ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20150831012553

DOCUMENT SIMILARITY IN REPEATEDLY TRANSLATED CORPORA

Vladimir Mateljan, Vedran Juričić, Dario Ogrizović

Preliminary communication

The paper analyses the changes in relationship between documents in textual corpus that occur due to the translation into another language. Authors analyzed the similarities between documents in original corpus, in Croatian, and compared them with the corresponding documents in translated corpus, in English. The changes were analyzed using two measures, chi-square test's P-value and new proposed measure, correction coefficient.

Keywords: *analysis; document similarity; multilingual; translated corpus; translation*

Sličnost dokumenata u uzastopno prevedenim korpusima

Prethodno priopćenje

Rad analizira promjene u odnosima dokumenata u tekstualnom korpusu koje nastaju zbog prevođenja na drugi jezik. Autori su analizirali sličnosti među dokumentima u originalnom korpusu, na hrvatskom jeziku, i usporedili ih sa sličnostima odgovarajućih dokumenata prevedenog korpusa, na engleskom jeziku. Promjene su analizirane koristeći dvije mjere, P-vrijednost hi kvadrat testa i novo uvedenu mjeru, koeficijent sličnosti.

Ključne riječi: *analiza; prevedeni korpus; prijevod; sličnost dokumenata; višejezičnost*

1 Introduction

Document similarity is a metric defined over a set of documents, where a distance between them is based on the likeness of their meaning or semantic content [1]. It is a problem of finding partial or whole overlap of elements (for example, paragraphs, sentences, parts of sentences) between two documents. The greater number of overlapping elements between two documents increases the probability of their similarity and is inversely proportional to their distance.

Documents are generally not compared directly on their content, but on their representation. Representations that can be used in document similarity are based on model from information retrieval field. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [2]. In order to avoid linearly scanning the text each time the document needs to be processed or compared to another document, documents are indexed in advance [2]. These indexes are then used to compare the documents by using one of known algorithms.

One of the models that can be used to compare documents is Boolean retrieval model that uses operators from Boolean algebra, and calculates the similarity based on binary document representation. In our system we use vector space model that represents documents as vectors in N-dimensional space [10]. Documents are then compared by using one of the similarity measure techniques, for example: Jaccard similarity measure, Euclidean distance measure or Cosine similarity measure.

Jaccard similarity measure or Tanimoto coefficient [4] measures similarity as the intersection divided by the union of the objects. Euclidean distance [1] is a square root of sum of squared distances between each vector dimension. In cosine similarity, the document similarity is measured as the cosine of the angle between vectors. The angle between vectors is in the interval [0, 90], so the cosine similarity value varies between 0 and 1.

Other measures can also be used to calculate similarity, like Greedy String Tilling [3], Dice coefficient [5], Levenshtein similarity [6] etc. The similarity measure that was used to compare documents in our algorithm is cosine similarity, because a significant number of preliminary tests showed that it has top performance in speed, precision and recall, and is optimum similarity measure for our system.

2 Text processing and similarity

This paper analyses the differences in similarities between original documents, written in Croatian, and their translated documents, written in English. The corpus that was used for analysis had 3000 documents, that is, 3000 articles that were collected from ten Croatian most popular news portals. There were no limitations on subject or tag of an article, only on its size: we did not collect articles shorter than 100 words.

The Croatian corpus is translated to the English corpus by using the Google Translate service [7]. This online service was chosen over others, for example Microsoft Bing Translator [8] or Yahoo! Babel Fish [9], because it shows better results when translating to English, and has free API that was more suitable for integration with our comparison system. Although it is a free service, it has limitations in the text length and the number of requests that can be made from one internet address, so the tests that were made had to consider that limitation.

2.1 Similarity within corpus

The similarity between two documents in the same corpus is calculated using the following algorithm:

1. Split documents D_1 and D_2 into sentences d_{11}, d_{12}, …, d_{1n} and d_{21}, d_{22}, …, d_{2m}. Store them so they can be used in other algorithms.
2. Calculate the similarity coefficient sim (d_{1k}, d_{2l}) between each sentence in the first document (d_{1k}) and

all sentences in the second document (d_{21}, d_{22}, …, d_{2m})

3. Calculate the similarity between each sentence in the first document and the second document

$$\mathrm{sim}(d_{1k}, D_2) = \max(\mathrm{sim}(d_{1k}, d_{21}), ..., \mathrm{sim}(d_{1k}, d_{2m})) \quad (1)$$

4. Sort the similarities sim (d_{1k}, D_2) descending
5. Select predefined number of similarities from the top of the list and calculate the similarity between documents D_1 and D_2

The second and the fifth step have a major impact on the processing speed and system's performance, especially on system's precision and recall. The second step preprocesses documents: the sentences are first converted to N-grams which are hashed and appropriately stored to increase system's speed. It also compares N-grams of the first document's sentence with N-grams of the second document.

