Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 12, pp.63216-63223, December, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

CASSIOPEIA MODEL AS AUTOMATIC SUMMARY EVALUATOR

*,1Luís Henrique Gonçalves de Aguiar and 2Marcus Vinícius Carvalho Guelpeli

1Graduate Program in Education –PPGEd, Federal University of Vales do Jequitinhonha and Mucuri, Brazil
2Department of Computing – DECOM, Federal University of Vales do Jequitinhonha and Mucuri, Brazil

ARTICLE INFO

Article History:
Received 18th September, 2017
Received in revised form
23rd October, 2017
Accepted 14th November, 2017
Published online 31st December, 2017

Key words:
Automatic Text Summarization,
Evaluation, Cassiopeia, ROUGE.

ABSTRACT

This paper proposes the use of the Cassipeia model as a new method for evaluating automatic summaries. Summary evaluation is an important task in the field of automatic text summarization, of which the most intuitive approach is conducted by human assessment. However, manual evaluation is both expensive and time consuming, therefore impractical. These difficulties led the researchers to seek automatic methods of evaluation. ROUGE is currently the most commonly used tool in the field, but each evaluated source text requires a human abstract, which also renders evaluation costly and limited. Simulations conducted in this study revealed that the evaluation performed by the Cassiopeia model is similar to the evaluation performed by the ROUGE tool; on the other hand, the use of the Cassiopeia model as an evaluator presented a few advantages, mainly the lack of human-made abstract in the evaluation processand independence from the domain and language.

Citation: Luís Henrique Gonçalves de Aguiar and Marcus Vinícius Carvalho Guelpeli, 2017. "Cassiopeia model as automatic summary evaluator", *International Journal of Current Research*, 9, (12), 63216-63223.

INTRODUCTION

With the appearance of a worldwide network of computers and its generalized use, the volume of information increases exponentially, thus rendering it increasingly hard to assimilate content of interest. A study conducted by the research team of the International Data Corporation (IDC) in 2014 showed that, until 2020, digital information will expand from 4.4 trillion gigabytes to 44 trillion gigabytes. The amount of digital information in 2020 is expected to be ten times greater in relation to 2013 (Jorge, 2015). Much of this information is represented by textdocuments, considering that this is the most natural way of storing information. Many of these documents are available in various documentary collections, such as: articles, books, blogs, websites, virtual environments, e-mails, magazines and journals, technical reports and others. In this context, tasks that are able to transform large amounts of documents into useful and organized knowledge become necessary. A solution, or at least an alternate path, to minimize this problem is to reduce the volume of available information throughgenerating abstracts from the original texts. In order to avoid dealing with large volumes of data, people tend to look for smaller, more compact versions: summaries. Creating abstracts by humans, according to Pardo (2007), is very laborious especially when dealing with a large volume of information. Since it is necessary to read and interpret the entire text so that its main ideas are extracted. Therefore, we are looking for automatic forms to produce these abstracts, an area that studies this topic is called Automatic Summarization (SA) of texts, which is a sub-area of Natural Language Processing (PLN) research. Automatic summarization is a task that consists in the automatic production of abstracts from one or more source texts, where the summary must contain the most relevant information of the source text (Garay, 2015). According to Pardo (2007), manmade abstracts are laborious and time-consuming, especially when dealing with a large volume of information, because it requires reading and interpreting the whole text in order to extract itsmain idea. In light of this, automatic forms are sought to generate these abstracts in the field called Automatic Text Summarization (AS), which is a sub-field of Natural Language Processing (NLP) research. Automatic summarization is a task that consists in the automatic generation of abstracts from one or more source texts, in which the summary must contain the most relevant information of the source text (Garay, 2015). One of the biggest challenges of the ASfield is generating summaries that preserve the most relevant information of the source text and, in order to do so, summaries must be evaluated. Abstract evaluation is an important task in the field of automatic text summarization (AS) and the most intuitive approach is abstract evaluation conducted by humans. However, manual evaluation is expensive and the results obtained are subjective and difficult to reproduce, for different evaluators may use different standards to evaluate the same summary. Difficulty in evaluation led researchers to look for automated methods to evaluate abstracts. Although automatic

Corresponding author:* **Luís Henrique Gonçalves de Aguiar,
Graduate Program in Education –PPGEd, Federal University of Vales do Jequitinhonha and Mucuri, Brazil.

evaluation does not suffer the drawbacks of manual evaluation, it still requires further research to become robust enough for a thorough evaluation.

