Computing and Informatics, Vol. 37, 2018, 509–532, doi: 10.4149/cai_2018_2_509

ASHUR: EVALUATION OF THE RELATION SUMMARY-CONTENT WITHOUT HUMAN REFERENCE USING ROUGE

Alan Ramírez-Noriega, Reyes Juárez-Ramírez
Samantha Jiménez, Sergio Inzunza

Universidad Autónoma de Baja California
Facultad de Ciencias Quimicas e Ingeniería
Calzada Universitaria 14418, Parque Industrial Internacional
Tijuana, Baja California, C.P. 22390 México
e-mail: {alan.david.ramirez.noriega, reyesjua, samantha.jimenez, sinzunza}@uabc.edu.mx

Yobani Martínez-Ramírez

Universidad Autónoma de Sinaloa
Facultad de Ingeniería Mochis
Fuente de Poseidon y Angel Flores s/n, Col. Jiquilpan
Los Mochis, Sinaloa, C.P. 81223 México
e-mail: yobani@uas.edu.mx

Abstract. In written documents, the summary is a brief description of important aspects of a text. The degree of similarity between the summary and the content of a document provides reliability about the summary. Some efforts have been done in order to automate the evaluation of a summary. ROUGE metrics can automatically evaluate a summary, but it needs a model summary built by humans. The goal of this study is to find a quantitative relation between an article content and its summary using ROUGE tests without a model summary built by humans. This work proposes a method for automatic text summarization to evaluate a summary (ASHuR) based on extraction of sentences. ASHuR extracts the best sentences of an article based on the frequency of concepts, cue-words, title words, and sentence length. Extracted sentences constitute the essence of the article; these sentences construct the model summary. We performed two experiments to assess the relia-

bility of ASHuR. The first experiment compared ASHuR against similar approaches based on sentences extraction; the experiment placed ASHuR in the first place in each applied test. The second experiment compared ASHuR against human-made summaries, which yielded a Pearson correlation value of 0.86. Assessments made to ASHuR show reliability to evaluate summaries written by users in collaborative sites (e.g. Wikipedia) or to review texts generated by students in online learning systems (e.g. Moodle).

Keywords: Text summarization, summary evaluation, ROUGE, sentences extraction

Mathematics Subject Classification 2010: 68-U15, 68-T50

1 INTRODUCTION

The objective of automatic text summarization is the reduction of an original text to a smaller number of sentences by means of a computer, while keeping the important ideas intact [8]. Many areas use automatic text summarization such as intelligent tutoring systems, telecommunication industry, information extraction, text mining, question answering, news broadcasting, and word processing tools [19, 30].

The information explosion on Internet requires a reduction in the amount of information size and an increase in information efficiency [30]. These activities become easier with automatic summarization because fewer lines may represent the most important information about a document. Thus, users can find the resources more quickly [2, 16].

A summary evaluation shows the high-points of the original text. Manual summary evaluation is the first option because human assessment guarantees achievement of the desired results. However, a text can have many useful summaries; these show the main disadvantages of a manual evaluation approach, as a different evaluator may not agree [20] in determining the correct summary. The manual comparison of peer summaries based on model summaries is an activity that requires much effort and time [25].

Development of evaluation methods for summarization is difficult. Human summaries vary for many reasons such as knowledge, biases, goals, and the intended audience [23]. There are methods to evaluate summaries such as ROUGE [12], BE [9], and Pyramid [23]. They are widely used in summarization to analyze summary content [3]. These methods need human impact to work efficiently, and are considered semi-automatic [16, 18].

Previous methods require a model summary or a set of model summaries to function. The extraction of a model summary is a time-consuming and expensive task [17]. It is necessary to have an ideal summary and the original text to automate this process in these evaluation systems completely.

The purpose of this article is to evaluate a summary without the human model input. Two phases divide the process: The first phase extracts the most representative sentences from the content through an algorithm based on frequencies of concepts, cue-words, title words, and sentence length. Despite being simple and not requiring an in-depth level of knowledge analysis, this technique is suitable for building summaries [16]. The second phase evaluates the original summary based on ROUGE metrics and the built summary in the first phase. The system is called ASHuR (Assessing Summaries without Human reference using ROUGE).

The remainder of the article is structured as follows: Section 2 describes related works. Section 3 explains related topics such as text summarization and tests to evaluate summaries. Section 4 outlines the proposed approach. Sections 5 and 6 describe two experiments together with the results and discussions. The final sections show conclusions and references.

2 RELATED WORK

There are studies related to the evaluation of previous summaries that have dealt with this problem. These studies have faced this issue because of the importance of a summary in the field of education, and its ability to provide a general idea of a lengthy document.

In [11] the authors proposed an integrated method to evaluate summaries using Latent Semantic Analysis (LSA) automatically. This method is based on a regression equation calculated with a corpus of a hundred summaries. It is validated on a different sample of summaries. The equation incorporates two parameters extracted from LSA: semantic similarity and vector length. The aim of this study was to use a simple and innovative LSA-based computational method to evaluate summaries reliably. Despite the efforts made in this article, the authors needed a training set for their algorithms to work. The training set is only for a common topic, which is the limit of this particular idea; a summary of 50 words works in only a few cases. A summary, limited to that number of words excludes many other situations where the evaluation system could be used.

FRESA [29] is a Framework for Evaluating Summaries Automatically, which includes document-based summary evaluation measures based on probabilities distribution. FRESA supports different n-grams and skips n-grams probability distributions. In addition, this environment evaluates summaries in various languages. This framework is an alternative to ROUGE in evaluating summaries based especially on the Jensen-Shanon divergence. FRESA takes the original text as a model, without requiring human intervention, and compares it to the abstract obtained automatically. Their system extracts phrases in evaluating the summary, however, human summaries give bad evaluation results because FRESA considers complete coincidences in sentences. FRESA metrics based on divergence are not perceived clearly and quickly. The conclusion is that values of the metric give a high value of

divergence between a text and its summary, this is always applicable to the phrases that are used in this system. Thus, FRESA associates values of great divergence regardless of the strategy used, including random compression. Therefore, there is not an adequate way of evaluating summaries [20].

Louis [18] presented and evaluated a suite of metrics which do not require gold-standard human summaries for evaluation. They proposed three evaluation techniques, two of which are model-free and do not rely on the gold standard for the assessment. The third technique improves standard automatic evaluations by expanding the set of available model summaries with chosen system summaries. SIMetrix is the tool used by these authors. The metrics of this system are based on the Kullback Leibler (KLD) and Jensen Shannon (JSD) divergence, in addition to the Fraction of Topic Words (FoTW). SIMetrix, is a very versatile system, and has a variety of tests to measure the relation between the summary and its content. Although SIMetrix shows good overall results in its tests, it has not excelled in the evaluation of summaries; ROUGE is the standard that is used in reporting automatic summarization evaluation results. However, SIMetrix is used in this investigation to validate summaries.