As it was already written, the similarity measure that was used to calculate similarity between N-gram is Cosine similarity. Similarity coefficients take a value from 0 to 1, where 0 indicates there is no similarity, and 1 indicates a complete match.

In the last, fifth step, the algorithm calculates the similarity between documents, based on calculated similarities between sentences. One solution is to calculate arithmetic mean between all similarity coefficients. The problem with this approach is that it will not detect similarities between large documents that partially overlap in a small number of sentences. For example, if comparing two documents with 1000 sentences, and documents have only 10 similar sentences, the majority of coefficients will be zero, and the similarity between documents will be around 0,1 %.

We have used approach that does not calculate arithmetic mean between all coefficients, but on only part of them. The number of sentences that are included in calculation was determined experimentally. It is shown that the best results are achieved when the mean is calculated on 6 sentences. Algorithm uses pondered arithmetic mean that considers not only similarity coefficients, but also sentence length and the number of overlapping N-grams in sentences that were used for calculation.

The final result of document comparison within the same corpus is an upper matrix of similarity *S*. Elements above diagonal represent similarity coefficients between each document in a corpus.

$$s_{ij}^{D} = \mathrm{sim}(D_i, D_j), s_{ij}^{D} = 0 \text{ when } i \geq j \quad (2)$$

2.2 Translation

Translation of each document D_i to T_i from Croatian to English was performed using this algorithm:

1. Segment document D_i into sentences d_{i1}, d_{i2}, …, d_{in}. The sentences were reused from the previous algorithm (step 1).
2. Translate sentences d_{i1}, d_{i2}, …, d_{in} to English. The result are sentences t_{i1}, t_{i2}, …, t_{in}
3. Form document *T* using sentences t_{i1}, t_{i2}, …, t_{in}

$$D_i \xrightarrow{\text{segment}} d_{i1}, ..., d_{in} \xrightarrow{\text{translate}} t_{i1}, ..., t_{in} \xrightarrow{\text{integrate}} T_i \in T \quad (3)$$

The sentences are segmented prior to translating because it was observed that Google Translate returns different results when the sentences are separated by different characters. For example, if sentences are separated by space, the translation is correct. But, if there is no space character between them, two scenarios are possible: the translation is correct, but differs from one when they are separated by spaces, and the second scenario in which the translation is incorrect. The incorrect translation of a sentence mainly contains words from the following sentence.

Translated (English) corpus is analyzed analogous, like the original (Croatian) corpus, as described in section 2.1. The result of this analysis is upper similarity matrix with elements

$$s_{ij}^{T} = \mathrm{sim}(T_i, T_j), s_{ij}^{T} = 0 \text{ when } i \geq j \quad (4)$$

2.3 Differences between corpora

Similarity between two documents in the original corpus is usually different from their similarity in the translated corpus. For example, it is possible that two original documents have very small similarity, but when they are translated to English, their similarity significantly increases. There are two major causes for this: the translation process and the comparison algorithm.

Google Translate service can generate quite different translations for a very small change in the text, for example, when words are reordered. The second cause is the comparison algorithm that cannot adequately adapt itself and its configuration so that it neutralizes changes that occur in translations. The algorithm calculates similarity by taking many parameters: N-Gram size, minimum number of overlapping N-Grams, overlapping N-gram ratio, sentence length, etc.

The problem is that translated sentences do not have equal numbers of words, or N-grams, as their originals. This affects the similarity, because it is based on a number of overlapping N-grams in sentences. For example, if algorithm compares sentence with 10 N-grams to sentence with 20 N-grams, and finds 5 overlapping N-grams, the calculated similarity is 0,5 (5/10). If those sentences are translated and shortened to 6 and 15 N-grams, the algorithm might find only 2 common N-grams, and so the similarity is 0,33 (2/6).

This research's main goal is to determine the differences in similarity coefficients between original and translated corpora. One of the approaches that can be used as a measure of differences between corpora is to calculate the maximum difference between similarity coefficients in original and translated corpus.

Tab. 1 shows an example of comparing corpora with three documents. It can be seen that the similarity between documents d_1 and d_2, and d_2 and d_3 is almost equal in both corpuses, but similarity between translated d_1 and d_3 is greater for 0,4 than between original documents. Calculated difference between corpora, using above described approach, is 0,4 (or 40 %), which is not appropriate, because other similarities are about the same.

