Comparison of State-of-the-Art Methods and Commercial Tools for Multi-Document Text Summarization

Yulia Ledeneva[1], René García Hernández[1], Grigori Sidorov[2],
Griselda Mathias Mendoza[1], Selene Vargas Flores[1], and Abraham García Aguilar[1]

[1] Universidad Autónoma del Estado de México,
Unidad Académica Profesional Tianguistenco,
Paraje el Tejocote San Pedro Tlantizapan, 52600, Estado de México, Mexico
[2] Natural Language and Text Processing Laboratory,
Center for Computing Research (CIC),
National Polytechnic Institute (IPN),
Av. Juan Dios Batiz, s/n, Zacatenco, 07738,
Mexico City, Mexico
yledeneva@yahoo.com, renearnulfo@hotmail.com, www.g-sidorov.org

Abstract. The final goal of Automatic Text Summarization (ATS) is to obtain tools that produce the most human-similar summary. Almost all the papers on ATS research area present a review of the state-of-the-art of one side of the issue, since only is reviewed the-state-of-the-art of the tools reported in papers. However, we found a great number of developed commercial tools which are not reported in papers (which is understandable by competitive reasons), but also have not been evaluated. The question is what commercial tools are good in comparison to paper-published tools. This paper gives a survey for 18 commercial tools and state-of-the-art methods for multi-document summarization task testing on a standard collection of documents which contains 59 collections of documents.

Keywords: Automatic multi-document summarization, Copernic summarizer, Microsoft office word summarizer, Svhoong summarizer, pertinence summarizer, Tool4noobs summarizer.

1 Introduction

According to recent researches the volume of information on the public Web are estimated at 167 terabytes, while the deep web to be 400 to 450 times larger, thus between 66,800 and 91,850 terabytes [1]. Such amount of information cannot be revised by a normal human, only with the help of computer by means of intelligent algorithms and tools. Such types of algorithms and methods have a big necessity and urgency to be developed, for example, automatic generation of text summaries for multi-document collection.

Humans tend to create summaries by generating and fusion of ideas, changing words and rephrasing long sentences. This manner of composing summaries is called abstractive summarization, contrary to the extracted summarization, where different text units (words, phrases, sentences, etc.) are extracted from the original collection of documents. The generation of extractive summaries does not require the understanding of the text.

Test summary is a short text transmitting the main ideas of the collections of documents without redundancy describing these ideas. In this paper, we take into account some important parameters to compare tools, which some of the tools permit to be changed depending of the resulted summary. These parameters are the size and format of the summary. The size of the summary depends on the user needs and can depend from the size of the original document. Therefore, the size of the summary must be the most flexible parameter. Another parameter is the format in which the summary is presented to the user, the most important key phrases or sentences can be highlighted within the summary or original document, without deleting the context in which such phrases occur. It is also desirable that the tools can work independently of domain and language of a given document, indeed it is not necessary that the original document is grammatically well written. We consider that to achieve a good summary, the tool should work mostly with text content and to a less degree with the document format. Also, a good tool to generate summaries should have a friendly interface. We consider that a good summary also must have coherence.

We consider two types of methods for generating multi-document summaries. There are commercial tools and the state-of-art methods. The differences are the prices for the commercial tools and lack of their description. For commercial tools, we consider two groups: installed and online tools (see the section 3).

Currently, there are commercial tools that automatically generate summaries compressing main ideas of a collection of documents. The first objective of the paper is to know which of the commercial tools produces summaries most similar to a human. The second objective is to compare the commercial tools to the state-of-the-art methods.

The paper is organized as follows. Section 2 summarizes the state-of-the-art of multi-document summarization methods. In Section 3, commercial tools are introduced. Section 4 presents the experimental settings and results. Sections 5 discuss obtained experimental results and give some conclusions of the paper.

2 State-of-the-Art Methods for Multi-Document Summarization

The following state-of-the-art methods are obtained promising results but not yet commercialized.