ROUGE is the evaluation system implemented as the de-facto standard; it is the most commonly used metric of content selection quality used in research papers because it is cheap and fast [21]. ASHuR evaluates a summary based on sentences extraction considering ROUGE as the evaluation system. This is an advantage that the related work does not have.

3 FUNDAMENTALS FOR TEXT SUMMARIZATION

3.1 ROUGE

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. ROUGE is a summary evaluation method that includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans [12].

This method has the following tests [13]:

- ROUGE-N: N-gram Co-Occurrence Statistics (versions ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4).
- ROUGE-L: Longest Common Subsequence.
- ROUGE-W: Weighted Longest Common Subsequence.
- ROUGE-S: Skip-Bigram Co-Occurrence Statistics.
- ROUGE-SU: Extension of ROUGE-S.

Document Understanding Conference (DUC), National Institute of Standards and Technology (NIST), and Text Analysis Conference (TAC) adopted ROUGE package for content-based evaluation [14, 27, 26, 28]. ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU tests have been used in many investigations to evaluate

experiments because they have a greater accord with the human evaluation [14, 26, 3, 27].

The typical information retrieval metrics are precision and recall [21], these metrics are used by ROUGE to evaluate summaries [12]. Precision (Equation (1)) is the number of sentences occurring in both the system and ideal summary divided by the number of sentences in the system summary. Recall (Equation (2)) is the number of sentences occurring in both the system and ideal summary divided by the number of sentences in the model summary [27].

$$\text{precision} = \frac{|\{\text{relevantObjects}\} \cap \{\text{retrievedObjects}\}|}{|\text{retrievedObjects}|}, \tag{1}$$

$$\text{recall} = \frac{|\{\text{relevantObjects}\} \cap \{\text{Objects}\}|}{|\text{relevantObjects}|}, \tag{2}$$

$$F_\beta = (1 + \beta^2) * \frac{\text{precision} * \text{recall}}{(\beta^2 * \text{precision}) + \text{recall}}. \tag{3}$$

The appeal of precision and recall as an evaluation measure is that after a human defines the gold standard sentence selection, it can be repeatedly used to evaluate automatically produced summaries by a simple comparison of sentence identifiers [21]. F-measure (Equation (3)) is a weighted harmonic mean of recall and precision. Where β is a variable to give preference either recall or precision, when $\beta > 1$ then the preference is given to precision, and when $\beta < 1$ then the preference is given to recall. This study used the F-measure for experiments.

3.2 SIMetrix

SIMetrix tool is a group of metrics to evaluate summaries [18]. Our investigation uses the SIMetrix model without a model summary.

The following SIMetrix metrics validate our proposal [18]:

- KLInputSummary: Kullback Leibler divergence between input and summary
- KLSummaryInput: Kullback Leibler divergence between summary and input. Since KL divergence is not symmetric, the features are computed both ways Input-Summary and Summary-Input. Both features above use smoothing.
- UnsmoothedJSD: Jensen Shannon (JS) divergence between input and summary. No smoothing.
- SmoothedJSD: A version with smoothing.
- CosineAllWords: Cosine similarity between all words in the input and summary.
- PercentTopicTokens: Proportion of tokens in the summary that are topic words of the input.

- FractionTopicWords: The fraction of topic words of the input that appear in the summary.
- TopicWordOverlap: Cosine similarity using all words of the summary but only the topic words from the input.

SIMetrix results showed that the strength of features vary considerably. The best metric is JS divergence, which compares the distribution of terms in the input and summary. According to the SIMetrix documentation, higher divergence scores indicate poor quality summaries. For the other metrics, higher scores indicate better summaries.

4 PROPOSED APPROACH

The proposed approach initially divides the article into its summary and its content. The system constructs the summary model based on the original content. Finally, ROUGE evaluates the model summary and the summary of the original article to obtain the summary assessment. Figure 1 displays this process.

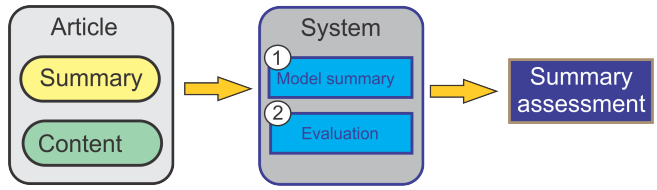


Figure 1. General diagram of the ASHuR evaluation process

4.1 Model Summary Module

This module creates a model summary, the following steps details the process:

1. Identification of raw sentences: This step obtains raw sentences from the content. A raw sentence is one taken from the original text without any special treatment. The ASHuR process begins with this set of sentences; after a process of cleaning, splitting, and scoring of the sentences, our system takes sentences from original raw sentences to produce the summary.
2. Determination of concepts frequencies: ASHuR applies a text cleaning process to raw sentences. Such process involves the following phases:

    - Tokenize sentences: The tokenization breaks down the sentences into a set of words [8] called tokens. The token is the minimal unit to analyze the text in this study.

- Delete stop-words: Stop-words are words that are insignificant in our method. Therefore, ASHuR eliminates stop-words from the original text. The stop-words list includes the most frequently occurring words in a text (e.g. a, the, of, etc.) [5].
- Apply stemming: The stemming technique uses the root form of a word. The primary objective is to assign equal importance to words having the same root. Thus, words expressed in their different forms are considered to be the same [8]. Our proposal uses Porter's algorithm to apply stemming; this is the most common method used in literature [24].

ASHuR gains word frequencies after the cleaning process. This information is useful to assess the impact of the sentence in the document. This phase obtains a processed version of raw sentences.

3. Identification of the article title: Words in the title always represent the main idea of the text. The title plays a particular role in ASHuR because sentences that have title words are more important than other sentences. The title follows the same cleaning process as the rest of the text.
4. Definition of signal words: This phase uses a technique where phrases or words determine the relevance of a sentence, these words are called signal words. There are different kinds of signal words, however ASHuR works with words related to importance such as greatness, conclusion, summary, etc. [16]. These words may be a good indicator of relevant information [4, 27]. This study employs a list of signal words based on [10].
5. Calculation of the sentences score: This phase calculates the score of each sentence based on frequencies and the amount of words. Title words and signal words found in the sentence also proportionally influence the score.
6. Selection of the best sentences: This phase chooses the sentences with the highest score while discarding the sentences which are too short. These sentences are in order according to their score. The total number of words in a sentence must be similar to the number of words of the original summary. ASHuR selects sentences representing the summary of the version of raw sentences.