Automatic evaluation measures were proposed, of which the best known and most widely used tool, according to Jorge (2015), is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) proposed by Lin (2004). The principle of ROUGEis to compare the number of shared words between the automatically produced summary and reference summaries, considered the ideal abstract, produced manually by a human expert. A limiting factor of ROUGE is the use of a human reference summary. This requirement imposes language and domain restrictions, often prevents the evaluation of large numbers of summaries, and requires a very time-consuming and expensive human labor. Studies in the field of automatic text summarization seek to improve its methods in order togenerate summaries that are increasingly similar to abstracts human-written. The evaluation of these automatic summaries plays an important role in this evolution, since improving quality in the production of these abstracts relies on efficient evaluation methods, in which it is possible to verify the use of the AS system and its suitability to specific tasks, as well as compare the results of different summarization methods. Considering the limitations of automatic summarization, both by human evaluation and by ROUGE, this paper presents the Cassiopeia model (Guelpeli, 2012) as an alternative for the evaluation of automated summaries. Cassiopeia is a text organizer, independent of language and domain. It uses automatic text summarization in the pre-processing step, in which the quality ofclustering is positively influenced according to the quality of the summarization. The model, unlike the ROUGE tool, does not need a human reference summary, whose use implies more agility and lower costs. Furthermore, Cassiopeia automatically generates Recall and Precision metrics, the same metrics used in ROUGE automatic summary evaluation. In view of the Cassiopeia model performance and considering the advantages of using the latter over the ROUGE tool, such as the use of human abstract in the evaluation process, and independent of domain and language. This paper proposes a systematic study that compares ROUGE and Cassiopeia models, with the purpose of showing that the Cassiopeia model can be used as an automatic summarizer and the evaluation performed by Cassiopeia is similar to the evaluation performed by the ROUGE tool.

Automatic Summarization

Automatic summarization is a sub-field within the Natural Language Processing (NLP) field, which aims to generate summaries from one or more source texts (Pardo, 2007). According to Garay (2015), the goal of AS is to recover the most important content of the original text and present it to the end user. Moreover, it is necessary to organize this information so that the main ideas in reference to the source text are maintained. Summaries are resources that are present in our daily lives. They are used to influence decisions when buying a book, reading a newspaper article, or even choosing which movie to watch in the theater. Summaries function as information sources for doubting readers about the title of the document, and they need to know if the text contains interesting reading material. Additionally, it helps readers who are interested in reading the whole text by offering them a preliminary view of the content, or even those who are interested in becoming familiar with a part of the research, by renderinginformation in an abbreviated way and thus saving time (Santos, 1996). The first studies in the AS field appeared about sixty years ago and deserve to be mentioned due to their relevance. One of the main studies is the keyword method by Luhn (1958), using word frequency to define the most important terms of the text. This principle is used until this day, as one wishes to select the most important words and sentencesin a text. Subsequently, other pioneering studies contributed to the field, such as: the relevance of sentence position in the text, highlighting the words contained in the title, in the first and last sentence of the text (Baxandale, 1958); the use of four types of lexical content indicators to identify relevant documents in a repository (Rath, 1961); application of other characteristics to select relevant sentences, such as the presence of pragmatic words, the occurrence of words from titles and headings and the sentence position in the text (Edmundson, 1969); the need to restrict domains to improve method results (Pollock; Zamora, 1975); classification of automatic summaries according to function, granularity, summarization technique, intended audience and number of documents to summarize (Hutchins, 1987); use of symbolic knowledge and statistical techniques for summarization (Hovy; Lin, 1999). Recent studies reveal the inclusion of new research approaches in the field of automatic text summarization. Fattah (2014) proposes a trainable automatic summarizer that uses a hybrid method for sentence selection. Several characteristics are taken into account, for instance: word similarity between sentences, word similarity between paragraphs, text format, term frequency, and sentence location. In Rocha and Guelpeli (2017), PragmaSUM was proposed, an automatic summarizer, independent of language and domain, which is based on the frequency of words in the text for sentence valuation. Yao (2015) proposes document summarization in several languages, thus creating a summary in a target language from text documents written in a different source language. In turn, He (2017) conducted a study exploring the temporal social context for twitter summaries, taking into account the difficulty in performing the summarization of tweets due to their size and their unstructured writing form. According to Gambhir and Gupta (2017), research in the field of automatic text summarization in recent years sought to improve existing approaches and developing new methods that produce higher quality summaries. Despite these efforts, the performance of automatic summarizers is still moderate and the quality of generated abstracts does not match that of a human made abstract.

Automatic summary evaluation

Summary evaluation is an increasingly studied and investigated subject due to difficulties in performing an adequate assessment, as explained by Pardo (2007). There may be summaries that are built differently and still be considered of good quality or suitable for a particular application. Human evaluation, which is usually most commonly used and considered most adequate, is expensive, time-consuming and prone to error. An alternative to this problem is automatic evaluation, which does not suffer the disadvantages of human assessment, but is still undergoing further research to become sufficiently robust for a complete evaluation. International conferences dealing with the evaluation of automatic summaries are held to promote research development. The TIPSTER Text Summarization Evaluation (SUMMAC) was held in the 1990s and was the first conference about the field. Other important conferences in the field of natural language processing that also deal with the evaluation of automatic summaries are the Document Understanding Conference (DUC), which was held from 2001 to 2007 and later replaced by the Text Analysis Conference (TAC). The TAC is an international conference consisting of a series of evaluation workshops that are held to promote research in the field of NLP and similar fields (Gambhir and Gupta, 2017). The most widely known and widely used measure according to Dias (2016) is the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) proposed by Lin (2004). The ROUGE tool is an evaluation metrics package that uses metrics to compare the amount of information shared between the automatically produced summary and its reference summary. The comparison is performed by overlapping sentences and counting *n-grams* between the automatic summary and the reference summary. The *n-gram* in the ROUGE measure, according to Tosta (2014), is considered a word or set of words occurring in sequence. According to Rino and Pardo (2007), this process is performed completely automatically, however, the use of ROUGE implies the construction of manual summaries, which demand human and financial resources. The metrics employed to compare the contents of the automatic summary withinformation from the reference summary are Recall metrics and Precision metrics. The Recall (R) metric, Equation 1, indicates how much of the reference summary is present in the automatic summary. In contrast, the Precision (P) metric, Equation 2, indicates how much of the automatic summary overlaps with the reference summary. In other words, Precision reveals the amount of information in the ideal summary that is in the automatic summary, while Coveragereveals the amount of information in the reference summary that was covered by the automatic summary (Jorge, 2015).

