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

International Journal of Current Research
Vol. 9, Issue, 07, pp.53935-53942, July, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

PRAGMASUM: AUTOMATIC TEXT SUMMARIZER BASED ON USER PROFILE

***Valdir Jr. Cordeiro Rocha and Marcus Vinicius Carvalho Guelpeli**

Universidade Federal dos Vales do Jequitinhonha e Mucuri - UFVJM

ARTICLE INFO

Article History:
Received 06th April, 2017
Received in revised form
10th May, 2017
Accepted 19th June, 2017
Published online 26th July, 2017

Key words:

Automatic text summarization, Keywords, PragmaSUM, Luhn, Zipf's Law, ROUGE.

ABSTRACT

This article presents the automatic text summarizer PragmaSUM, which is independent from the language and knowledge domain of the source text, based on the Cassiopeia algorithm, which uses Luhn's distribution and Zipf's Law to select words in the text used for classifying sentences and generating the summary. A corpus is created for tests in Portuguese, composed of scientific articles from 10 different knowledge domains, for evaluating summaries generated by BLMSumm, GistSumm and PragmaSUM summarizers. Performance was observed using Recall, Precision and F-Measure metrics present in the assessment tool ROUGE. The end of the article presents the results of the summary assessment generated by the summarizers and PragmaSUM by employing two forms of summarization: with keywords for classifying sentences in the source text without using these words and by comparing summarizers. It was observed that using keywords in automatic text summarization allows for personalization of the summary according to the users' needs by fetching sentences that really correspond to their interest domain.

Citation: Valdir Jr. Cordeiro Rocha and Marcus Vinicius Carvalho Guelpeli, 2017. "PragmaSUM: automatic text summarizer based on user profile", *International Journal of Current Research*, 9, (07), 53935-53942.

INTRODUCTION

There are currently several sources of text information available online that are fetched in several ways and with different aims, thus rendering it impossible to assimilate everything. Selecting the ones that best correspond to public interest facilitates information processing and recovery (Guelpeli, 2012). With the large scale of information flow currently being generated, it becomes impossible to read every available text, since there is a limit to a person's ability and it requires a huge amount of time and effort. Another option is to present this information in condensed form, rendering it easier and quicker to read the complete text regarding content analysis and decision. Automatic Summarization (AS) is one of the areas dedicated to this field of research. According to Oliveira (2014), AS is a sub-field of Natural Language Processing (NLP), which automatically generates summaries that aim to reduce the volume of information without losing the original quality of the source text. Academic literature calls the system that conducts this task automatic summarization.

Oliveira (2014) highlights that, although it is currently possible to find several automatic summarizers, whether commercial or in literature, most are written in the English language. Considering the limitations found in the AS field, this article presents the creation of an automatic summarizer that is independent from the language and knowledge domain of the source text and is adaptable to the user's profile. The implemented summarizer, PragmaSUM (Pragmatic Summarizer), employs the word value method presented by Guelpeli (2012). PragmaSUM is an extractive automatic text summarizer; i.e., it defines what the main sentences of the text will be and selects them to compose the generated summary, without altering its composition. In order to measure PragmaSUM, a corpus was built in Portuguese, assessment tests were conducted for automatic summarizers, and the metrics Recall, Precision and F-Measure were applied using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) tool (Lin and Hovy 2003). This research aims to contribute by mainly presenting a model that improves the precision of generated automatic summaries. The article is divided into six sections. Section 2 presents the state of the art. The third section presents PragmaSUM and its method for sentence valuation. Section 4 presents the methodology used for creating the corpus and its statistics, as well as the summarizers used and metrics used in summary assessment. Section 5 describes the results achieved and the sixth section concludes the research and suggests future studies.

State of the art

According to Guelpeli (2012), AS has been formalized since 1950, and the initial milestone of this research was the method using Luhn's (1958) keywords. Also in 1958, Baxendale writes about the importance of the first and last sentence of the

Corresponding author:* **Valdir Jr. Cordeiro Rocha,
Universidade Federal dos Vales do Jequitinhonha e Mucuri - UFVJM