4.2 Evaluation Module

The first module of the summarization system generates the summary of the original article. ROUGE metrics then compare the generated summary with the model summary. Figure 2 represents the complete process of ASHuR.

For the evaluation part of the process, this study employs the following ROUGE tests: ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU. This study calculates the mean result of ROUGE tests to obtain a single result, however, another option could be to take a ROUGE test to represent the evaluation of the summary. We consider that ROUGE is a useful tool for the tasks assessment and that a new algorithm for this assessment is not necessary.

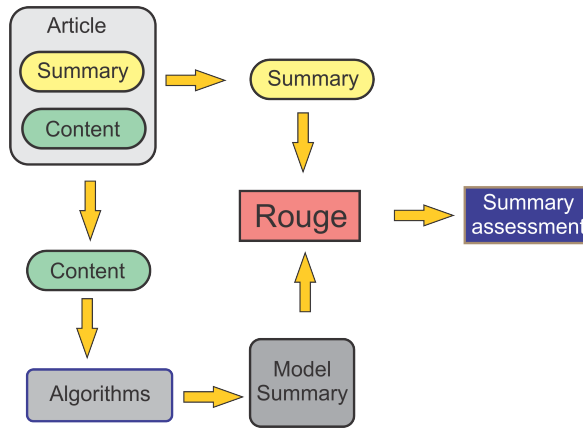


Figure 2. Specific diagram of the ASHuR evaluation process

4.3 Formal Representation of the Model

The process of ASHuR is represented in the following definition and equations:

Definition 1. Let $R_{procesed} = \{T_1, \ldots, T_{|R_{procesed}|}\}$ be the set that represents an article R_{raw} with a text cleaning process in sentences. Each element in $R_{procesed}$ has the form $T_j = \{t_1, \ldots, t_{|T|}\}$, where elements in T represent words in a sentence.

Definition 2. Let $S = \{c_1, \ldots, c_{|S|}\}$ be the set that represents the score of sentences found in a document, where each element in S represents an ordered pair of the form $c_i = (r, d)$, the element r represents the score of the sentence and the element d represents the number of words in a sentence.

After of previous definitions, for each element $T \in R_{procesed}$, then $S_T \leftarrow (r_T, |T|)$, where the score is calculated by the Equation (4) based on Equations (5), (6), and (7). Equation (6) uses the variable a to represent a value for signal words, these words are represented by the set W. Equation (7) uses the variable b to represent a value for title words, these words are represented by the set I. The variable $|T|$ represents the number of elements in T.

$$r_T = f \cdot g \cdot l, \tag{4}$$

$$f = \sum_{t \in T} Freq(t)/|T|, \tag{5}$$

$$g = \begin{cases} a, & \text{if } |W \cap T| > 0, \\ 1, & \text{otherwise}, \end{cases} \tag{6}$$

$$l = \begin{cases} b, & \text{if } |I \cap T| > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

Let $S_{sort} = \{x_1, \ldots, x_{|S_{sort}|}\}$ be the set S ordered by the element r of the pair ordered x_k, sentences representing the summary of the article are taken from the set S_{sort}. The Algorithm 1 displays the process to obtain sentences that represent the summary. The variable a represents the sum of words in each x appended to $S_{summary}$, also, the variable max represents the maximum number of words of the summary and min represents then minimum number of words considered by sentence. The generated summary is represented in $S_{summary}$, this is used to evaluate other summaries.

Algorithm 1 Process to obtain the most important sentences

```
for all x ∈ S_sort do
    if (a < max) and (x_d > min) then
        a ← a + x_d
        append x to S_summary
    end if
end for
```

5 COMPARISON TO SIMILAR APPROACHES USING MODEL SUMMARY

5.1 Experiment

This experiment compares ASHuR to nine Summarization Systems (SS) based on sentence extraction. The aim is to assess the quality of the extracted sentences against similar approaches using model summaries. We selected SS as represented in Table 1 for the experiment because literature references to them and they are freely available.

None of the SS selected have algorithms available to be implemented. Only the applications have been published. Some of the systems are web applications, while others are applications for the Windows operating system. Others are applications for the Linux operating system. This setback complicates the automation of the evaluation process, therefore, the sample size for this iteration is not as extensive as desired.

This experiment uses research articles to perform the comparison between SS because expert researchers review these kind of documents before the publication, so that articles have quality in the abstract (summary) as well as the content. This experiment considers the abstract as the model summary of ROUGE.

Id	System
1	ASHuR
2	Autosummarizer [1]
3	Freesummarizer [6]
4	IBM Many Aspects Document Summarization Tool (furthest) [15]
5	IBM Many Aspects Document Summarization Tool (Greedyexp) [15]
6	IBM Many Aspects Document Summarization Tool (K-Median) [15]
7	IBM Many Aspects Document Summarization Tool (SVD) [15]
8	Online summarize tool [22]
9	Open text summarizer [31]
10	Swesum [7]

Table 1. Summarization systems

The test data is contained in 40 articles selected from the special issue "Social Identity and Addictive Behavior" in the Journal of Addictive Behaviors Reports [1], Volumes 1 (June 2015), 2 (December 2015), 3 (June 2016), 4 (December 2016), and 5 (June 2017). We chose this journal because it considers theoretical aspects with few equations that can hinder the work of summarization systems.

The preparation phase of documents deleted the abstract and the references, the rest of the article remained intact. The prepared papers were submitted to each SS to build its summary. The next phase compared generated summaries and the model summaries. Each algorithm made a summary per article which was contrasted with the corresponding original summary.

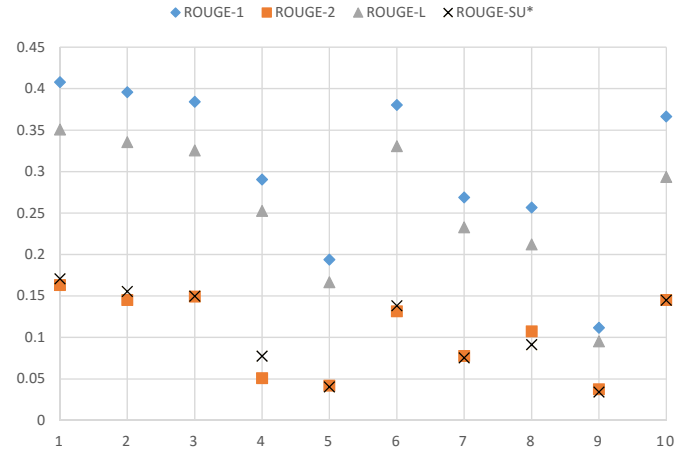
The eight tests of ROUGE evaluated results of SS considering the F-measure. The ROUGE tests result is a value between 0 and 1, the closer to one the better the summary.