Table 1 Document comparison – example coefficients

	1	2	3
1	0	0,52	0,23
2	0	0	0,02
3	0	0	0

	1	2	3
1	0	0,41	0,63
2	0	0	0,01
3	0	0	0

Difference can be calculated by using arithmetic mean between all similarity coefficients, but, as majority of coefficients in corpus is zero, the difference would be very small, even when some document pairs significantly differ.

2.3.1. Difference in frequencies

Above two approaches showed that the difference should not be analyzed directly by comparing coefficients, but it should be the characteristic of a whole corpus. Authors analyzed the differences between distributions of similarity coefficients between Croatian and English corpus. A distribution of similarity within the same corpus is formed by putting coefficients into the classes, so that classes contain the frequencies of similarity coefficients in certain range.

As it was already mentioned, similarity coefficients take a value from 0 to 1, so there is an infinite number of possible classes. If the width is too high, there is an inadequate number of classes to perform quality analysis; if the width is set too small, most classes contain very few elements, or are empty. For example, if the width is set to 0,001, there are 95 % empty classes in our case. Authors assessed the optimal width to 0,01, so that coefficients are distributed to 100 classes.

The difference in distribution of similarity coefficients between corpora is measured by using Chi square test. Chi square test for optimum algorithm configuration:

H_0: There is no difference between frequencies

H_1: There exists a difference between frequencies

$\chi^2 = 2{,}48$, $df = 15$

Critical $\chi^2_{0{,}05}(15) = 24{,}99$

$p = 0{,}99$, $p > 0{,}05$

The calculated value is above critical value, so we can reject H_1 hypothesis and conclude that there is no statistically significant difference in distributions. We called this test a successful test.

This test was run using optimum parameters. We ran a complete test (forming distributions and calculating p value) 1000 times, for every possible combination of parameters. It was determined that out of 1000 tests, 78,3 % of them were successful, that is the calculated p-value was above the critical value in 783 tests.

2.3.2 Correction coefficient

It was assumed that there exists a linear relationship between similarity coefficients of the original and the translated documents. The relationship then can be represented by using the following formula:

$$s_{ij}^{T} = e \times s_{ij}^{D} + f \qquad (5)$$

where e and f are parameters, s_{ij}^{T} is a similarity coefficient between documents d_i and d_j in the translated corpus, and s_{ij}^{D} a coefficient between them in the original corpus. The difference between similarity coefficients can be minimized by setting parameters e and f to the specific values, which were calculated by using method of least squares. The formula can be used to transform similarity coefficients between original and translated corpus and vice versa.

As it turned out, our formula needed a correction, because some of transformed coefficients were greater than one, that is, the similarity between documents was greater than 100 %. Corrected formula contains only parameter e, the coefficient of correction.

$$s_{ij}^{T} = e \times s_{ij}^{D} \qquad (6)$$

By using method of least squares, we calculated 0,93 as a value for the parameter e. This value can be used to correct or at least decrease possible translation or algorithm errors. For example, if similarity between two translated documents is 0,5, the corrected value is 0,5×0,93=0,47, which is the most probable value for similarity between the original documents. By multiplying it with similarity coefficients, the difference between similarity coefficients of original and translated corpora should be reduced.

To test this assumption, we recalculated the p-values for the original and the corrected translated similarity coefficients. Out of 1000 tests, this time 88,1 % of them were successful as opposed to 78,3 % when no correction was applied.

The correction coefficient can also be used to assess the difference in distributions. If its value is 1,0, it means that there is a complete match between coefficients of similarity of original and translated corpus, and that there is no need for correction. Analogous, lesser or greater values means that there exists a difference.

3 Retranslation

In previous tests, the original documents were translated to English, and we observed the differences between corpora. We also analyzed the changes that occur when documents are retranslated to Croatian, then again to English, etc. We expected that documents' content would be significantly changed, and that it would have negative impact on relationship between documents and their similarities.

In order to analyze the impact, we randomly selected 500 documents from the original corpus and repeatedly translated them to English and back, until we had a total of 20 corpora. We had to limit the number of translated documents and translation steps because of the Google Translate limitations, and because of the translation and document analysis processing time.

We analyzed the percentage of successful tests (based on Chi square test's p-value) and the value of correction coefficients. Fig. 1 shows the results of analysis. Full lines represent *p*-values, and dotted line represents correction coefficients. Vertical axis represents the value of those two measures, and horizontal axis represents the ordinal of translation process. For example, values for the first translation processes are the result of analysis of the original corpus and first translation of documents; for the second translation process values are calculated by analyzing the original corpus and the second translation of documents, etc.