Maximal Frequent Sequences (MFSs): This work presents a method to generate extractive summaries from a multi-document based on statistics, which is independent of the domain and language. Ledeneva *et al.* [2, 3] experimentally shows that the words which are parts of bigrams (2-word sequences) which are repeated more than once in the text are good terms to describe the content of that text, so also called the maximal frequent sequences (sequences of words that are repeated a number of times and also are not contained in other frequent sequences). This work also shows that the frequency of the term as ranking of terms gives good results (while only count the occurrences of a term in repeated bigrams). Ledeneva *et al.* applies a method which has 4 stages for generating the summary. These steps are term selection, term weighting, sentence weighting and sentence selection. In term selection step, SFMs, repetitive bigrams (must appear at least twice in the text), and unigrams (simply words) are extracted. In term weighting step, the frequency of the term is used, which is the number of times the term occurs in the text. In sentence weighting, only the weight of all the terms contained in that sentence is calculated. Finally, sentence selection that composes the summary is performed by two criteria. First, the *k* sentences with bigger weight are selected. Second, *k* sentences with bigger weight are selected and completed with the first sentences (similar to baseline heuristic) that appear in the document (combined version). The best result is obtained with combined version when k=1, reaching 47% of similarity with the summaries made by a human.

MFS 1st method: In [4] are shown that MFSs are good descriptors if the lengths of the descriptors are considered, and words that derived from MFSs are good descriptors if the frequency is considered. For term selection option *W*, the term weighting option *f* showed best performance in all experiments.

MFS 2nd method: MFSs with higher threshold generate better summaries than MFSs with lower threshold [5]. It can be explained on reason that there exist in the language multiword expressions that can express the same content in the more compact way which can be detected more precisely using higher.

Then, we observed that, in contrast to MFSs, FSs is important if are extracted with lower threshold. It can be explained because there exist in the language a lot of single word or at least an abbreviation to express an important meaning.

Such single words or abbreviations should be considered as bearing the more important meaning with lower threshold because we need to extract more single words or abbreviations for knowing if they can be used for composing a summary.

The third hypotheses we explore in this work, is that MFSs represent in a better way the summarized content of collection of documents than FSs because their (MFSs) probability to bear important meaning is higher. It can happen because there are too many non-maximal FSs in comparison to MFSs.

MFSs using clustering sentence algorithms: In the previous method, sentences which have bigger weight are selected for composing the summary. However, if the sentences that are chosen in that order may include very similar sentences and do not provide new information in the summary. The work [6] uses a clustering algorithm based on MFSs to make some groups of sentences, from which the most representative sentence from each group are selected to compose the summary.

TextRank graph based algorithm: Mihalcea [7, 8] constructs the graph to represent the text, so the nodes are words or other text sequences interconnected by vertices with meaningful relationships and the vertices are added to the graph for each sentence in the text. A relation of similarity is defined to establish the connections between sentences, where the relationship between two sentences can be seen as a process of "recommendation".

For the sentences extraction task, the goal is to qualify whole sentences and order them from most to least importance. A sentence that points to a certain concept in the text gives to the reader a "recommendation" to refer to other sentences in the text that point to the same concepts, and then a link can be established between any two sentences that share a common content. Since this method can determine the importance of each of the sentences, it was used to generate multi-documents summaries.

Baseline configuration: The baseline configuration consists of taken the first *n* sentences of the document to complete the summary to required size. This configuration supposes that the most important information of a collection of documents is in the first sections of the document. This simple heuristic has been shown to generate very good summaries in the field of news documents [9].

Baseline random configuration: This heuristic far from seeking to obtain best summaries attempts to determine the quality of the summaries when only a set of sentences are taken randomly as a summary. The idea is to determine if obtained results are significant comparing to other “intelligent” methods [3].

Best DUC systems: The best top 5 systems from 17 systems in DUC 2002 for multi-document summarization task are described in [9], see Figure 6 for more details.

3 Commercial Tools for Multi-Document Summarization

Currently, there are several commercial tools that help us in generating automatic summaries, among the most popular are the following tools. In this paper, the tools considered to analyze and compare are: Svhoong Summarizer, Pertinence Summarizer, Tool4noobs Summarizer, Copernic Summarizer, Microsoft Office Word Summarizer 2003 and Microsoft Office Word Summarizer 2007, and Microsoft Office Word Summarizer 7.