5.2 Result

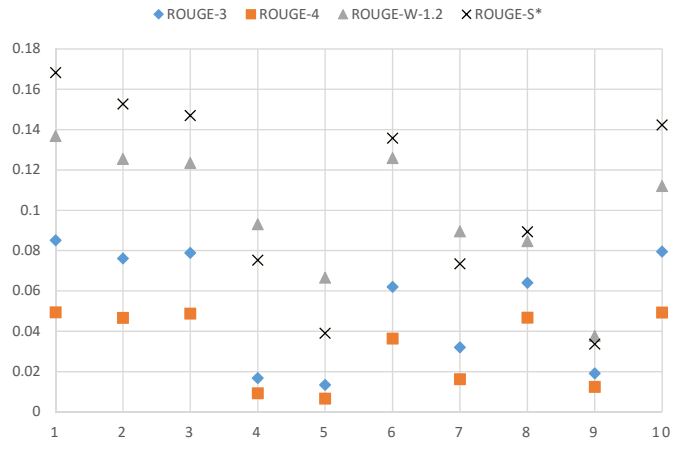
The results of the ROUGE evaluation applied to SS are displayed in Figure 3. Two groups organize the information; group 1 presents the most commonly used tests (see Figure 3 a)), and group 2 presents the rest of tests (see Figure 3 b)). The x-axis deploys Identifiers of SS and the y-axis represents the values reached by the tests. Graphs of results present ROUGE tests by a figure; rhombus, square, triangle, or cross, so, tests can be differentiated.

Means results obtained by SS in ROUGE tests are displayed in Figure 4. This figure shows the values reached in the x-axis and SS in the y-axis. The best-positioned systems are *ASHuR*, *Autosummarizer*, and *Freesummarizer* in that order. The worst positioned are *OpenTextSummarizer* and *IBM_GREEDYEXP*.

[1] `http://www.sciencedirect.com/science/journal/23528532/vsi`



a) Group 1



b) Group 2

Figure 3. Results of summarization systems considering ROUGE

5.3 Discussion

ASHuR obtained higher results in each test than the rest of SS (see Figure 3). This showed that our method achieved sentences more representative of the content of the original text.

Test files contained tables in text format, the systems positioned in the first places dealt with this point correctly. However, other systems such as *OpenTextSummarizer* had problems with the tables, which led to poor evaluation results.

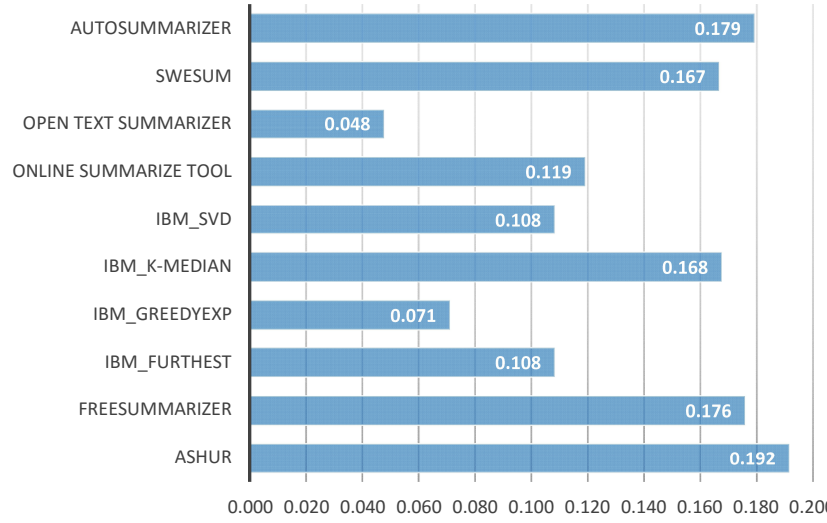


Figure 4. Means of summarization systems in ROUGE tests

It is complicated to achieve values close to one (the ideal value) in some ROUGE tests, but it is simpler in others tests. ROUGE-1 and ROUGE-L are tests that obtain higher scores than the rest of tests. ROUGE-4 is the most complicated test to overcome, this test on average had the lowest results.

The SS position of each ROUGE test varies a few places; accordingly, the ROUGE's tests maintain consistency and regularity in results, even though the score of the systems are similar. This means the summary evaluation of a system will not be in the first positions in a test and the last positions in another.

6 COMPARISON AGAINST HUMAN SUMMARY WITHOUT MODEL SUMMARY

6.1 Experiment

SS did not intervene in this second experiment because the first experiment verified that ASHuR obtains more precise results. The aim is to demonstrate that the ASHuR summary is similar to human summaries. This activity was realized with SIMetrix (Summary Input similarity Metrics) [18]. This tool analyses a text summary through similarity metrics (Section 3.2). SIMetrix is a system that allows, unlike ROUGE, to perform summary evaluations without a summary model. However, important conferences as DUC or TASC do not consider it relevant because they trust to the evaluation of ROUGE.

SIMetrix does not have ROUGE support, however, ROUGE needs a model summary to evaluate other texts. Thus, SIMetrix evaluates summaries in this ex-

periment because it does not require a model summary. The objective of this project is to generate a version of ROUGE to assess abstracts without human intervention in the same way as SIMetrix but with the support of ROUGE.

This study focuses on unstructured documents such as Wikipedia documents. This experiment considers the Wikipedia branch in the category *Main topic classifications* for test data. This category is the main one in the hierarchy of Wikipedia. The rest of the categories is derived from this one. The main category has 10 subcategories, and these contain other categories (see Table 2). This paper contemplates the direct categories of *Main topic classifications*. The categorization described corresponds to the Wikipedia version of October 1, 2016.

Categories	**Sub-Categories**	**Pages**
Main Topic Classifications	10	14
Geography	26	75
Nature	26	15
Reference works	39	25
Health	45	13
History	32	27
Philosophy	18	51
Science and technology	9	7
Humanities	33	49
Mathematics	21	12
People	34	2
Total	286	290

Table 2. Category main topic classifications of Wikipedia

The main category and sub-categories in the Table 2 contain 290 articles. This experiment did not consider articles with the following characteristics:

1. Articles without a summary (e.g. the article Caribmap[2]).
2. Articles that describe a list of other pages (e.g. the article Lost History[3]).
3. Articles that are in two or more of the considered categories (e.g. the article People[4]). Articles that met the desired characteristics were 196.