$$\mathrm{R} = \frac{\mathrm{NSA}\ \in \mathrm{NSR}}{\mathrm{NSR}} \tag{1}$$

$$\mathrm{P} = \frac{\mathrm{NSA}\ \in \mathrm{NSR}}{\mathrm{NSA}} \tag{2}$$

Whereas $NSA$is the number of sentences in the automatic summary and *NSR* is the number of sentences in the references summary. The precision and coverage metrics are complementary; for this reason, the *F-Measure* (F) is used, Equation 8, which represents the harmonic mean between Coverage and Precision metrics. F-Measure results vary in an interval of [0 1], in which the closer to 1, the better the summary evaluation.

$$\mathrm{F} = 2 * \frac{\mathrm{R} * \mathrm{P}}{\mathrm{R} + \mathrm{P}} \tag{3}$$

In addition to the ROUGE tool, some other works were proposed to evaluate automatic summaries. The pyramid method proposed by Nenkova and Passonneaum (2004) considers the frequency with which information occurs simultaneously in the reference summary as the principle to identify the most important information of a text. These text fragments that are considered relevant are referred to in this method as summary content units (SCU). Thus, the SCU, which is contained in a greater number of reference summaries, is granted more weight and is positioned in a pyramid according to its score, as higher scores are placed on top. The automatic summaries are evaluated comparing them with the SCUs and are considered more informative those which have the largest number of SCU near the top of the pyramid. The Basic Elements (BE) method (Traz; Hovy, 2008) conducts sentence segmentation of the reference summary into small content units, using a manual process. In light of these content units, a human judge evaluates the percentage of information that the automatic summary was able to cover in relation to the reference summary. The AutoSlummENG summary evaluator (AUTOmatic Summary Evaluation based on N-gram Graphs) (Giannakopoulos et al., 2008) performs the evaluation automatically. The evaluation is based on comparing the graphical representation of the n-grams of the automatic summaries with the reference summaries. AutoSlummENG is independent of language due to its statistical characteristics.

Corpus

The term corpus is used to refer to a collection of electronically stored written text documents processed by a computer with linguistic research objectives. When building a text corpus, one must make a selection of representative data, i.e., which consists of a corpus of linguistic evidence that can support generalizations and can be used to test hypotheses. According to Oliveira and Guelpeli (2014), the word corpus (plural, corpora) originates from the Latin word for body, a set of texts, that inCorpus Linguisticsdetermines a collection of selected and organized texts. According to Sardinha (2000), Corpus Linguistics is responsible for the collection and exploration of corpora, or a set of judiciously collected data with the aim of being used for research of a language or linguistic variety. For Aluísio and Almeida (2006), a computerized corpus observes a set of considerations that influences the validity and reliability of research based on a corpus. Its creation is a repetitive process, starting with text selection, based on significant criteria for research (external criteria), follows with empirical investigations of language, or linguistic variety under analysis (internal criteria) and concluding with a review of the whole project.

Cassiopeia Model

The Cassiopeia Guelpeli model (2012) is a hierarchical text clusterer that conducts clustering in textual bases of different domains. The structure of Cassiopeia consists of three macro stages: pre-processing, processing and post-processing (Figure 1). The model is composed of two main processes: summarization in the pre-processing stage, which aims to decrease the number of words, maintaining the most important information, clustering in the processing stage, which conducts clustering according to text similarity.

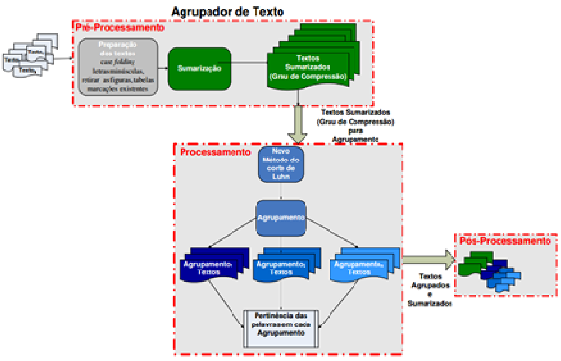
Agrupador de Texto
Pré-Processamento
Texto
Preparação dos textos case folding letrasminúsculas, retirar as figuras, tabelas marcações existentes
Sumarização
Textos Sumarizados (Grau de Compressão)
Textos Sumarizados (Grau de Compressão) para Agrupamento
Processamento
Novo Método do corte de Luhn
Agrupamento
Agrupamento Textos
Agrupamento Textos
Agrupamento Textos
Pertinência das palavras em cada Agrupamento
Textos Agrupados e Sumarizados
Pós-Processamento

Figure 1. Model Cassiopeia (Guelpeli, 2012)

The Cassiopeia process begins with texts inputs at the preprocessing stage, in which the texts are cleaned and prepared for computer processing. The main goal in this stage is to reduce word volume, thus gaining qualitative and quantitative improvement for processing. For this, the use of the summarization technique was proposed in the preprocessing stage, thus reducing dimensionality and data spaces. The application of this technique bestowed the model with an advantage in quality and processing, and allowed the use of stopwords in the texts, making the Cassiopeia model independent of language. After the pre-processing stage is concluded, the processing stage begins, which uses the clustering process to perform the clustering of the texts according to similarity. Due to the proposed text summarization, it was possible to create a different solution for the Cassiopeia model to perform the processing step. Luhn (1958) proposed the Luhn cut-off grids (Figure 2), in which an upper limit and a lower cut limit applied to the zipf curve (ZIPF, 1949) were defined.