Svhoong Summarizer: This summarizer is available online. The text should be copied to the web page [11]. The final summary are the text underlined in the same page. The usage of this summarizer it is very tedious because the final summary should be saved sentence by sentence. It offers some options to generate summaries of different size, with the following percentages 10, 20, 30, 40, 50, 60, 70, 80, 90 of the original text. In this work, the percentages were calculated and approximated according to available option of the tool. This step was made manually and took long time. It is available for 21 languages.

Pertinence Summarizer: This summarizer is available online [12]. With each document, Pertinence calculated percentages automatically depending on the number of word in the document. For example, 1% (34 words), 5% (171 words), 10% (342 words), etc. There are 3 forms to introduce text: copy to the web page, examine the document (this option was utilized in this work), or introduce the address of the web page. The tool does not have an option for 100 words, thus the percentage were automatically calculated for the given collection. It is available for 12 languages.

Tool4noobs Summarizer: This summarizer is available online [13]. It all integer percentages from 1 to 100 of the original text can be introduced. The original text should be copied to the web page. A least one sentence should be introduced. It permits to introduce 50 lines of text. The percentages were re-calculated, because the difference to 100 should be calculated.

This tool uses three steps to generate summaries: extraction of text; identification of the key-words in the text and its relevance; identification of sentences with key-words and generation of summary.

Copernic Summarizer. This software was developed exclusively for the generation of automatic summaries. It is a flexible and suitable tool. In this paper, we use version 2.1 which was installed on the Microsoft Windows operating system. It offers different options to generate summaries from multiples documents:

- 5%, 10%, 25% and 50% of words of the original collection of documents;
- 100, 250 and 1000 words.

According to [10], Copernic Summarizer uses the following methods:

1. Statistical model (S-Model). This model is used in order to find the vocabulary of the text.
2. Knowledge Intensive Processes (K-Process). Consider the way in which human make summary texts by taking into account the following steps:
3. Language detection. It detects the language (English, German, French or Spanish) of the document for applying specific processes.
4. The limits of sentence recognition.
5. Concept extraction. Copernic Summarizer uses machine learning techniques to extract keywords.
6. Document Segmentation. Copernic Summarizer organizes the information that it can be divided into larger related segments.
7. Sentence Selection. Sentences are selected according to their importance (weight) discarding those that decrease readability and coherence.

Table 1. Parameters of Commercial Tools.

Tool	Language	Price	Characteristic
Copernic Summarizer	Franch, English, Dutch, Spanish	$59 US	Installed
Microsoft Word 2003	Multilingual	$830.78	Installed
Microsoft Word 2007	Multilingual	$1400.00	Installed
OTS	Multilingual	Free software	Installed
Pertinence	Multilingual	Free software	Online
Tool4noobs	Multilingual	Free software	Online
Shvoong	Multilingual	Free software	Online

Microsoft Office Word Summarizer. This tool can be found in versions of Microsoft Office Word 2003 and Microsoft Office Word 2007. This tool can generate summaries of 10 or 20 sentences, 100 or 500 words (or less) or in percentages of 10%, 25%, 50% and 75% of words of the original document. If some of the percentages are not appropriate, the user can change as needed. This tool offers various ways of visualizing summaries. One is highlighting the color of important sentences in the original document.

The summary created by this tool is the result of an analysis of key words; the selection of these is done by assigning a score to each word. The most frequent words in the document will have highest scores which will be considered as important. The sentences containing these words will be included in the summary.

4 Experimental Results

For comparing the above mentioned applications the collection Document Understanding Conference (DUC) 2002 [9] was used, which was created by the National Institute of Standards and Technology (NIST) for the usage by researchers in the area of automatic text summarization. This collection has the data set of 60 document collections which consist of 567 news articles of different length about technology, food, politics, finance, etc. For each document in the collection was created two summaries by two human experts with a minimum length of 100 words.