The comparison process consisted of obtaining a summary of articles for each treatment (ASHuR and human) and comparing it with the content using SIMetrix. Firstly, ASHUR generated its summary, and this was compared with the content to obtain a summary-content relation measure. We then examined the original abstract of the article (human summary) with the content getting another measure of summary-content relation. The hypothesis is that a high positive correlation

[2] `https://en.wikipedia.org/wiki/Caribmap`
[3] `https://en.wikipedia.org/wiki/Lost_history`
[4] `https://en.wikipedia.org/wiki/People`

will be achieved using the Pearson test between both measures of relation. Each summary (ASHuR and Human) was evaluated against its content by 8 SIMetrix tests.

This experiment separated tests in two clusters according to their form of evaluation. Tests that consider that the closer to 0 a result is, the better correlation will exist (B1), and tests that consider that the closer to 1 a result is, the better correlation will exist (B2). The group B1 contemplates the KLInputSummary, KLSummaryInput, UnsmoothedJSD, and SmoothedJSD tests. Group B2 contemplates the CosineAllWords, PercentTopicTokens, FractionTopicWords, and TopicWordOverlap tests.

6.2 Results

The evaluation of ASHuR results and human summaries was contrasted 196 times, one for each article. Figure 5 shows the general results organized by test and treatment (ASHuR and human).

Boxplots represent the data distribution in summarized form in Figure 5, the vertical line inside the rectangle represents the data median. The x-axis represents the applied test and the treatment, ASHuR tests are described at the label end with the letter A, and human tests with the letter H. The y-axis shows the values scale of tests; these vary according to the group of applied tests.

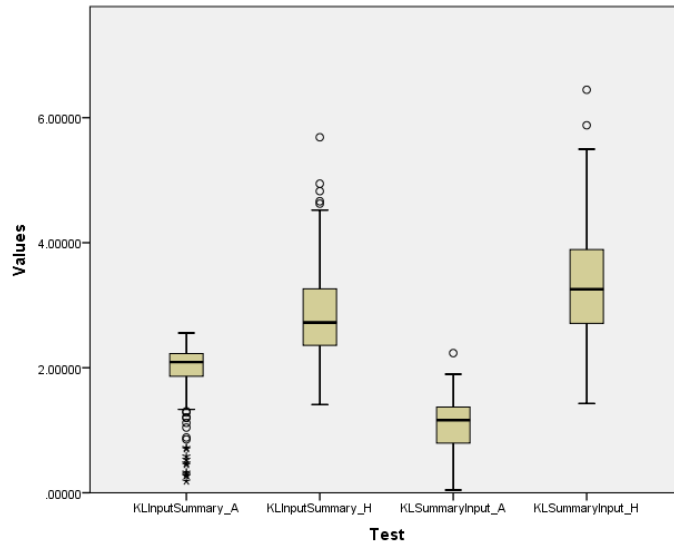
Results of Figure 5 a) represent block B1 tests, the closer to zero the means, the better the summary will be evaluated. Figure 5 b) displays results of block B2 tests ASHuR obtained values closer to zero in each of the tests, however, it also got more outliers.

The closer to 1 the means of block B2 are, the summaries will be better. The best-performing tests are CosineAllWords and TopicWorldOverlap of ASHuR. Tests evaluate summaries based on different aspects; this causes some tests to obtain results closer to zero and others more distant.

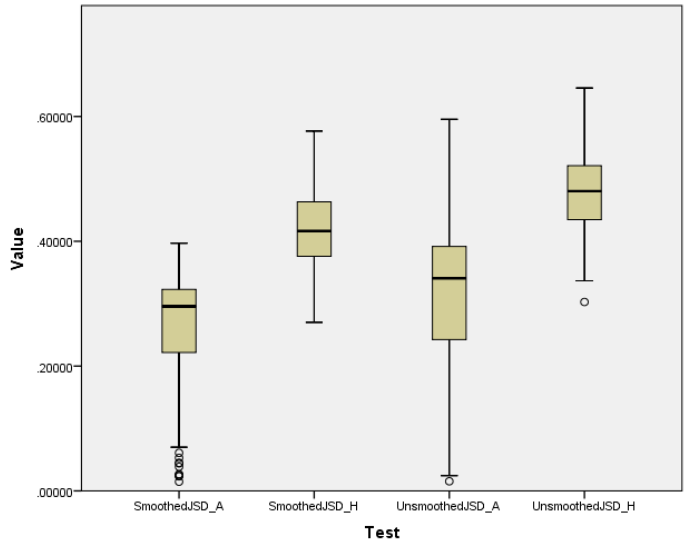
The mean and standard deviation of the tests B1 and B2 are represented in Table 3. This table shows the information according to the test group and the type of treatment (ASHuR and Human).

The Spearman correlation test compared results of ASHuR and the human considering the groups B1 and B2. Figure 6 shows results of 196 evaluations that represent each article, the x-axis represents tests blocks and the y-axis values. Although the data from block B1 are less dispersed than block B2, most of the B2 data are closer to 1, which means that block B2 has most acceptable results than block B1.

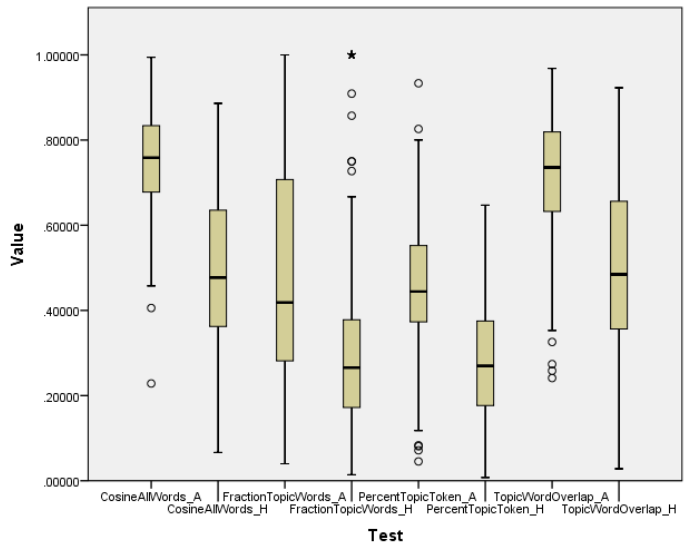
The descriptive data of evaluations are described in the Table 4. Outliers are commonly treated in some way to observe the impact of these on the outcome, for that reason data are analyzed with and without them. The block B1 had no outliers because of the low dispersion of data, however, the block B2 obtained some anomalous values.



a) Tests block B1 (part 1)



b) Tests block B1 (part 2)



c) Tests block B2

Figure 5. Data distribution of the evaluation summary-content organized by test blocks, block B1 – the best result is zero, block B2 – the best result is one

6.3 Discussion

Tests groups B1 and B2 showed a similar behavior in their results (see Figure 5). The ASHuR evaluation gave more desirable results in each test compared to the evaluation of human generated summaries. Results of ASHuR in the group B1 were closer to zero than the results of the human. The results of ASHuR in the block B2 are closer to one than human results.