original text; in 1969, Edmundson addresses the computational choice of sentences as the greatest potential for transmitting meaning to the original text; in 1975, Pollock and Zamora reinforce the relevance of domain restriction; in 1987, Hutchins classifies summaries as indicative, informative, and critical; in 1993, Maybury suggests the use of a hybrid approach; in 1997, Marcu explores the rhetorical associations among sentences in the text; also in 1997, Hovy and Lin explore the use of symbolic knowledge and statistical techniques for summarization; in 1999, Sparck states that the taxonomy created by Hutchins (1987) is a contingent factor for establishing its applicability and creating a consistent assessment of this process. Nenkova and McKeown (2012) and Vanderwende *et al*. (2007) recently conducted studies using statistical methods combined with other text characteristics, such as word frequency, TF-IDF term weighting, position of sentences, relation between title and signal phrases, etc. Other approaches consider semantic associations between words and combine them with similar characteristics in the process of sentence similarity. Examples of these approaches are, among others, latent semantic analysis by Gong and Liu (2001), topic signatures by Lin and Hovy (2001), and sentence grouping by He, Qin, and Liu (2012). Nóbrega and Pardo (2016) propose enriching the summary using the text subject based on text segments. The main idea is that a text can be segmented into smaller ideas, or its subtopics, in order for each subtopic of the text to be represented by a text segment which is coherent with one or more sentences in the same line. According to Tabassum and Oliveira (2015), during recent years, AS research for sets of documents have attracted greater interest in graph-based approaches and based on topic Bayesian models. They argue that Bayesian models incorporate the concept of latent topics in n-gram language models. The research published in the AS field indicate two methodologies: the superficial method, which uses statistical processes, and the thorough method, which is composed of linguistic models. In addition to those methods, a hybrid approach is proposed that uses the former methods combined for AS. AS studies indicate two distinct categories for obtaining summaries: extractive summarization and abstractive summarization. According to TABASSUM and OLIVEIRA (2015), the extractive methods select a subset of words, phrases or sentences existing in the source text to compose the summary. Whereas the abstraction-based method creates a compact version by transmitting the summarized meaning of the source text. These methods build an internal semantic representation and then use natural language processing techniques in order to create summaries that are more similar to those produced by human beings.

Rouge

According to Oliveira (2014), ROUGE is an automatic package of summary assessment that compares the quality of summaries generated by automatic summarizers with those conducted by human beings. This tool is adopted in international conferences dedicated to the theme, such as the Text Analysis Conference (TAC), held annually in the United States of America and sponsored by the US Department of Defense. The use of automated assessment is justified by the enormous quantity of texts being analyzed and by the high costs they would ensue if conducted by specialists. In order to perform an assessment, ROUGE considers the number of n-grams, i.e., a given sequence of words, which share the summaries generated automatically and those created manually. The n-grams occupy an interval between 0 and 1 and, the closer to 1 is the result, the more the automatic summary is similar to the compared human summary (LUCHI and RIBEIRO 2011). Figure 1 shows a simplified structure of the evaluation process with ROUGE.

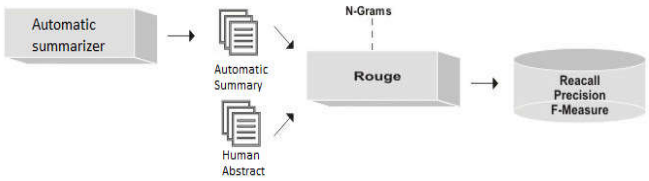

Figure 1. Simplified summary assessment using the ROUGE tool (DELGADO, 2010)

In order to assess the generated summaries, ROUGE uses the statistical metrics Recall, Precision and F-Measure. Recall indicates how the manual summary remains in the automatic summary, Precision indicates how much of the automatic summary coincides with the manual summary and F-Measure indicates the harmonic average between recall and precision. The result of the calculation lies at an interval between 0 and 1 and the closer to 1, the more the automatic summary is similar to the compared human summary (Guelpeli 2012; Luchi and Ribeiro 2011).

PragmaSUM

PragmaSUM is an automatic text summarizer that operates regardless of the text's language or knowledge domain. The summarizer was developed in Java programming language, using the development environment Eclipse for Windows operating system. Because it was developed in Java, PragmaSUM can be executed in any operational system installed with Java virtual machine (JVM). PragmaSUM uses the technique presented by Guelpeli (2012) to evaluate sentences from the source text in the text cluster model Cassiopeia, which presents the method to reduce high dimensionality and sparse data, Luhn algorithm (LUHN 1958), based on the Zipf curve. According to Guelpeli (2012), Zipf's Law (Figure 2) is a specific statistic distribution found in rare stochastic phenomena. The frequency distribution of word occurrence in a text is illustrated as the Y axis representing frequency and the X-axis containing the value of the relative position of this word according to how often it appears in the source text.

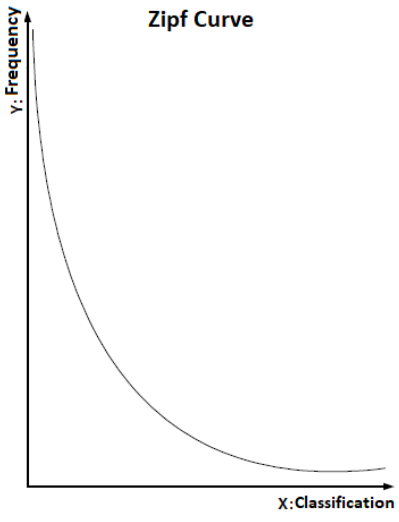

Figure 2. Zipf curve (Guelpeli, 2012)

Luhn proposed, as illustrated by Figure 3, that it is possible to define an upper and lower cut-off, called Luhn's algorithm. Thus, Luhn proposed a technique for finding relevant terms, assuming that the most significant terms for discriminating the content of the document are placed in an imaginary peak, positioned in between two cut-off points, according to Figure 3. The first cut-off, known as upper cut-off, aims to remove the stopwords; i.e., the words that most often appear in the text. The second cut-off is designed to decrease the number of very specific words, found only once in the documents and that contribute to the large number of sparse data in the matrix representation. Thus three distinct areas are generated: trivial or basic information is found more frequently in area I; interesting information is found in area II; white noises are found in area III.

