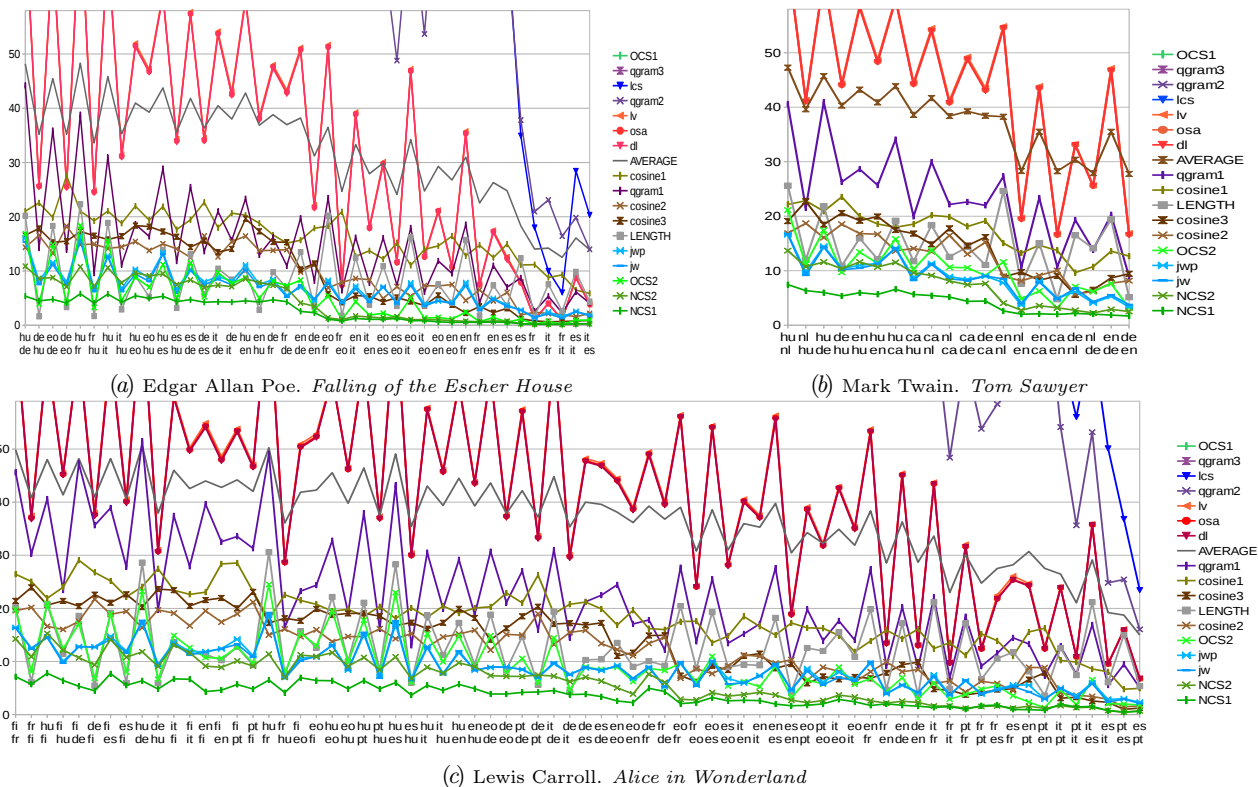
























FIGURE 4. Metric errors when the correct answer is shorter (errors larger 50% are not shown)



FIGURE 5. Metric errors when the correct answer is longer (errors larger 50% are not shown)