The results dispersion of block B1, in Figure 5 a), shows more concise data for the evaluation of ASHuR. On the contrary, the human evaluation data are more compact in Figure 5 b), even though ASHuR data achieved a better score. Table 3 shows this information more precisely. Block B2 shows that human data are less spread than ASHuR data in most tests, however, human data do not receive a superior evaluation than ASHuR data.

The evaluation of results indicates that ASHuR generates better summaries than humans. However, these results are provided by automated tests that do not evaluate consistency and congruence of text sentences. Our best results are due to a system based on phrase extraction that is favored by this type of evaluation system. In spite of this, we made tests to put in context real results. If negative results had been obtained at this stage, it would have meant a poor phrase extraction that would have nothing to do with the important aspects of the text.

Test Group	Test	Mean	SD
B1	KLInputSummary_H	2.853	0.678
	KLInputSummary_A	1.939	0.486
	KLSummaryInput_H	3.302	0.878
	KLSummaryInput_A	1.065	0.434
	UnsmoothedJSD_H	0.479	0.057
	UnsmoothedJSD_A	0.313	0.117
	SmoothedJSD_H	0.420	0.060
	SmoothedJSD_A	0.265	0.087
B2	CosineAllWords_H	0.485	0.180
	CosineAllWords_A	0.751	0.124
	PercentTopicTokens_H	0.278	0.141
	PercentTopicTokens_A	0.447	0.161
	FractionTopicWords_H	0.317	0.218
	FractionTopicWords_A	0.504	0.280
	TopicWordOverlap_H	0.490	0.202
	TopicWordOverlap_A	0.712	0.148

Table 3. Mean and standard deviation of SIMetrix test

	B1		B2	
Descriptives	**All Values**	**Without Outliers**	**All Values**	**Without Outliers**
Correlations mean	0.812	0.812	0.865	0.901
Trimmed mean (5 %)	0.813	0.813	0.890	0.913
Median	0.822	0.822	0.943	0.951
Standard deviation	0.095	0.095	0.184	0.114
Minimum	0.536	0.536	−0.121	0.545
Maximum	0.999	0.999	0.999	0.999
P-value (mean)	0.188	0.198	0.135	0.099
P-value (trimmed mean)	0.187	0.187	0.110	0.087

Table 4. Descriptive data of the evaluation of correlation tests

The experiment applied the Pearson test to measure the degree of correlation between evaluations of ASHuR and the human. Results (see Table 4) show an average correlation between ASHuR-Human summaries of 0.812 for the block B1, this correlation means that the ASHuR summaries are 81.2 % similar to the human summaries according to the applied tests. However, the p-value obtained of 0.188 was not as good as we would wanted.

The block B2 showed an average correlation of 0.865 and an average trimmed to 95 % of 0.890, this indicates that there are 5 % of anomalous values that are negatively affecting results. When the average correlation is calculated without outliers then an average of 0.901 is obtained and an average trimmed to 95 % of 0.913.

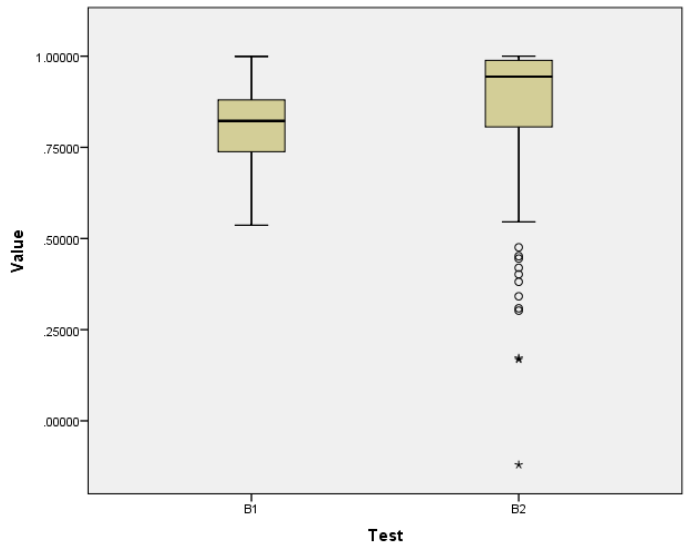


Figure 6. Results concentration of blocks B1 and B2

The block tests B2 shows a higher correlation than block test B1, even though the data of the block B2 are more dispersed than block B1. The block B2 is more prone to generating outliers. Outliers are caused by a significant difference between the results of ASHuR and the human result. These events occurred for the following reasons:

- Different use of words: Although the human summary is correct, it is poorly evaluated because different words are used in the writing of the summary and the content (4 cases).
- Different use to the summary section: The summary section has a different function than summarizing the document content, e.g., describes the use of the article instead of the content (3 cases).
- Inadequate sentence extraction: ASHuR performed an inappropriate phrases extraction due to established design characteristics of the algorithm (3 cases).
- Short summary: The summary is too short, limited to few words, this causes ASHuR only select a sentence that inappropriately represents the content (2 cases).

The proper treatment of these events will give more accurate results to ASHuR in future versions of our algorithm.

7 EVALUATION WITH ASHUR AND ROUGE

This section shows the evaluation of 21 articles of Wikipedia considering the ROUGE evaluation based on ASHuR. These articles are concepts related to Object Oriented Programming (OOP). The procedure consisted of three steps:

1. to obtain the summary using ASHuR,
2. to take the human summary from the Wikipedia article, and
3. to evaluate the human summary with ROUGE considering the ASHuR summary as a model.

ROUGE tests – ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU* – assessed articles summaries. Figure 7 shows the results of the evaluation of each Wikipedia article. Values of the graph represent the F-measure on the y-axis. The x-axis displays articles represented by an identifier. These identifiers are represented in Table 5.