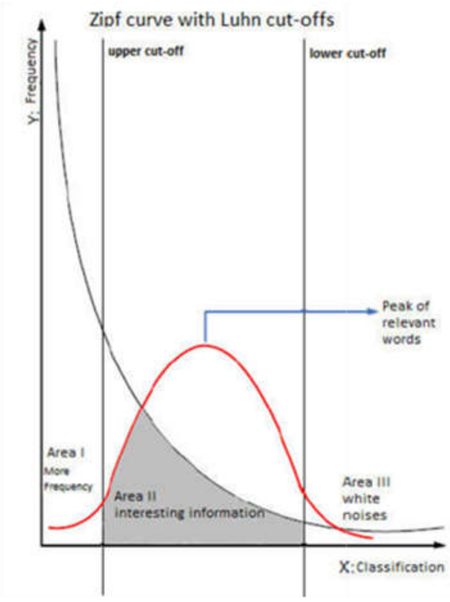

Figure 2. Zipf Curve with LuhnCut-Offs (Guelpeli, 2012)

The Cassiopeia model defines a new method based on the Luhn cut-off, in which a medium cut-off is proposed in the distribution of word frequency (Figure 3). In order to render this Luhn cut-off variation viable, centroids were used, as a form of representing the sample space, and for organizing texts in clusters, the hierarchical agglomerative method and the Cliques algorithm, in order to guarantee similarity between clustered texts.

The model uses relative frequency (RF) as an instrument to characterize and evaluate the relevance of a set of words in the document that will be clustered. Relevancy is defined according to the frequency it is found in the document. Based on the weight of the words, obtained in the relative frequency, the average of the total words in the document is calculated. In this step, the model uses truncation, with a maximum size of 50 positions for the word vectors (Figure 2), creating a cut-off that represents the average frequency of the words obtained with the calculations, and then organizes the vectors of words (GUELPELI, 2012). In the post-processing stage the model presents a hierarchical structure with texts grouped according to similarity and summarized as output. Thus it provides better evaluation in comparison to other text clusters. Due to summarization, the texts have a much smaller number of sentences with a high degree of informality. According to Guelpeli (2012) the pre-processing step directly influences the clustering step, and the better the quality of the summarization algorithm, the better the text clusters in the Cassiopeia model. This way, the model proved to be a viable solution in the evaluation of automatic summaries. Moreover, Cassiopeia automatically generates Recall and Precision metrics, the same metrics used in the evaluation of automatic summaries by ROUGE.

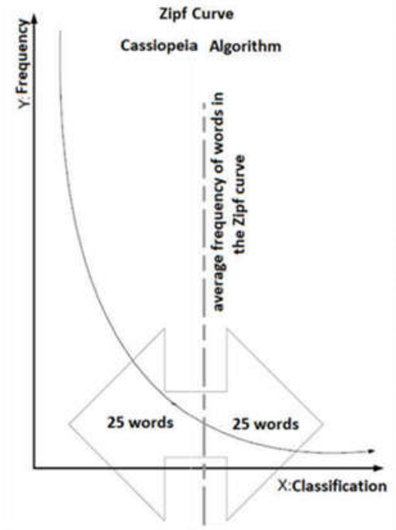

Figure 3. Word selection of the cut-off (Guelpeli, 2012)

MATERIALS AND METHODS

This paper proposes the application of the Cassiopeia model (Guelpeli, 2012), as an automatic summarizer. The experiments developed during this research can be organized into three basic stages: Corpus creation, summarization, and automatic summary evaluation. In the first stage, we used the educational corpus (Aguiar; Rocha; Guelpel, 2017) built for this research. The educational corpus was created using the methodological steps for compiling a corpus defined by Aluísio and Almeida (2006). It is composed of scientific articles in Portuguese in the Education domain, divided into 10 categories based on the classification by the governmenthigher education institute (CAPES) for the general field of Education. The selection of scientific articles composing the corpus was justified by the need to obtain a reference summary for each text written by a specialist to be used in the ROUGE tool. These will be used in summarization simulations and automatic summarization evaluations in this paper and in studies by the Research Group on Text Mining and Natural Language Processing and Machine Learning (Mineração de Textos e Processamento de Linguagem Natural e Aprendizado de Máquina - MTPLNAM). Scientific articles meet these requirements through the abstract written by the author. The articles were selected from the Scientific Electronic Library on-line (SCIELO[1]) database. The corpus consists of 500 articles, 50 articles for each category. Statistical information was calculated using the software *FineCount 2.6 free*[2]. The corpus consists of 2,999,646 words in total and Table 1 condenses the statistics of the texts separated in the 10 categories that compose the corpus.