References

- [1] W. W. Cohen, P. Ravikumar, S. Fienberg. “A comparison of string distance metrics for name-matching tasks”, *IIWEB'03 Proceedings of the 2003 International Conference on Information Integration on the Web* (August 09–10, 2003, Acapulco, Mexico), 2003, pp. 73–78.  [URL](#) [↑]₅₆₂
- [2] K. Branting. “A comparative evaluation of name-matching algorithms”, *ICAIL '03 Proceedings of the 9th international conference on Artificial intelligence and law* (June 24–28, 2003, Scotland, United Kingdom), 2003, pp. 224–232.  [dbi](#) [↑]₅₆₂
- [3] P. Christen. “A comparison of personal name matching: Techniques and practical issues”, *Proceedings of the Sixth IEEE International Conference on Data Mining — Workshops (ICDMW'06)* (December 18–22, 2006, Hong Kong, China), IEEE, New York, 2006, pp. 290–294.  [dbi](#) [↑]₅₆₂
- [4] G. Recchia, M. Louwerse. “A comparison of string similarity measures for toponym matching”, *COMP '13 Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place* (November 05–08, 2013, Orlando FL, USA), 2013, pp. 54–61.   [URL](#) [dbi](#) [↑]₅₆₂
- [5] N. Gali, R. Marinescu-Istodor, P. Fränti. “Similarity measures for title matching”, 2016 23rd International Conference on Pattern Recognition (ICPR) (December 4–8, 2016, Cancun, México).  [dbi](#) [↑]₅₆₂
- [6] Yufei Sun, Liangli Ma, Shuang Wang. “A comparative evaluation of string similarity metrics for ontology alignment”, *Journal of Information & Computational Science*, **12**:3 (2015), pp. 957–964.   [URL](#) [dbi](#) [↑]₅₆₂
- [7] M. del Pilar Angeles, A. Espino Gamez. “Comparison of methods Hamming Distance, Jaro, and Monge–Elkan”, *DBKDA 2015: The Seventh International Conference on Advances in Databases, Knowledge, and Data Applications* (May 24–29, 2015, Rome, Italy).  [URL](#) [↑]₅₆₂
- [8] C. Varol, C. Bayrak. “Hybrid matching algorithm for personal names”, *ACM Journal of Data and Information Quality*, **3**:4 (2012), 8.  [dbi](#) [↑]₅₆₂
- [9] M. P. J. van der Loo. “The stringdist package for approximate string matching”, *R Journal*, **6**:1 (2014), pp. 111–122.  [URL](#) [↑]₅₆₄
- [10] S. V. Znamenskij. “Simple essential improvements to ROUGE-W algorithm”, *Journal of Siberian Federal University. Mathematics & Physics*, **8**:4 (2015), pp. 258–270.  [dbi](#) [↑]₅₆₄
- [11] S. V. Znamenskij. “A belief framework for similarity evaluation of textual or structured data, similarity search and applications”, *Similarity Search and Applications, SISAP 2015, Lecture Notes in Computer Science*, vol. **9371**, eds. G. Amato, R. Connor, F. Falchi, C. Gennaro, 2015, pp. 138–149.  [dbi](#) [↑]₅₆₄
- [12] S. V. Znamenskij. “A model and algorithm for sequence alignment”, *Program systems: theory and applications*, **6**:1 (2015), pp. 189–197.   [dbi](#) [URL](#) [↑]₅₆₄
- [13] S. V. Znamenskij. “Models and axioms for similarity metrics”, *Program systems: theory and applications*, **8**:4(35) (2017), pp. 349–360 (in Russian).   [dbi](#) [URL](#) [↑]₅₆₄

- [14] S. V. Znamenskij. “From similarity to distance: axiomatic set, monotonic transformations and metric determinacy”, *Journal of Siberian Federal University. Mathematics & Physics*, **11**:3 (2018), pp. 331–341.  [doi](#) [↑]₅₆₄
- [15] M. M. Deza, E. Deza. *Encyclopedia of distances*, Springer-Verlag, Berlin, 2009, 583 pp.  [URL](#)  [doi](#) [↑]₅₆₅
- [16] S. V. Znamenskij, V. A. Dyachenko. “An alternative model of the strings similarity”, DAMDID/RCDL 2017 (Moscow, Russia, October 9–13, 2017), CEUR Workshop Proceedings, vol. **2022**, Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains, eds. L. Kalinichenko, Y. Manolopoulos, N. Skvortsov, V. Sukhomlin, pp. 177–183 (in Russian).  [URL](#) [↑]₅₆₅
- [17] A. Islam, D. Inkpen. “Semantic text similarity using corpus-based word similarity and string similarity”, *ACM Transactions on Knowledge Discovery from Data*, **2**:2 (2008), 10, 25 pp.  [doi](#) [↑]₅₇₀

Received

17.04.2018

Revised

03.12.2018

Published


28.12.2018

Recommended by

dr. Evgeny Kurshev

Sample citation of this publication:

Sergej Znamenskij. “Stable assessment of the quality of similarity algorithms of character strings and their normalizations”. *Program Systems: Theory and Applications*, 2018, **9**:4(39), pp. 561–578.

 [doi](#) 10.25209/2079-3316-2018-9-4-561-578

 [http://psta.psir.ru/read/psta2018_4_561-578.pdf](#)

The same article in Russian:  [doi](#) 10.25209/2079-3316-2018-9-4-579-596

About the author:**Sergej Vital'evich Znamenskij**

Scientific interests migrated from functional analysis and complex analogues of convexity to the foundations of the development of collaborative software, similarity metrics and interpolation theories



0000-0001-8845-7627

e-mail: svz@latex.pereslavl.ru

Эта же статья по-русски:  [doi](#) 10.25209/2079-3316-2018-9-4-579-596