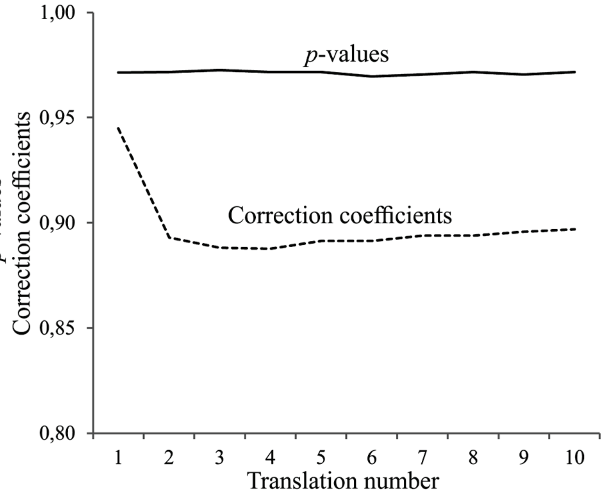

Figure 1 Relation between *p*-values, correction coefficient and translation number

It can be seen that the correction coefficient decreases in first two translations, and that it is almost constant afterwards. The percentage of successful tests almost does not change at all. The greatest difference in correction coefficient from the second to the tenth translation is 0,06, and the difference in percentage of successful tests is 0,003. Differences are negligible if we consider that there were 19 translations from the original to the final corpus.

Decrement of correction coefficient in the second translation can be explained by the fact that the translation introduces certain modifications in documents' content. Because the changes that occurred as a result of the first translation affect the second translation, the correction coefficient between second translation and the original will be smaller than between first translation and the original corpus. Smaller correction coefficient means greater difference between similarity coefficients, caused by modifying document content.

However, after the second translation, the correction coefficient almost does not change, so it can be concluded that the results of later translations also do not change. Accordingly, the content of a document in the third translation is almost identical to the content of this document in the fourth, fifth, etc. translation. This also means that their original documents, in Croatian, are almost identical.

3 Conclusion

This paper proposes the measure for determining differences between original and translated corpora by calculating the correction coefficient between them. When two corpora are identical, that is, when there are no changes in document relations, the correction coefficient has value 1. Larger deviations from this value mean greater differences in corpora.

Authors have also shown that the differences that occur by document translations can be reduced by multiplying similarity coefficients with the correction coefficient. The correction coefficient reduces the difference between two documents in original corpus and their translated document in the other corpus. This is proven by analyzing Chi square test's *p*-values with and without using the correction coefficient.

4 References

[1] Deshpande, R.; Vaze, K.; Rathod, S.; Jarhad, T. Comparative Study of Document Similarity Algorithms and Clustering Algorithms for Sentiment Analysis. // International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 3, 5(2014), pp. 196-199.

[2] Manning, C. D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA, 2008. DOI: 10.1017/CBO9780511809071

[3] Wise, M. J. String similarity via greedy string tiling and running Karp-Rabin matching. Online Preprint. 1993.

[4] Niwattanakul, S.; Singthongchai, J.; Naenudorn, E.; Wanapu, S. Using of Jaccard coefficient for keywords similarity. // Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol. 1, 2013.

[5] Thada, V.; Jaglan, V. Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm. // International Journal of Innovations in Engineering and Technology. 2, 4(2013), pp. 202-205.

[6] Fred, A. L.; Leitao, J. M. A comparative study of string dissimilarity measures in structural clustering. // International Conference on Advances in Pattern Recognition, Springer London, 1999, pp. 385-394. DOI: 10.1007/978-1-4471-0833-7_39

[7] Google Translate. https://translate.google.com, 04.04.2015.

[8] Bing Translator. https://www.bing.com/translator, 04.04.2015.

[9] Babelfish. http://www.babelfish.com/, 04.04.2015.

[10] Salton, G.; Wong, A.; Yang, C. S. A vector space model for automatic indexing. // Communications of the ACM. 18, 11(1975), pp. 613-620. DOI: 10.1145/361219.361220

Authors' addresses

Vladimir Mateljan
University of Zagreb,
Faculty of Humanities and Social Sciences,
Ivana Lučića 3, 10000 Zagreb, Croatia
vmatelja@ffzg.hr

Vedran Juričić
University of Zagreb,
Faculty of Humanities and Social Sciences,
Ivana Lučića 3, 10000 Zagreb, Croatia
vjuricic@ffzg.hr

Dario Ogrizović
University of Rijeka, Faculty of Maritime Studies,
Studentska 2, 51000 Rijeka, Croatia
dario@pfri.hr