The parallel texts of books translations in the quality evaluation of basic models and algorithms for the similarity of symbol strings

© Sergej V. Znamenskij
Ailamazyan Program Systems Institute of RAS,
Veskovo village, Yaroslavl region, Russia
svz@latex.pereslavl.ru

Abstract. This numeric evaluation of string metric accuracy is based on the following idea: taking the paragraph of text in one language sort all paragraphs of the document in other language by similarity with given paragraph string and consider place of the right translation as the value of the evaluation score.

Such a search of proper translation provides an objective and reproducible quality assessment for known similarity metrics and shows the most sensitive ones.

Keywords: string similarity, data analysis, similarity metric, distance metric, numeric evaluation, quality assessment.

1 Introduction

The evaluation of distance or similarity of symbol strings plays important role in processing of huge data of various nature and requires significant computational resources. Comparison of models and algorithms for such evaluation heavily depends on test sets of similar strings, which can come from different sources [1]. They are usually either private or unpublished data arrays (as in [2-5]), or manually formated linguistic corpuses or thesauri (as in [6]). Public availability of test data sets is necessary for the reproducibility of experiments and for independent assessment of the quality of the initial data. The high labor input as a rule limits their volume and availability. The generated data tests [7] are useful but can't replace the testing of real world data.

Let's exploit the remarkable ability to access parallel texts of the book in different languages and use them for the evaluation of the quality of similarity metrics. Several dozens of such books were kindly selected and provided to researchers on the site http://www.farkastranslations.com/bilingual_books.php by Hungarian programmer and translator Andras Farkas.

2 Aim, objects and measure of quality

How does the model, algorithm, and metric normalization affect the effectiveness of the metric (measure) of similarity on symbol strings? In search of a transparent answer to this question, we can confine ourselves to well-known algorithms with widely used executable or verifyed executable code and with a well-described model that does not require empirical selection of parameters.

Well-known metrics included in the popular stringdist package of R are involved in the tests and are described in detail in [8]. Also the NCS similarity metric is ivolved that is promoted by the author as a more effective alternative to LCS, proposed and investigated in [9-11]. For the experiment we used the code in C, published in [11], and run it from Perl XS, which is competetive in speed to the stringdist package.

A simple and clear measure of the effectiveness of the similarity metric is its *accuracy*, that we define as the probability of guessing by its values the translation of a known fragment of text into an unfamiliar language among other fragments of text in the same unfamiliar language. In practice this probability is calculated as the ratio of the number of fragments with a smaller value (or with a larger distance if the distance metric is used) to the total number of fragments that are not a translation.

3 Technique of experiments

Perl XS was used for basic processing. To calculate metrics from stringdist, Perl generated and ran a script on R. Since not all metric calculation routines support utf8, transliteration of diacritics was required. For this purpose, the Text::Unidecode packages were used.

The full source code of the executed script and custom

Submitted to the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

perl XS module are public available over this link (4k).

The preliminary tests confirmed the negative effect of metric normalization of similarity metrics. The usual normalization of the LCS (length of the longest common subsequence) similarity metric in stringdist made this metric much less efficient than the others. This effect was already known and was explained in details with examples in [12] and investigated in [13].

3 The experiment with close lengths

Normalization of string similarity metric is carried out by recalculating the values of the metric according to the heuristic formula, taking into account the length of the strings. Normalization of string similarity metric is carried out by recalculating the values of the metric according to the heuristic formula, taking into account the length of the strings. Since the proper formula for the metric normalization is an unresolved problem (for more details, see [13]), than many comparisons were excluded from consideration, in which a significant difference in the lengths of a and b lines can affect result more than the metric itself.
Since the metric's effectiveness is not related to a specific set of values, than the empirical formulas

$$mLCS = max(a,b) - LCS \quad (1)$$

$$mNCS = max(a,b) + \frac{|a-b|}{4} - NCS \quad (2)$$

were used instead of usual normalizing the similarity metrics. The formula (1) is a simple monotonic linearization of the Daniel Bakkelund's metric [14].
The Edgar Poe's book "Fall of the House of Usher» was used for evaluation. It is presented by a parallel text of 265-270 medium-sized fragments of 195-228 bytes in 7 languages, which gives more than 5000 different links of related texts among more than 1500000 possible couples of unrelated texts.