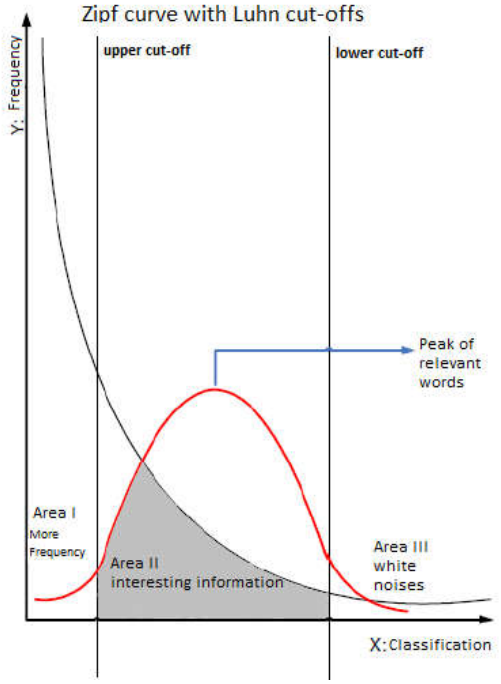

Figure 3. Zipf curve with Luhn cut-offs (Guelpeli, 2012)

Considering the information above, Guelpeli (2012) proposes a technique in his text clustering model, Cassiopeia, for reducing high dimensionality and sparse data. According to Guelpeli (2012), based on the weight of the words, obtained in relative frequency, the average is calculated over the sum of words in the document. In this stage, due to high dimensionality, the model uses truncation; i.e., a maximum of 50 positions for the word vectors, since, according to Wives (2004), a larger value is not necessary, so the cut-off represents the average frequency of the words obtained with the calculations and then organizes word vectors (Figure 4). In order to accomplish this, PragmaSUM calculates the average frequency of the words in the source text and chooses the word that has approximately the same value as the average frequency and selects the 25 words above and 25 words below this average, creating a cut-off with 50 words that will be used in the sentence evaluation of the text.

The compression ratio of the summary can be chosen by the user, and is calculated according to the total number of sentences present in the source text. Equation 1 presents the calculation used by PragmaSUM.

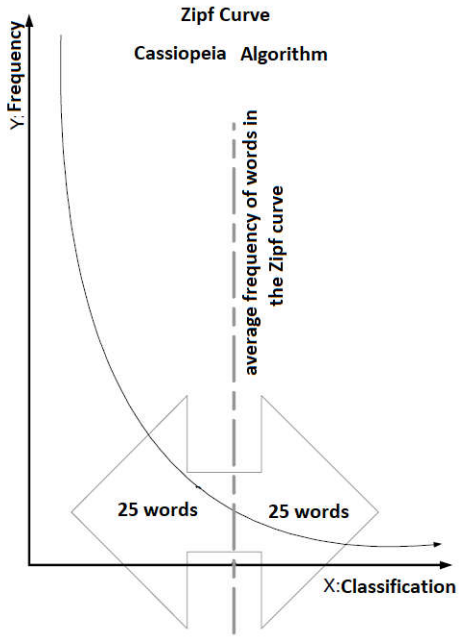

Figure 4. Word selection from the cut-off (Guelpeli, 2012)

$$SS = \frac{ST * TC}{100} \quad (1)$$

Where: SS is the number of sentences in the summary, ST is the number of sentences in the source text, and TC is the percentage of the text that will be present in the summary, given from the difference between 100 and the chosen compression rate. If SS is not a whole number, the closest whole number is selected. In order to personalize the summary, the PragmaSUM user can choose five words, which are classified according to their position: the first word has the highest value and the last word has the lowest, as can be seen in Table 1.

Table 1. Value of Profile Words

Position	Value
1st	6
2nd	5
3rd	4
4th	3
5th	2

The use of these words for sentence assessment of the source text is important for personalizing the obtained summary, thus generating more precise summaries, according to the user's profile. Section 5 presents the results of the precision obtained by using these words. Figure 5 presents the PragmaSUM interface.

Methodology

This section describes the methodology used for the tests conducted with the summarizers, the construction of a test corpus, the chosen domains, its organization and the sources from which they were extracted. The choice of scientific articles was made considering whether they possess summaries and key words that were used in the summarization by PragmaSUM and in the assessment performed by the ROUGE tool.