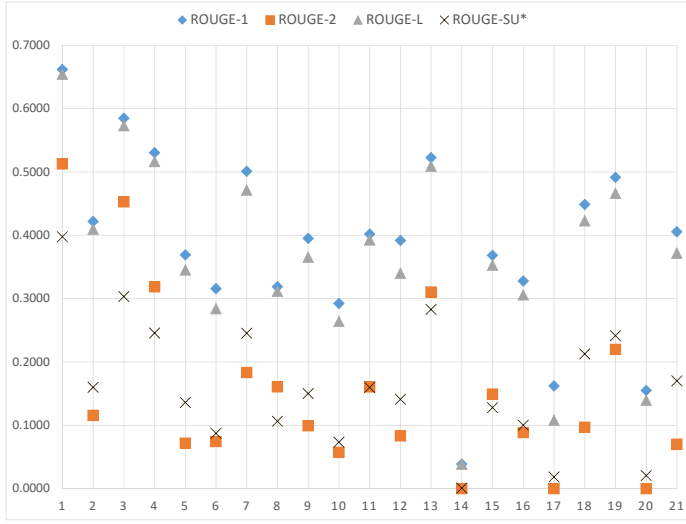


Figure 7. Results of the summaries evaluation with ROUGE and ASHuR

The summary score will depend on the test or tests considered. If a flexible evaluation is necessary, ROUGE-1 is chosen; if a harder evaluation is required, ROUGE-SU* could be used. This study employed an average of four tests, Table 5 shows results. According to the results, some articles present high probabilities to

contain an inadequate summary. Articles *overriding*[5], *composition*[6], and *persistence*[7] have the lowest evaluations and they have a probability greater than 80 % to be inadequate or at least they have a lower level than the rest of the articles.

We analyzed these articles in detail to determine why their assessments are so low:

- The article *overriding* uses too many sample code in the content, most of the text is used to explain it. The summary is adequate, but this complement the content instead of functioning as a set of ideas that represent the content.
- The article *composition* has a very short summary based on two statements, this makes the summary evaluation problematic.
- The article *persistence*, although it has an accurate summary, it is relatively short, the content does not utilize words used in the summary.

Id	Wikipedia Article	Mean
1	Abstract type	0.557
2	Abstraction	0.277
3	Access modifier	0.479
4	Attribute	0.403
5	Class	0.231
6	Concurrency	0.190
7	Constructor	0.350
8	Encapsulation	0.224
9	Overloading	0.252
10	Hiding	0.172
11	Inheritance	0.279
12	Package	0.239
13	Method	0.406
14	Overriding	0.020
15	Modularity	0.250
16	Object	0.206
17	Composition	0.072
18	OOP	0.295
19	Parameters	0.355
20	Persistence	0.079
21	Scope	0.254

Table 5. Average of ROUGE tests for Wikipedia articles

ASHuR can review documents to identify cases where the summary is inadequate to the content by an alert signal.

[5] https://en.wikipedia.org/wiki/Method_overriding
[6] https://en.wikipedia.org/wiki/Object_composition
[7] https://en.wikipedia.org/wiki/Persistence_(computer_science)

8 CONCLUSIONS

This study presents ASHuR, an algorithm to measure the relation summary-content quantitatively without a model summary using ROUGE. According to the classification given in [16], our method works as follows: based on text, works with a single document, extracts information about text with an indicative proposal, considers only one language at the time (mono-lingual), gives an evaluation without ideal summary made by humans. This investigation worked with Wikipedia articles, but ASHuR can be applied to documents with a defined structure by content and summary.

ASHuR consists of two modules. The first module builds a model summary based on content, and the second module evaluates the original summary with the model summary created. ASHuR ranked in the first place among nine SS based on sentences extraction. In another experiment, our method achieved high correlation, based on the Pearson test, between ASHuR summary and human summary.

This study shows that a text can be evaluated without a model summary based on the proposed approach. We realize that the comparison based on human summaries is the best, however, when humans are not available, our proposal could be a good option.

According to evaluations performed in the experiment, the summary assessment implemented with our approach is an approximation with encouraging results. The project gives the possibility of evaluating summaries at the moment; one or multiple model summaries are not needed. Thus, ASHuR can evaluate a summary written by users in collaborative sites (e.g. Wikipedia) or can review texts written by students stored in online repository (e.g. Moodle).

For future work, we propose to solve problems such as synonyms, anaphora, proportion summary – content according to the length and term distribution. These would improve the algorithm and the precision of the sentences. This study considers adding to ASHUR the option of offering recommendations to improve its summary, considering the most common problems encountered in the evaluation.

REFERENCES

[1] AS: Autosummarizer. 2016, `http://autosummarizer.com/`.

[2] Bagalkotkar, A.—Kandelwal, A.—Pandey, S.—Kamath, S. S.: A Novel Technique for Efficient Text Document Summarization as a Service. 2013 Third International Conference on Advances in Computing and Communications (ICACC), 2013, pp. 50–53, doi: 10.1109/ICACC.2013.17.

[3] Dang, H. T.—Owczarzak, K. K.: Overview of the TAC 2008 Update Summarization Task. Text Analysis Conference (TAC 2008), 2008, pp. 1–16.

[4] Ferreira, R.—Freitas, F.—de Souza Cabral, L.—Lins, R. D.—Lima, R.—Franca, G.—Simske, S. J.—Favaro, L.: A Context Based Text Summarization

System. 11th IAPR International Workshop on Document Analysis Systems (DAS), 2014, pp. 66–70, doi: 10.1109/DAS.2014.19.

[5] Fox, C.: A Stop List for General Text. ACM SIGIR Forum, Vol. 24, 1989, No. 1-2, pp. 19–21, doi: 10.1145/378881.378888.

[6] FS: Free Summarizer. 2016, http://freesummarizer.com/.

[7] Hassel, M.—Dalianis, H.: SweSum – Automatic Text Summarizer. 2016.

[8] Hingu, D.—Shah, D.—Udmale, S. S.: Automatic Text Summarization of Wikipedia Articles. Proceedings of International Conference on Communication, Information and Computing Technology (ICCICT 2015), 2015, pp. 15–18, doi: 10.1109/ICCICT.2015.7045732.

[9] Hovy, E.—Lin, C.-Y.—Zhou, L.—Fukumoto, J.: Automated Summarization Evaluation with Basic Elements. Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006, pp. 899–902.

[10] Kress, J. E.—Fry, E. B.: The Reading Teacher's Book of Lists. 6th edition, 2015.