All the used couples were divided into 8 groups with a different ratio of lengths by the value of factor

$$\delta = max\left(\frac{a}{b}, \frac{b}{a}\right) - 1 \quad (3)$$

and in each part of the experiment was carried out independently. The results are shown in Fig. 1-8, where the effectiveness of metrics is marked on the vertical axis, and pairs of languages are arranged horizontally in order of increasing effectiveness.

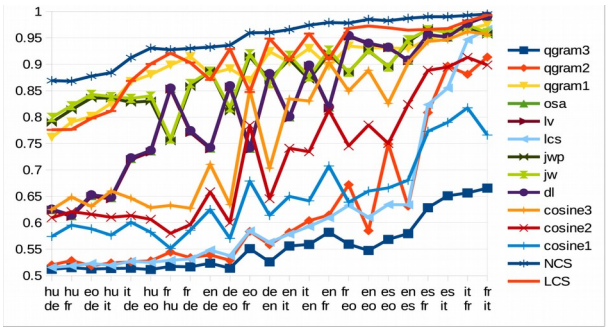

Figure 1 The accuracy of metrics (Edgar Poe, δ<0.01).

Figure 2 The accuracy of metrics (Edgar Poe, 0.01<δ<0.02).

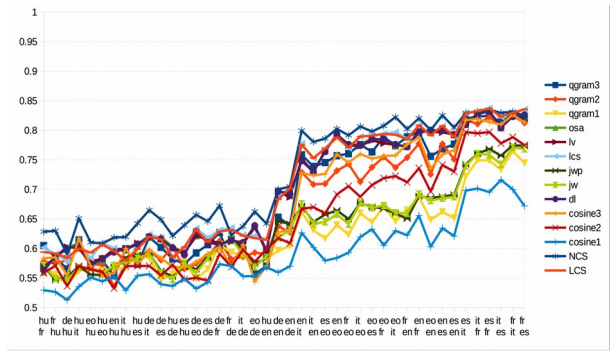

Figure 3 The accuracy of metrics (Edgar Poe, 0.02<δ<0.05)

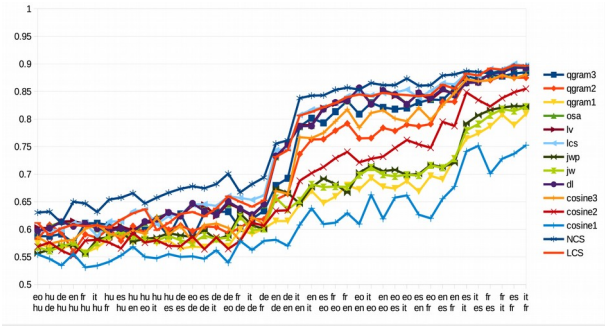

Figure 4 The accuracy of metrics (Edgar Poe, 0.05<δ<0.1).

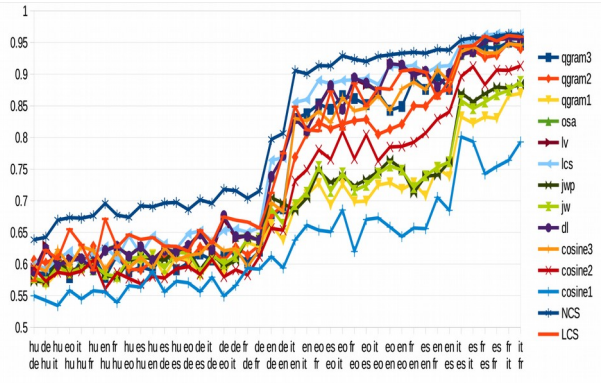

Figure 5 The accuracy of metrics (Edgar Poe, 0.1<δ<0.5)