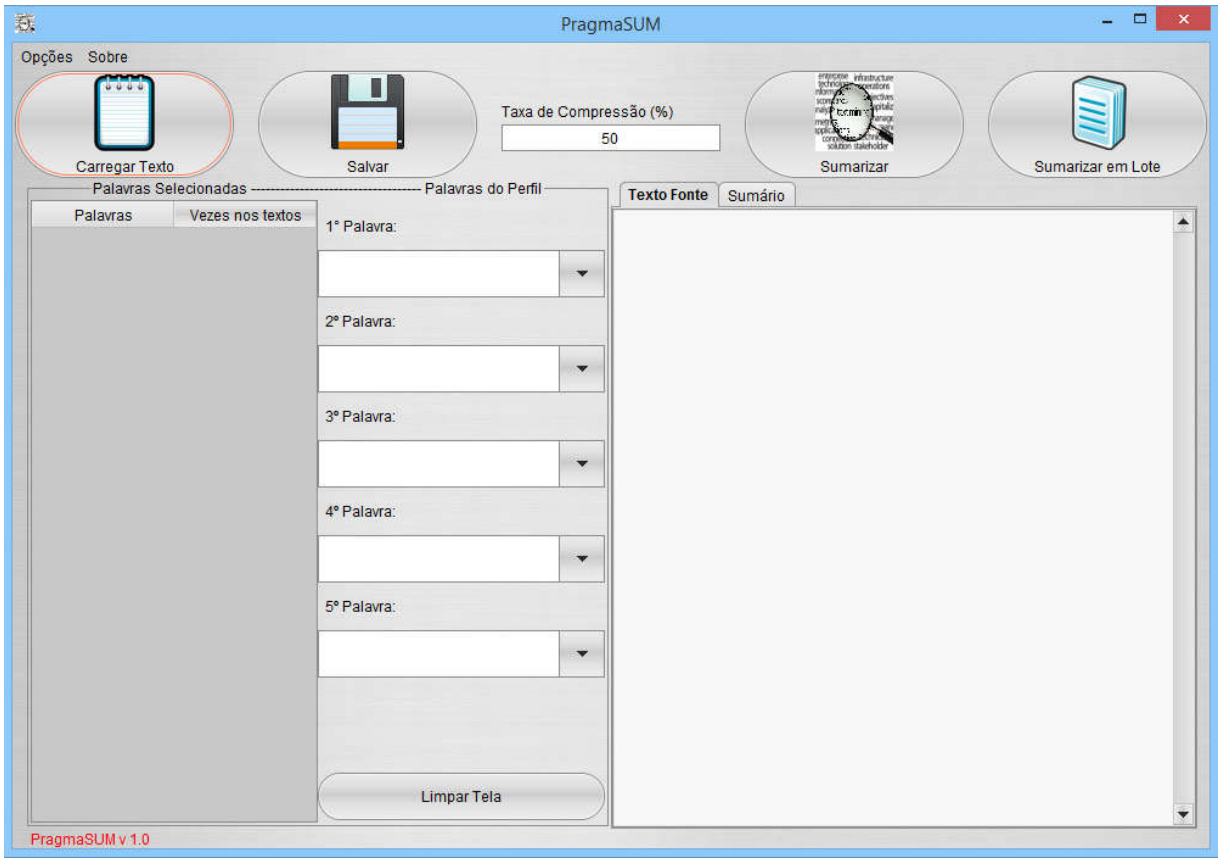

Figure 5. PragmaSUM Interface

Table 2. Corpus statistics created for tests

Files	Characters	Characters and spaces	Words	Words and numbers	Sentences	Average number of words per text
Business	2132914	2634837	424805	435561	25362	8496
Agronomy	684889	873911	137806	146621	13384	2756
Physical Education	1469394	1841347	298644	306003	23392	5972
Engineering	1563482	2008393	313304	322307	19588	6266
Philosophy	1761247	2202861	371484	377061	26260	7429
Physiotherapy	769856	924618	153446	160104	21129	3068
Geography	1482445	1775764	298354	306404	31658	5967
History	2057161	2469530	425457	432658	40889	8509
Medicine	849235	1016013	168232	176688	19571	3364
Psychology	1529982	1830251	305705	311233	32760	6114.1
Total	14300605	17577525	2897237	2974640	253993	57941
Standard deviation	511808.2	630245.2	105794.1	105857.2	7947.024	2116.011
Total average	1430060.5	1757752.5	289723.7	297464	25399.3	5794.1

All scientific articles were taken from the Web of Science database, totaling 500 texts separated into 10 different knowledge domains, each one containing 50 articles and divided into 3 text files, summary, key words, and article content. Table 2 shows the corpus statistics for each knowledge domain. The free version of the software FineCount 2.6 was employed for creating statistics. After the corpus was created, simulations were conducted in order to evaluate the performance of automatic summarizers BLMSumm, GistSumm and PragmaSUM. BLMSumm (Oliveira and Guelpeli, 2011) is independent of language and domain of the source text and uses different sentence and algorithm classification methods for generating summaries. Since there is no study about the best algorithm used by BLMSumm, in this article, all summaries generated by BLMSumm were conducted from the combination of the sentence classification method TF-ISF with the randomly picked algorithm Subida de Encosta (Hill Climb).