[1] Available at: http://www.scielo.br/

[2] Available at: http://www.tilti.com/software-for-translators/finecount/

Table 1. Corpus statistics created for the experiments

Categories	Words	Words and Numbers	Sentences	Average words per text
Special Education (SE)	274182	281052	10267	5483,64
Continuing Education (CE)	281035	285382	14274	5620,7
Preschool Education (PS)	281091	286480	23248	5621,82
Teaching-Learning (TL)	275404	279809	13299	5508,08
Philosophy of Education (PE)	313779	317365	18554	6275,58
History of Education (HE)	392515	400849	14646	7850,3
Education Politics (EP)	371540	379529	12607	7430,8
Psychology of Education (PD)	266666	271889	13157	5333,32
Sociology of Education (SE)	329689	334182	17144	6593,78
EducationalTechnology (ET)	213745	218345	7688	4274,9
Total	2999646	3054882	144884	59992,92
General Average	299964,6	305488,2	14488,4	5999,29

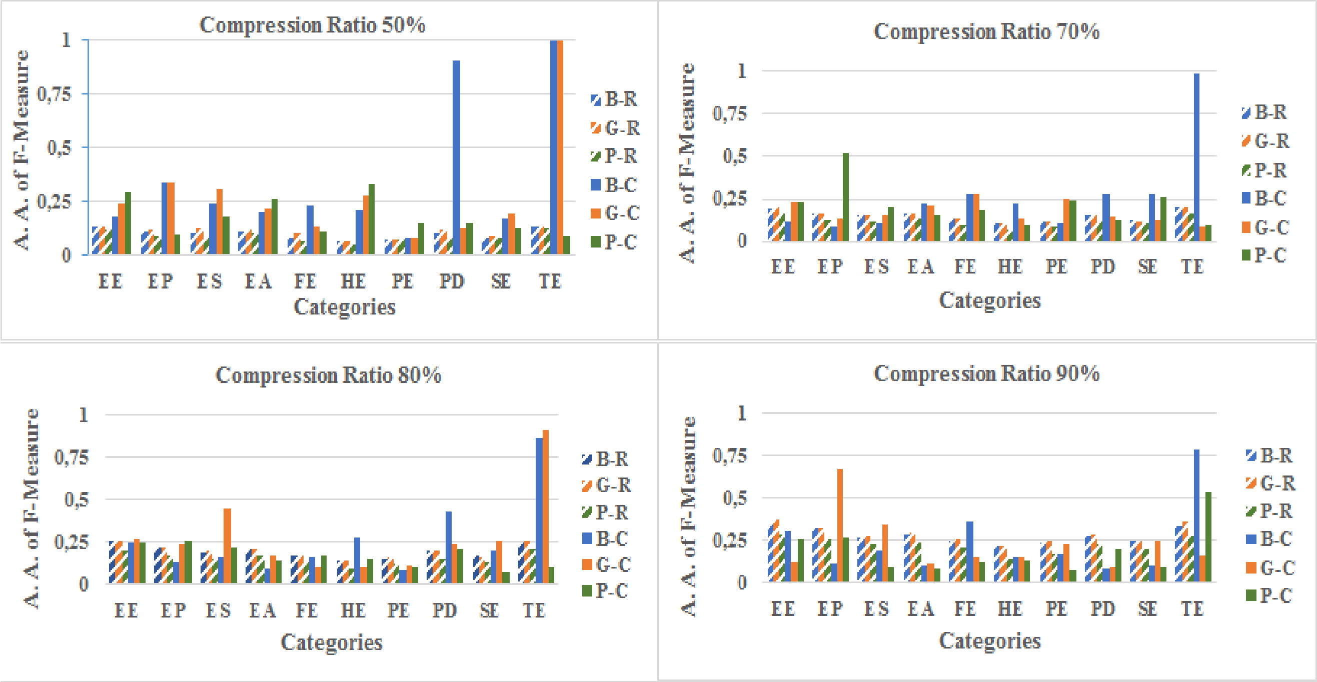

Figure 4. Comparison of accumulated average from F-Measure

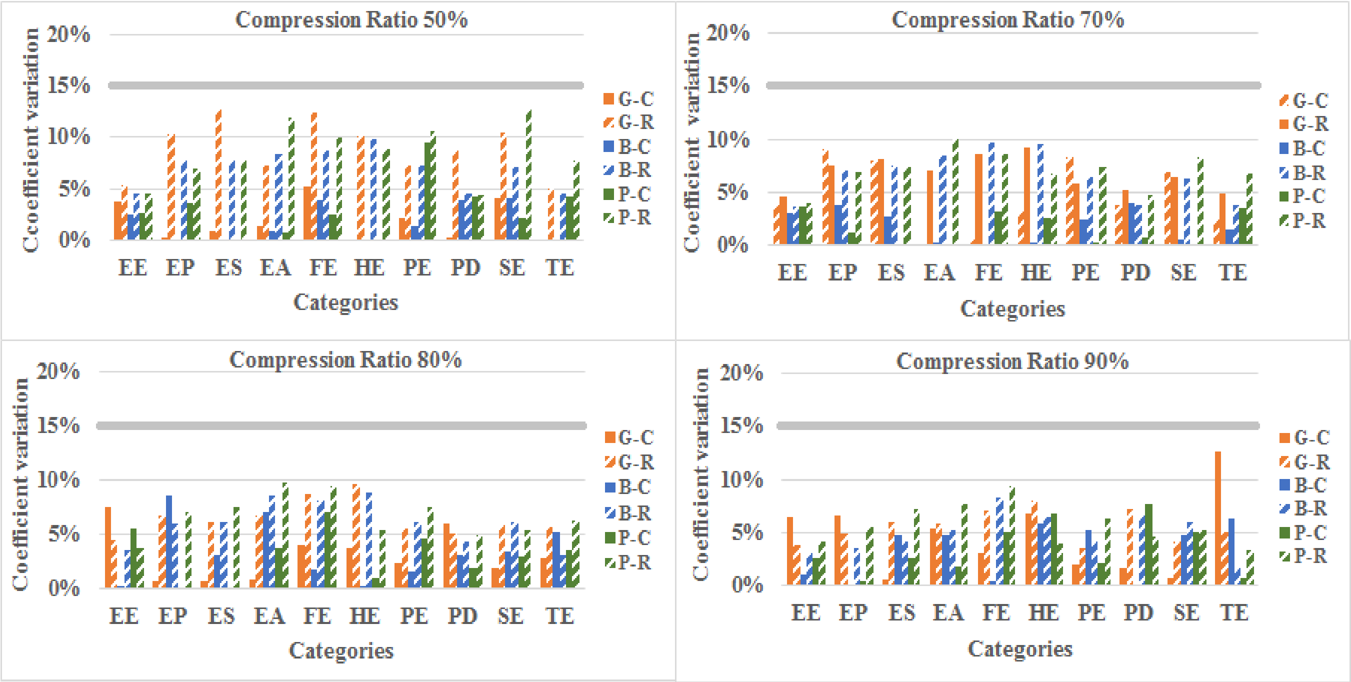

Figure 5. Comparison of Coefficients of Variation

In the automatic summarization process, we used three summarizers found in academic literature: BLMsumm (Oliveira; Guelpeli, 2011), GistSumm (Pardo, 2002) and PragmaSUM (Rocha, Guelpeli, 2017). The selection of these summarizers was based on some of the criteria of this research, such as the free use of the tool and the ability to perform text summarization in Portuguese. Another criterion was the possibility of defining compression percentages that allowed a compression range of 50% to 90%. The summarization of the corpus was performed using four compression rates: 90%, 80%, 70% and 50%, thus summaries were generated with the corresponding values of 10%, 20%, 30% and 50% compression, respectively, in relation to the original text. Thus, we summarized the 500 source texts of the 10 corpus categories, using the 4 compression rates and the 3 summarizers. The results of the summarization process showed that a sum of 6 thousand automatic summaries were obtained.

The automatic summary evaluation process was performed by the ROUGE tool and the Cassiopeia model. Six thousand automatic summaries were generated in the process of summarizing the Education corpus, divided into 10 categories and undergoing 4 compression rates 90%, 80%, 70% and 50%. The evaluations were carried out in batch, in both evaluators, including the automatic summaries of the same category and compression ratio in each batch. The metrics obtained in the evaluation process were Recall, Precision and F-Measure. For each evaluation process batch a 50-fold repetition was performed, in