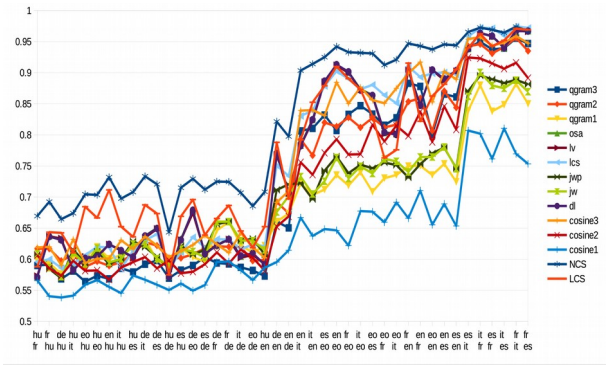

Figure 6 The accuracy of metrics (Edgar Poe, $0.5<\delta<1$)

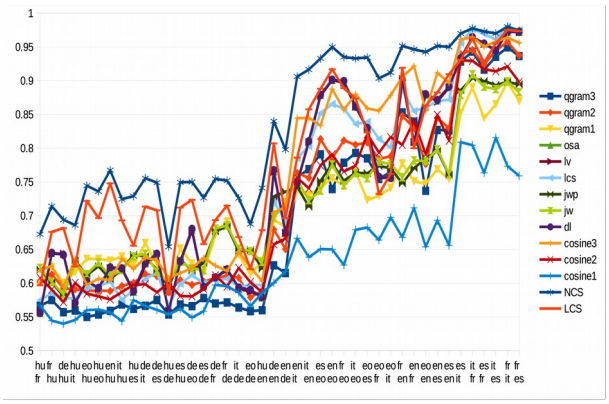

Figure 7 The accuracy of metrics (Edgar Poe, $1<\delta<2$)

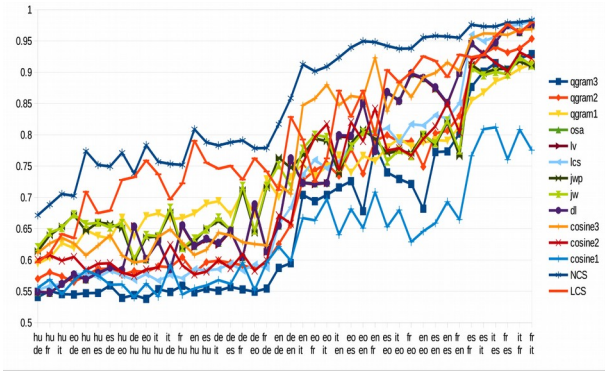

Figure 8 The accuracy of metrics (Edgar Poe,$2<\delta<4$)

The graphs clearly show the following patterns:

1. The leader and outsider among the metrics practically do not depend on either languages, or on the ratio of lengths.
2. Despite the limited ratio of lengths, the performance of normalizations of LCS and lcs is either very close, or significantly different in favor of LCS, indicating the remaining significance of the factor of dependence on the lengths of normalization lcs.
3. The empirical formulas (1) and (2) are quite effective.

5 The experiment with equal lengths

The figures above convincingly show that the effects of normalization (ie, correction of the dependence of the metric on lengths by different formulas), unfortunately can strongly and unpredictably distort the results of experiments on the quality of models and algorithms.

Therefore, for the next experiment, take the thicker Tom Sawyer book with more than 4000 fragments of medium length 122-140 characters in five languages each, which gives more than 10,000 bound pairs of fragments and compare just fragments of coincident length. The results are presented on figures 9-16.

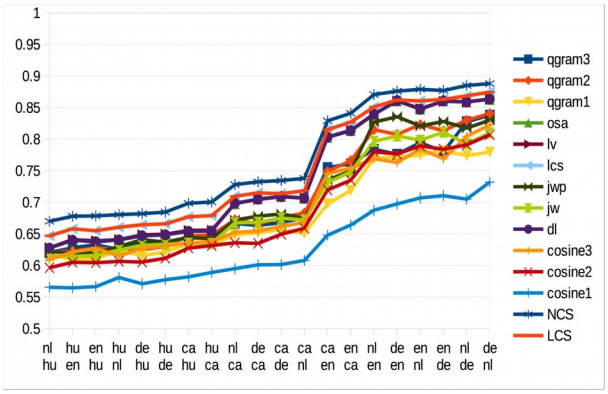

Figure 9 The accuracy of metrics (Mark Twain $\delta<0.01$) .