GistSumm (Pardo 2002, 2005) is a summarizer based on the text's main idea through which it is possible to identify the sentence that best represents the main idea of the text, which Pardo (2002) calls gist sentence. GistSumm uses a superficial approach, i.e. statistical methods, to identify the gist sentence or the sentence that resembles it the most. Two forms of summarization were used with PragmaSUM: the first used 5 key words from each corpus text by seeking personalization of the summary; the second used the same form without using these key words. Summarization of the texts was conducted with four different compression rates: 50%, 70%, 80% and 90%. Altogether, 8 thousand summarizations were performed (500 source texts * 4 compression rates * 4 summarization methods). After the summaries were generated, the ROUGE tool was used for assessing them. In view of the results presented, statistical tests were used in order to verify if there is a significant difference among samples. These statistical results were separated by domain and compression rate and

they will be presented in the following section. It is worth highlighting that the manual summaries and key words used in this work were extracted from scientific articles, thus they may not correspond to the content of the text and influence the results obtained.

RESULTS

Due to the large number of data generated by measures used to calculate the efficiency of the summaries and aiming to compare the precision obtained by PragmaSUM by using key words in the text, only the graphs comparing domains of the Precision metric will be presented in this article. All other results, as well as the created corpus and the summarizers used are available at http://goo.gl/xFH1zg. The Precision metric indicates the rate in which the automatic summary coincides with the manual summary, and way in which the use of key words aims to personalize the summary according to its incidence in the source text becomes an ideal metric for analyzing the performance of the algorithm used by PragmaSUM. Table 3 and Figure 6 present the Precision results of the Precision comparison of all domains with the compression rate of 50%. It can be observed that PragmaSUM has a slight advantage over the version without key words, with the exception of the Agronomy and Engineering domains. GistSumm presents the lowest results.

Table 4 and Figure 7 present the results of Precision when comparing all domains with the compression rate of 70%. The same conclusion made from previous results can be made here. The difference noted between both versions of PragmaSUM is now greater.

Table 3. Precision comparison among all domains with 50% compression

	BLMSumm	GistSumm	PragmaSUM	PragmaSUM_no_keywords
Business	0.8888	0.79366	0.91603	0.9093
Agronomy	0.8402	0.65554	0.8768	0.88035
Physical Education	0.86777	0.7467	0.89809	0.89319
Engineering	0.8719	0.72004	0.90354	0.90865
Philosophy	0.82195	0.74733	0.8522	0.83615
Physiotherapy	0.84055	0.65235	0.87466	0.86192
Geography	0.86642	0.75347	0.90368	0.89426
History	0.845	0.77344	0.8774	0.87338
Medicine	0.80604	0.70122	0.84976	0.84527
Psychology	0.85757	0.76213	0.88959	0.88076