order to obtain results with less imprecision. These metrics were measured in the ROUGE tool by comparing the reference summary written by the author and the automatic summary. In turn, the metrics were obtained in the Cassiopeia model, through text clustering. The metric used to compare the evaluations performed by the ROUGE tool and the Cassiopeia modelwas F-Measure, due to its importance. The results of this metric will be presented through its accumulated average. This will favor the interpretation and comparison of the results. Result dispersions were also calculated to perform comparisons. The dispersion has the purpose of measuring the degree of variability of the values around the mean average. The metric used to calculate the result dispersion of both evaluators was the coefficient of variation (CV). The coefficient of variation is a relative dispersion measure useful for comparing the concentration degree in relative terms. It can be calculated through a non-negative data set (Equation 4), expressed as a percentage, and describes the standard deviation relative to the mean average, where the lower the variation coefficient value, the more cohesive the data will be (Triola, 2014).

$$CV = \frac{\sigma}{\bar{X}} * 100\% \qquad (4)$$

Where CV is the coefficient of variation,$\sigma$is the standard deviation of data from the series and $\bar{X}$is the data average. For calculating the coefficient of variation, the 50 *F-Measure* values generated in the batch evaluation by Cassiopeia and ROUGE was used in each evaluation.

RESULTS

The evaluation process of automatic summaries was conducted by the ROUGE tool and by the Cassiopeia model, making use of the automatic summaries generated in the summarization stage of the Education corpus. The calculations of the mean accumulated averages from *F-Measure* and the result dispersions were organized according to the four compression rates (50%, 70%, 80% and 90%), divided into ten categories that compose the Education corpus and three summarizers, PragmaSUM, GistSumm and BlmSumm, which are being evaluated. Figure 4 presents the comparison of accumulated averages in *F-Measure* and Figure 5 presents the comparison of dispersion percentage. In the presentation of results, each summarizer will be represented by the first letter of their name along with the first letter of the evaluator. For example, the PragmaSUM summarizer evaluated by Cassiopeia and ROUGE will be represented, respectively, by: “P-C” and “P-R”. As for GistSumm summarizer, it will be represented by: “G-C” and “G-R”, and the BlmSumm summarizer by: “B-C” and “B-R”.