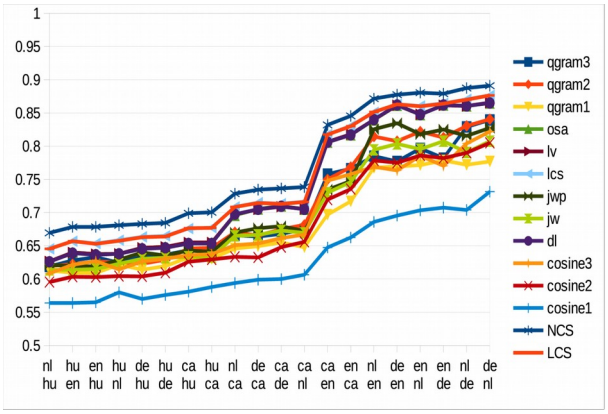

Figure 10 The accuracy of metrics (Mark Twain $0.01<\delta<0.02$)

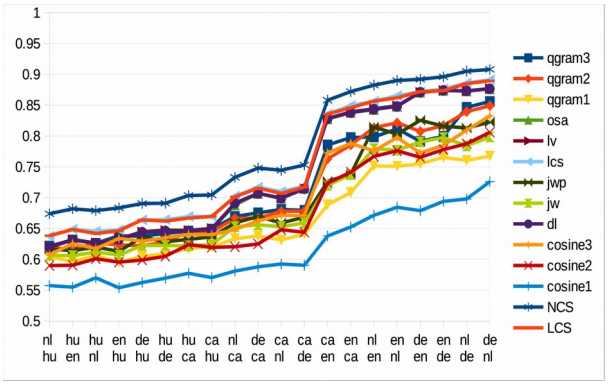

Figure 11 The accuracy of metrics (Mark Twain $0.02<\delta<0.05$)

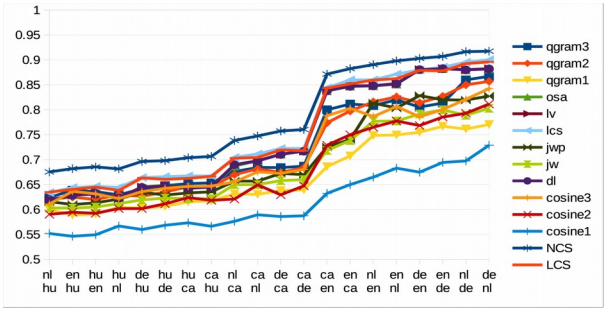

Figure 12 The accuracy of metrics (Mark Twain, $0.05<\delta<0.1$).

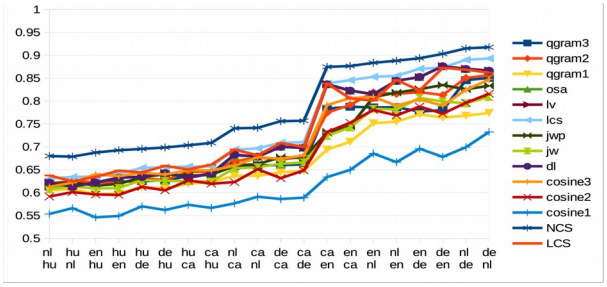

Figure 13 The accuracy of metrics (Mark Twain, $0.1<\delta<0.5$)

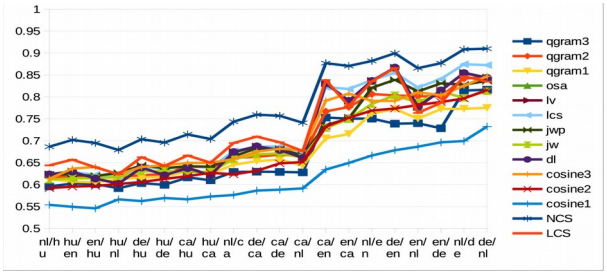

Figure 14 The accuracy of metrics (Mark Twain, $0.5<\delta<1$)

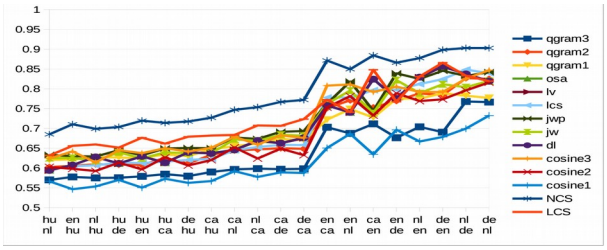

Figure 15 The accuracy of metrics (Mark Twain, $1<\delta<2$)