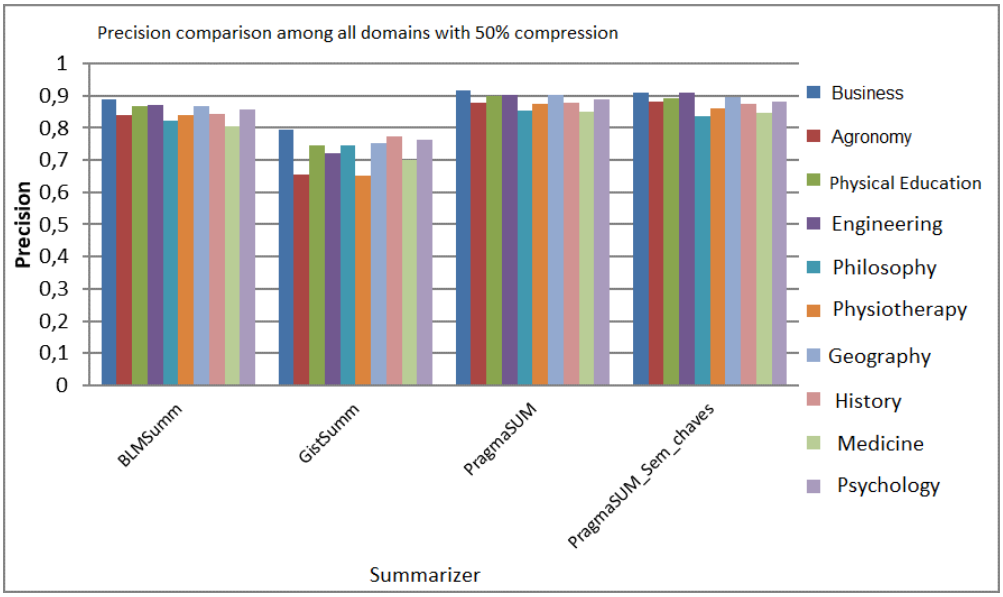

Figure 6. Precision comparison among all domains with 50% compression

Table 4. Precision comparison among all domains with 70% compression

	BLMSumm	GistSumm	PragmaSUM	PragmaSUM_no_keywords
Business	0.82124	0.78366	0.88092	0.86515
Agronomy	0.73472	0.64171	0.80941	0.81219
Physical Education	0.78361	0.7349	0.86464	0.85832
Engineering	0.80894	0.7142	0.87489	0.87875
Philosophy	0.7523	0.72787	0.817	0.79745
Physiotherapy	0.71061	0.65051	0.79785	0.77221
Geography	0.77601	0.74222	0.86345	0.82431
History	0.75894	0.75827	0.83957	0.83483
Medicine	0.70487	0.66985	0.7861	0.77244
Psychology	0.76809	0.73778	0.85436	0.82675

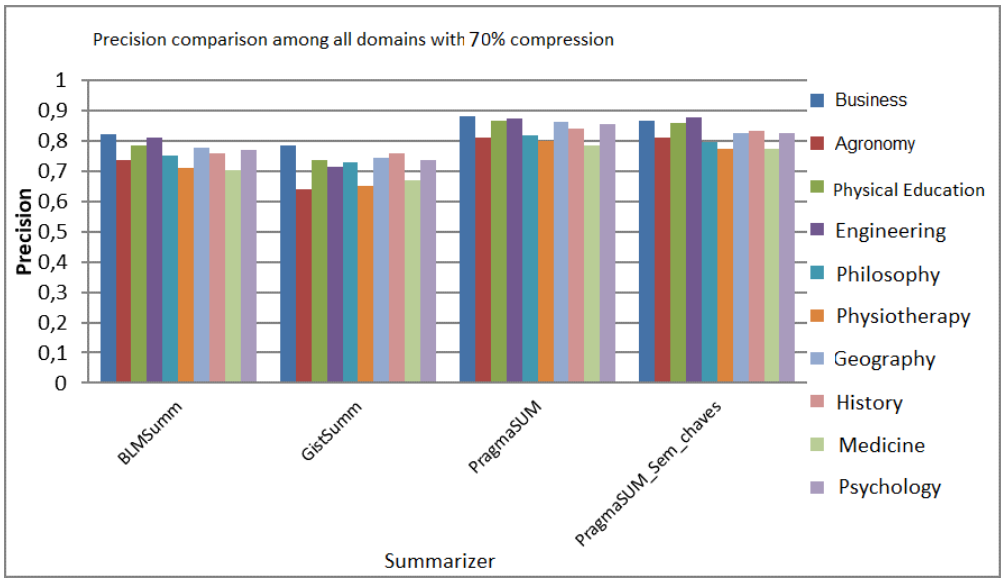

Figure 7. Precision comparison among all domains with 70% compression

Table 5 and Figure 8 present the results of the Precision metric comparing all domains with the compression rate of 80%. PragmaSUM is only surpassed by it keyword-less version in the Engineering domain and identical results were achieved in the Medicine domain. GistSumm results are more similar to BLMSumm results, surpassing them in 3 domains. Table 6 and Figure 9 present the Precision results regarding comparison among all domains with a 90% compression rate. It was observed that PragmaSUM gains considerable advantage over its keyword-less version, except in the Engineering domain. GistSumm results are very similar to BLMSumm results, surpassing them in 7 domains.

In view of the results obtained from PragmaSUM, the use of key words in the text summarization improved the performance obtained in the Precision metric. This improvement was observed mainly with the largest compression rates applied. It is worth highlighting that there was no previous text analysis whatsoever regarding the influence of key words in its content, which could influence the results achieved by the analyzed metrics. A factor that deserves attention is the comparison between both methods used in summarization by PragmaSUM. There was considerable improvement after using key words in most domains, particularly when the compression rate used was

Table 5. Precision comparison among all domains with 80% compression

	BLMSumm	GistSumm	PragmaSUM	PragmaSUM_no_keywords
Business	0.75515	0.7479	0.84415	0.8249
Agronomy	0.63662	0.59751	0.74753	0.73551
Physical Education	0.71027	0.68868	0.82033	0.81386
Engineering	0.73183	0.69611	0.84401	0.84718
Philosophy	0.68261	0.67582	0.78883	0.75485
Physiotherapy	0.62178	0.61129	0.72907	0.68788
Geography	0.69616	0.69757	0.81836	0.77992
History	0.69714	0.72001	0.79955	0.79716
Medicine	0.61346	0.60983	0.71372	0.71372
Psychology	0.70087	0.70345	0.81329	0.76467