DISCUSSION

Results show that there are no large differences in the values of the accumulated means of the F-Measure obtained in the evaluations carried out by the Cassiopeia model and the ROUGE tool. When analyzing the results of Cassiopeia and ROUGE in isolation, it is possible to verify that the values obtained for each summarizer within a category present close results. This pattern can be observed in most categories and compression ratios. When analyzing the result dispersions, it was found that all the results of the coefficient of variation for ROUGE and Cassiopeia evaluations in the ten categories of the Education corpus, summarized by the three summarizers, in 4 compression ratios,obtained values lower than 15%. The results show that the ROUGE and Cassiopeia model evaluations obtained homogeneous results, i.e., with low dispersion. The highest dispersion coefficient achieved in the results was 13% under the compression ratios of 50% and 90%. With a 50%compression ratio, the categories Preschool Education and Sociology of Education for GistSumm and PragmaSUMsummarizers, respectively, obtained this value in the ROUGE evaluation. At a 90% compression level, this value was obtained by the category Educational Technology with the GistSumm summarizer in the Cassiopeia evaluation. With compression of 70% and 80%, the highest CV value was 10% in the ROUGE evaluation. Viewing the result dispersion obtained by Cassiopeia and ROUGE separately, it is possible to notice that Cassiopeia obtained results containing less dispersion. As a parameter the number of results that reached a coefficient of variation equal to or lower than 5%, it is possible to identify that 97% of the results of Cassiopeia and 23% in the tool Rouge obtained this value with 50% compression. At the compression ratio of 70%, 87% of the Cassiopeia results and 27% of the Rouge presented dispersion of less than or equal to 5%. At the 80% compression level, 83% of the Cassiopeia results and 33% in the Rouge tool obtained a CV of 5% or less. Finally, in the compression ratio of 90%, 73% of the results of Cassiopeia and 56% of the Rouge achieved this number. The results reveals that evaluations by the ROUGE tool and the Cassiopeia model obtained results with low dispersion; in the Cassiopeia model, it generally obtained a lower coefficient of variation.

Conclusion

The use of the Cassiopeia model as an automatic summary evaluator showed that the evaluation carried out by the model compared to the ROUGE tool obtained similar results. Results revealed that there are no major differences in the values obtained in both processes. It can also be observed that the cumulative *F-Measure* mean for each summarizer within a category are also similar in value. This balance can be verified both in the evaluation carried out by the Cassiopeia model and in the evaluation performed by the ROUGE tool. The dispersion in the results, calculated by the coefficient of variation,reveals that the percentage of dispersion of the accumulated mean averages of the *F-Measure* for both evaluators is low, which shows that the evaluations carried out by Cassiopeia and ROUGE produced homogeneous results, all of which were lower than 15%. According to Martins (2011) and Correa (2003), if the value of the CV is less than or equal to 15%, the data is said to have low dispersion;if it varies between 15% and 30%, it is considered medium dispersion; above 30%, data is said to present high dispersion. When comparing the coefficients of variation of both evaluators, it is possible to verify that the Cassiopeia modelmostly achieved results containing less dispersion. The Cassiopeia model obtained, in 85% of the evaluation results, dispersion equal to or less than 5%. In contrast, the tool ROUGE obtained 35%. This comparison shows that although the two evaluators achieved results with low dispersion, Cassiopeia presented more cohesive results, which maintains the coherence of the clustering technique. The use of the Cassiopeia model as an automatic summary evaluator showed that the evaluation performed by the model compared to the ROUGE tool is similar. Furthermore, the application of Cassiopeia as an evaluator presented some advantages compared to the ROUGE tool. The main advantage is not using amanual summary in the evaluation process, which allows the evaluation of large numbers of summaries, presenting low costs and greater agility. Another advantage enjoyed by the Cassiopeia model is itsindependence of the domain and language, as shown in the studies by Guelpeli (2012). The Cassiopeia model uses summarization in the pre-processing stage to reduce the volume of words and reduce dimensionality, thus making it impossible to use the stopwords list, rendering the model independent of the language in which the text is written. In turn, domain independence occurs due to the adoption of a new method called Cassiopeia, used in the processing step of the Cassiopeia model. In light of the results obtained in this research, it was revealed that the evaluation performed by the Cassiopeia model is similar to the evaluation performed by the ROUGE tool. Moreover, considering the advantages in the use of Cassiopeia, it is possible to state that the Cassiopeia model can be used as an automatic summarizer.

Future Research

In view of the continuation of this work, future studies are suggested, such as comparing other automatic summarizer softwarewith the evaluations carried out by the Cassiopeia model and the ROUGE tool. Another future research suggestion is to carry out the comparison of the evaluation carried out by the Cassiopeia model with an evaluation performed by human specialists, which is considered the ideal evaluation. In order to more thoroughly investigate the application of the Cassiopeia model as an automatic summarizer, one of the subsequent steps is to apply it as an evaluator of an English-language corpus.

REFERENCES

Aguiar, L. H. G. D. A.; Rocha, V. J. C. and Guelpeli, M. V. 2017. Uma coleção de artigos científicos de Português compondo um *Corpus* no domínio educacional. *Plurais Revista Multidisciplinar*, v. 2, n. 1, p. 60–74.

Aluísio, S. M. and Almeida, G. M. D. B. 2006. O que é e como se constrói um *corpus* ? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. *Calidoscópio*, v. 4, n. 3, p. 155–177.