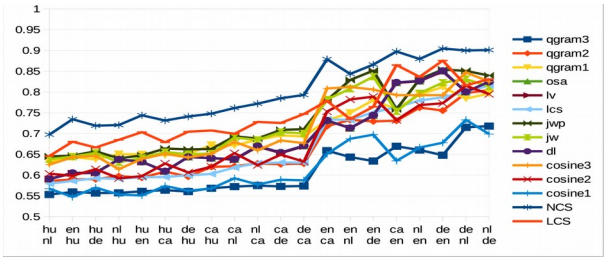

Figure 16 The accuracy of metrics (Mark Twain, $2<\delta<4$)

On the new graphs, the sensibilities of the common based metrics (LCS and lcs; also osa, lv and dl).
Figures 8-12 demonstrate both the high stability of the ranking of metrics to the change text lengths, and the less pronounced, but also stable dependence on language couples. Unfortunately, the length of the translation rarely differs from the length of the original by more than 20%, and therefore statistics for widely differing lengths are not so convincing.
Less than 3% of available texts were used. However, the quadratic computational complexity of the problem hinders their processing.
The calculation of the first experiment on the PC took less than four hours, the second was done in several days. The next "Three musketeers" book failed to be calculated with the same tools.

3 Conclusions

The stable ranking of the string similarity algorithms for quality was obtained. There are no clear signs that the picture can significantly change when experimenting with new books and languages. The tests show that the optimal choice of the metric depends not so much on the data (books and specific languages) as on the model laid down in the basis of the algorithm and on the correct consideration of differences in the lengths of the arguments.

References

[1] W. W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. In In International Joint Conference on Artificial Intelligence (IJCAI) 18, Workshop on Information Integration on the Web, 2003.

[2] Branting, K. “A comparative evaluation of name-matching algorithms.” in ICAIL, 2003, pp. 224–232.

[3] Christen, P. 2006. A comparison of personal name matching: Techniques and practical issues. In Data Mining Workshops, Sixth IEEE International Conference on Data Mining (Hong Kong, December 18-22, 2006). IEEE, New York, 290-294.

[4] G. Recchia, and MM. Louwerse. A Comparison of String Similarity Measures for Toponym Matching. In COMP@ SIGSPATIAL, 54-61, 2013.

[5] Najlah Gali, Radu Mariescu-Istodor, Pasi Fränti. Similarity Measures for Title Matching. 2016 23rd International Conference on Pattern Recognition (ICPR) Cancún Center, Cancún, México, December 4-8, 2016

[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) 957–964 (wordnet)

[7] Maria del Pilar Angeles, Adrian Espino-Gamez. Comparison of methods Hamming Distance, Jaro, and Monge-Elkan. International Conference on Advances in Databases, Knowledge, and Data Applications. DBKDA 2015., Rome, Italy (generated)

[8] M.P.J. van der (2014). The stringdist package for approximate string matching. R Journal 6(1) pp 111-122

[9] Znamenskij S. V. Simple essential improvements to ROUGE-W algorithm Journal of Siberian Federal University. Mathematics & Physics, 4 (2015) 258–270.

[10] Znamenskij S. V., A Belief Framework for Similarity Evaluation of Textual or Structured Data, Similarity Search and Applications LNCS 9371 (2015), 138–149. doi: 10.1007/978-3-319-25087-8_13

[11] Znamenskij S. V.: A model and algorithm for sequence alignment. Program systems: theory and applications 6:1, 189–197 (2015). doi:10.25209/2079-3316-2015-6-1-189-197

[12] Znamenskij S., Dyachenko V. 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 (in Russian).

[13] Znamenskij Sergej V. From Similarity to Distance: Axoim Set, Monotonic Transformatons and Metric Determinacy Journal of Siberian Federal University. Mathematics & Physics 2018, 11(3), 1–12 doi:10.17516/1997-1397-2018-11-3-1-11.

[14] Bakkelund D. “An lcs-based string metric,” University of Oslo, (2009).