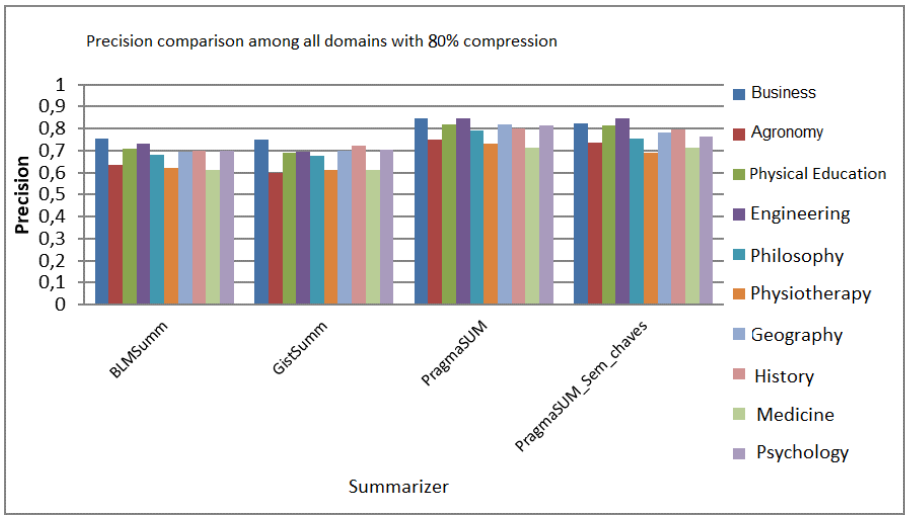

Figure 8. Precision comparison among all domains with 80% compression

Table 6. Precision comparison among all domains with 90% compression

	BLMSumm	GistSumm	PragmaSUM	PragmaSUM_no_keywords
Business	0.75515	0.7479	0.84415	0.8249
Agronomy	0.63662	0.59751	0.74753	0.73551
Physical Education	0.71027	0.68868	0.82033	0.81386
Engineering	0.73183	0.69611	0.84401	0.84718
Philosophy	0.68261	0.67582	0.78883	0.75485
Physiotherapy	0.62178	0.61129	0.72907	0.68788
Geography	0.69616	0.69757	0.81836	0.77992
History	0.69714	0.72001	0.79955	0.79716
Medicine	0.61346	0.60983	0.71372	0.71372
Psychology	0.70087	0.70345	0.81329	0.76467

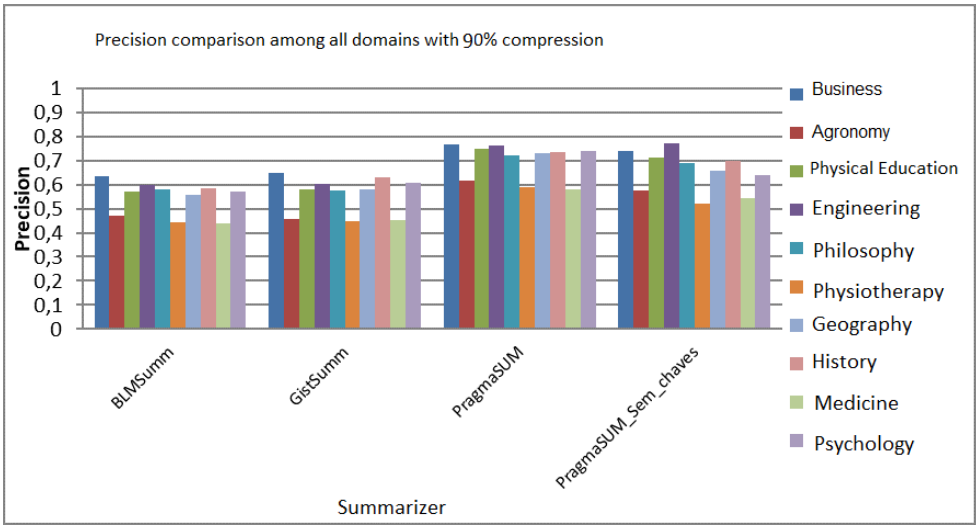

Figure 9. Precision comparison among all domains with 90% compression

80% or 90%. When summary size decreases, the selection of a few sentences becomes necessary, thus increasing the chance that sentences containing key words replace other sentences in the summary.

The results of the Kendall rank correlation coefficient used to measure ordinal associations in statistics demonstrated that all the tables generated by the StatPlus software were satisfactory, thus it is possible to state that there was a significant difference among experimented samples; i.e., all results previously presented were confirmed. It can be stated that PragmaSUM, by employing key words, achieved success in 33 out of 40 analyzed samples; i.e., an 82.5% success rate. PragmaSUM was surpassed by its version without key words in all the compression rates of the Engineering domain, and when compression rates were 50% and 70% in the Agronomy domain, in addition to presenting identical values when the compression rate was 80% in the Medicine domain.