Baxendale, P. B. 1958. Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development*, v. 2, n. 4, p. 354–361, out.

Correa, S. M. B. B. 2003. Probabilidade e estatística. PUC Minas Virtuais, p. 116.

Dias, M. S. 2016. Investigação de modelos de coerência local para sumários multidocumento Investigação de modelos de coerência local para sumários multidocumento. São Carlos: Tese (Doutorado em Ciência de Computação e Matemática Computacional) - Universiade de São Paulo.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, v. 16, n. 2, p. 264–285.

Fattah, M. A. 2014. A hybrid machine learning model for multi-document summarization. Applied Intelligence, v. 40, n. 4, p. 592–600, 20.

Gambhir, M. and Gupta, V. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, v. 47, n. 1, p. 1–66.

Garay, Y. B. A. 2015. Sumarização multidocumento com base em aspectos informativos. São Carlos: Dissertação (Mestrado em Ciência de Computação e Matemática Computacional) - Universiade de São Paulo.

Giannakopoulos, G. *et al.* 2008. Summarization system evaluation revisited. ACM Transactions on Speech and Language Processing, v. 5, n. 3, p. 1–39.

Guelpeli, M. V. C. 2012. Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização. Niterói: Tese (Doutorado em Computação) - Univerisade Federal Fluminense.

HE, R. *et al.* 2017. Twitter summarization with social-temporal context. World Wide Web, v. 20, n. 2, p. 267–290.

Hovy, E. and Lin, C. 1999. Automated Text Summarization in SUMMARIST. Advances in Automatic Text Summarization, p. 18–24.

Hutchins, J. and John, 1987. Summarization: Some problems and Methods. (K. P. Jones, Ed.) Meaning: the frontier of informatics. Anais...London: Aslib.

Jorge, M. L. R. C. 2015. Modelagem gerativa para sumarização automática multidocumento. São Carlos: Tese (Doutorado - Programa de Pós-Graduação em Ciência de Computação e Matemática Computacional) - Universidade de São Paulo.

Lin, C. 2004. Looking for a Few Good Metrics : ROUGE and its Evaluation. Tokyo: Proceedings of NTCIR Workshop.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165.

Martins, G. A. 2011. Estatística geral e aplicada. 4. ed. São Paulo: Atlas.

Nenkova, A and Passonneau, R. 2004. Evaluating content selection in summarization: The pyramid method. Proceedings of HLT-NAACL, v. p. 145–152.

Oliveira, M. A. DE and Guelpeli, M. V. BL MSumm, 2011. Métodos de Busca Local e Metaheurísticas na Sumarização de Textos. Proceedings of the ENIA - VIII Encontro Nacional de Inteligência Artificial, p. 285–298.

Oliveira, R.R. and Guelpeli, M.V.C. 2014. Building a Corpus in Italian Written Language. In:6th International Conference on Corpus Linguistics (CILC2014). Las Palmas de Gran Canaria, Espanha.

Pardo, T. A. S. 2002. GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos. Technical Report. NILC-TR, v. 25p, p. 02–13.

Pardo, T. A. S. 2007. Sumarização Automática: Principais Conceitos e Sistemas para o Português Brasileiro. São Paulo: Relatório Técnico. Universidade de São Paulo - USP, Universidade Federal de São Carlos - UFSCar, Universidade Estadual Paulista – UNESP.

Pollock, J. J. and Zamora, A. 1975. Automatic Abstracting Research at Chemical Abstracts Service. *Journal of Chemical Information and Modeling*, v. 15, n. 4, p. 226–232.

Rath, G. J., Resnick, A. and Savage, T. R. 1961. The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. *American Documentation*, v. 12, n. 2, p. 139–141.

Rocha, V. JR. Cordeiro and Guelpeli, M. V. C. Pragmasum : automatic text summarizer based on user pr profile. *International Journal of Current Research*, v. 9, n. 7, p. 8, 2017.

Santos, M. Dos, 1996. The textual organization of research paper abstracts in applied linguistics. Text-Interdisciplinary Journal for the Study of, v. 16, n. 4, p. 481–499.

Sardinha, T. B. 2004. Linguistica de *Corpus*. DELTA: Documentação e Estudos em Linguística Teórica e Aplicada, v. 20, n. 2, p. 364–365.

Tosta, F. E. S. 2014. Aplicação de conhecimento léxico-conceitual na sumarização multidocumento multilíngue. São Carlos: Dissertação (Mestrado em Linguistica) Universidade Federal de São Carlos.

Tratz, S. and Hovy, E. 2008. Summarization evaluation using transformed basic elements. Proceedings of the 1st Text Analysis Conference.

Triola, M. F. 2014. Introdução à estatística: atualização da tecnologia. 11^{o} ed. Rio de Janeiro: Tradução e revião técnica Ana Maria Lima de Farias, Vera Regina Lima de Farias e Flores LTC.

Yao, J. G.; Wan, X. and Xiao, J. 2015. Phrase-based Compressive Cross-Language Summarization. Conference on Empirical Methods in Natural Language Processing. Anais...Lisboa, Portugal: Association for Computational Linguistics.

Zipf, G. K. 1949. Human behavior and the principle of least effort : an introduction to human ecology. Oxford: Addison-Wesley Press.

****