[11] León, J. A.—Olmos, R.—Escudero, I.—Jorge-Botana, G.—Perry, D.: Exploring the Assessment of Summaries: Using Latent Semantic Analysis to Grade Summaries Written by Spanish Students. Procedia – Social and Behavioral Sciences, Vol. 83, 2013, pp. 151–155, doi: 10.1016/j.sbspro.2013.06.029.

[12] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Vol. 1, 2004, pp. 25–26.

[13] Lin, C.-Y.: Looking for a Few Good Metrics: Automatic Summarization Evaluation – How Many Samples Are Enough. Proceedings of the NTCIR Workshop-4, 2004, pp. 1765–1776.

[14] Liu, F.—Liu, Y.: Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, 2010, No. 1, pp. 187–196.

[15] Liu, K.—Terzi, E.—Grandison, T.: ManyAspects: A System for Highlighting Diverse Concepts in Documents. Proceedings of the 34th International Conference on Very Large Data Bases (PVLDB), Vol. 1, 2008, No. 2, pp. 1444–1447, doi: 10.14778/1454159.1454196.

[16] Lloret, E.—Palomar, M.: Text Summarisation in Progress: A Literature Review. Artificial Intelligence Review, Vol. 37, 2012, No. 1, pp. 1–41, doi: 10.1007/s10462-011-9216-z.

[17] Louis, A.—Nenkova, A.: Automatic Summary Evaluation Without Human Models. TAC, 2008.

[18] Louis, A.—Nenkova, A.: Automatically Assessing Machine Summary Content Without a Gold Standard. Computational Linguistics, Vol. 39, 2013, No. 2, pp. 267–300.

[19] Masoumi, S.—Feizi-Derakhshi, M.-R.—Tabatabaei, R.: TabSum – A New Persian Text Summarizer. Journal of Mathematics and Computer Science, Vol. 11, 2014, pp. 330–342.

[20] Molina, A.—Torres-Moreno, J.-M.: El Test de Turing para la Evaluación de Resumen Automático de Texto. Linguamática, Vol. 7, 2015, No. 2, pp. 45–55 (in Spanish).

[21] Nenkova, A.—McKeown, K.: Automatic Summarization. Foundations and Trends in Information Retrieval, Vol. 5, 2011, No. 2-3, pp. 103–233, doi: 10.1561/1500000015.

[22] OST: Online Summarize Tool. 2016, `https://www.tools4noobs.com/summarize/`.

[23] Passonneau, R. J.—Nenkova, A.—McKeown, K.—Sigelman, S.: Applying the Pyramid Method in DUC 2005. Proceedings of the Document Understanding Conference (DUC), Vancouver, BC, Canada, 2005, pp. 1–8.

[24] Porter, M. F.: An Algorithm for Suffix Stripping. Program, Vol. 14, 1980, No. 3, pp. 130–137, doi: 10.1108/eb046814.

[25] Saggion, H.—Torres-Moreno, J.-M.—da Cunha, I.—SanJuan, E.—Velázquez-Morales, P.: Multilingual Summarization Evaluation Without Human Models. Coling, 2010, Poster Volume, pp. 1059–1067.

[26] Sankarasubramaniam, Y.—Ramanathan, K.—Ghosh, S.: Text Summarization Using Wikipedia. Information Processing and Management, Vol. 50, 2014, No. 3, pp. 443–461, doi: 10.1016/j.ipm.2014.02.001.

[27] Steinberger, J.—Ježek, K.: Evaluation Measures for Text Summarization. Computing and Informatics, Vol. 28, 2009, No. 2, pp. 251–275.

[28] Torres-Moreno, J. M.—Saggion, H.—da Cunha, I.—SanJuan, E.—Velázquez-Morales, P.: Summary Evaluation with and Without References. Polibits Research Journal on Computer Science and Computer Engineering and Applications, Vol. 42, 2010, pp. 13–19, doi: 10.17562/PB-42-2.

[29] Torres-Moreno, J. M.—Saggion, H.—da Cunha, I.—Velázquez-Morales, P.—SanJuan, E.: Évaluation Automatique de Résumés Avec et Sans Référence. TALN 2010, Montréal, Canada, 2010, Vol. 1, pp. 19–23 (in French).

[30] Ubul, A.—Atlam, E.-S.—Kitagawa, H.—Fuketa, M.—Morita, K.—Aoe, J.-I.: An Efficient Method of Summarizing Documents Using Impression Measurements. Computing and Informatics, Vol. 32, 2013, No. 2, pp. 371–391.

[31] Yatsko, V. A.—Vishnyakov, T. N.: A Method for Evaluating Modern Systems of Automatic Text Summarization. Automatic Documentation and Mathematical Linguistics, Vol. 41, 2007, No. 3, pp. 93–103, doi: 10.3103/S0005105507030041.

Alan Ramírez-Noriega acquired his Master's degree in applied informatics at Universidad Autónoma de Sinaloa in 2014 and his Ph.D. degree in computer science from the Universidad Autónoma de Baja California in 2017. The main areas of interest are intelligent tutoring systems, knowledge representation, and text mining.

Reyes Juárez-Ramírez received his Master's degree in computer science from the Scientific Research and Higher Education Center in Ensenada in 2000, and his Ph.D. degree in computer science from the Universidad Autónoma de Baja California in 2008. He is currently Professor and Researcher at the Faculty of Chemical Sciences and Engineering, Autonomous University of Baja California. He has two main areas of interest: software engineering and human-computer interaction.

Samantha Jiménez is Ph.D. student at Universidad Autónoma de Baja California, Tijuana, México. She received her Bachelor's degree in 2011 in computer systems and her Master's degree in engineering in 2013 from the University of Colima, Colima, México. Her research interests are in the areas of human computer-interaction, dialogue systems, affective computing, multi-agent systems and evolutionary computing.

Sergio Inzunza received his Master's degree in computer science from the Autonomous University of Baja California in México in 2014. He is currently Ph.D. student in computer engineering, were he focused on creating tools for developers to model user and context information as a way to improve recommender systems.

Yobani Martínez-Ramírez received his Master's degree in computer science from Centro Investigación Científica y de Educación Superior de Ensenada (CICESE) and his Doctorate of Educational Technology from Centro Universitario Mar de Cortés. Since 1997 he has worked with the Universidad Autónoma de Sinaloa (UAS). He is currently Professor and Full-Time Researcher at the Faculty of Engineering of the UAS Mochis with PROMEP profile recognition (teacher improvement program) by the Ministry of Education (SEP). The areas of generation and application of knowledge of interest are implementing innovative systems and educational technology.