Conclusion

This article's main goal was to present the automatic text summarizer, PragmaSUM. For this reason, a test corpus was created composed of scientific articles in Portuguese and ten different domains. The tools BLMSumm, GistSumm and two forms of PragmaSUM summarizations were evaluated, with and without the use of key words. Tests were conducted by generating summaries with four different levels of compression, 50%, 70%, 80% and 90%. Results were evaluated through the ROUGE tool using Recall, Precision and F-Measure metrics. A comparison between domain results showed that, according to the Precision metric, the Business domain achieved the best results, with all summarizers, with the exception of the keyword-less PragmaSUM version, which presented the Engineering domain with higher results. PragmaSUM was only surpassed in the engineering domain and, when the compression rate was 50% and 70%, in the agronomy domain, where its keyword-less version possessed slight advantage. The presented results were confirmed with the use of Friedman's ANOVA statistical tests and Kendall's rank correlation coefficient. It was possible to note, when observing the four applied compression rates, that automatic summarizers achieved more homogenous results, in both domains, when the compression rate was 50% and much variation when the rate was 90%. It was also observed that, as demonstrated in the results, automatic summarizers presented a trend: the larger the compression applied, the poorer the results, with great variation, with the exception of PragmaSUM, which did not suffer great loss with increase in compression. As previously mentioned, there was no type of analysis regarding the influence of key words over the content of the text. Since the Business domain obtained the best results from Precision, this may signify that, in this domain, key words have greater importance in the text and, in the Medicine and Physiotherapy domain, they have less importance, since they obtained the lowest results. Another factor that can be considered in the performance of each domain is its size, as can be observed in Table 2. Generally speaking, the largest domains obtained the best results, and the smallest obtained the worst. This may be due to the fact that, with higher compression rates, there is greater loss in results, since smaller texts consequently generate smaller summaries. As mentioned earlier, the fact that PragmaSUM is successful with more elevated compression rates with the use of key words is

precisely because its summary is composed of a larger number of important sentences in the text; i.e., sentences that contain the key words chosen by the author. It can thus be concluded that the use of key words in automatic text summarization can personalize the summary according to the user's needs, by extracting sentences that really correspond to the domain of interest.

Future research

Future research can include the creation of a corpus in different languages and in more domains for a wider evaluation scope. Another suggestion is to improve the personalization method of the summary through key words by employing machine learning tools, thus eliminating human interaction in the process and allowing PragmaSUM to learn the linguistic profile of the user. It would be interesting to conduct a study about the relevance of using key words in the content of scientific articles, thus rendering it possible to analyze which criterion researchers adopt to choose key words in their writing, whether they really are relevant to the text or if they are chosen to guide research studies conducted by search algorithms on the internet for their articles. Furthermore, it would be worthwhile studying the amount of these key words in the indexation of scientific articles.

REFERENCES

Delgado, C.H, C.E Viana, and M.C.V Guelpeli. 2010. Comparando sumários de referência humanos com extratos ideais no processo de avaliação de sumários extrativos. IADIS Ibero-Americana.

Gong, Y and X Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, page 19–25.

Guelpeli, M.V.C. 2012. Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização. Ph.D. thesis, Universidade Federal Fluminense.

He, R, B Qin, and T Liu. 2012. A novel approach to update summarization using evolutionary manifold-ranking and spectral clustering. Expert Systems with Applications.

Lin, C.Y. and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. In the Proceedings of the Language Technology Conference.

Lin, CY and E Hovy. 2001. The automated acquisition of topic signatures for text summarization. international conference on computational linguistic.

Luchi, D. and E. Ribeiro. 2011. Sumarização automática de textos via ranqueamento de sentenças. Universidade Federal do Espírito Santo – UFES.

Luhn, H.P. 1958. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2:157–165.

Nenkova, A and K McKeown. 2012. Miningtext data, chapter A Survey of Text Summarization Techniques. US: Springer.

Nóbrega, F.A.A. and T.A.S. Pardo. 2016. Improving content selection for update summarization with subtopic-enriched sentence ranking functions. International Journal of Computational Linguistics and Applications (IJCLA)., 7(2):111–128.

Oliveira, M. A. and M. V. C. Guelpeli. 2011. Blmsumm – métodos de busca local e metaheurísticas na sumarização de textos. ENIA - VIII Encontro Nacional de Inteligência Artificial, pages 287 – 298.

Oliveira, R.R. 2014. A criação de um corpus de textos em italiano e sua utilização na avaliação de sumarizadores automáticos. Monografia(Graduação em Sistemas de Informação) – Universidade Federal dos Vales do Jequitinhonha e Mucuri,.

Pardo, T.A.S. 2002. Gistsumm: Um sumarizador automático baseado na ideia principal de textos. Technical report, Universidade Federal de São Carlos.

Pardo, T.A.S. 2005. Gistsumm – gist summarizer: Extensões e novas funcionalidades. Technical report, Universidade Federal de São Carlos.

Tabassum, Shazia and Eugenio Oliveira. 2015. A review of recent progress in multi document summarization. Doctoral Symposium in Informatics Engineering.

Vanderwende, L., H. Suzuki, C. Brockett, and A Nenkova. 2007. Beyond sumbasic: Taskfocused summarization with sentence simplification and lexical expansion. Information Processing & Management.

Wives, L. K. 2004. Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos. Ph.D. thesis, Universidade Federal do Rio Grande do Sul.

****