ROUGE 1.5.5 evaluation toolkit, proposed by Lin [14, 15], is the tool used for the automatic comparison of summaries. In particular, we use n-gram statistics (where n = 1), which has the ability to measure similarity and determine the quality of an automatic summary compared to the both summaries created by a human. We use this tool, to compare quality of the generated summaries by commercial tools and the state-of-the-state methods.

4.1 Configuration of Experiments

Commercial tools were evaluated in the operating system Windows XP Professional Service Pack 2 (SP2). Each file was manually selected and applied to generated summary of 100 words. In the case of Microsoft Office Word 2003, 2007 and version 7 is not possible to use the option of 100 word summaries because sometimes produces summaries less than 100 word, which decrease the quality of summaries. For such problem was necessary to calculate the adequate percentage to produce a summary with minimum 100 words, calculated as follows: $(Number\_of\_diserable\_wrods / Number\_of\_total\_words) * 100$.

4.2 Quality of Online Commercial Tools

The evaluation results of commercial tools are realized using ROUGE. The best obtained result was for online commercial tools Shvoong Summarizer.

4.3 Quality of Installed Commercial Tools (case Microsoft Office Word)

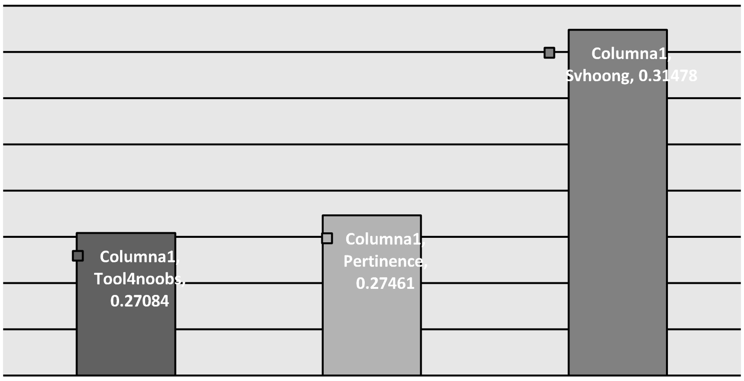

Figure 1. The evaluation results of online commercial tools testing using the operating system Windows XP, Vista,and 7.

Microsoft Office Word 2007 using the operating system Windows XP was the best obtained result than using the operating system Windows Vista or Windows 7 (see Figure 2). Microsoft Office Word 2007 using the operating system Windows 7 obtained less result than using Microsoft Word 2003 using different operating systems (see Figure 2).

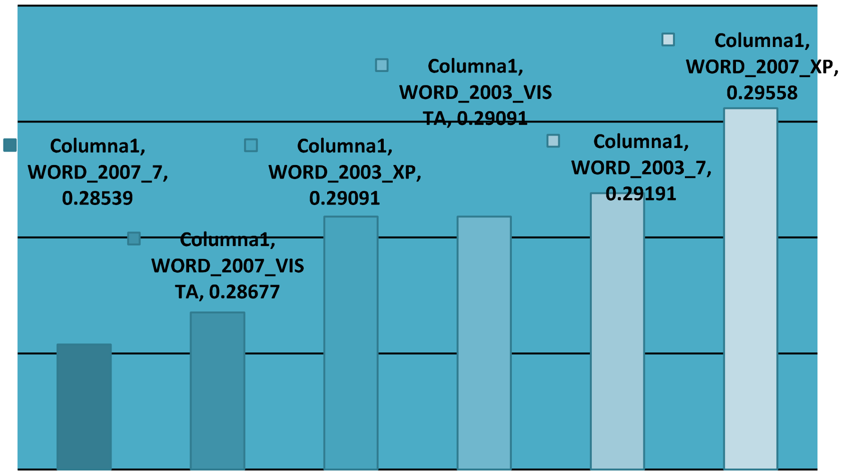

Figure 2. The evaluation results of different versions of installed commercial tool Microsoft Office Word tesitng in the operating system Windows XP, Vista, and 7.

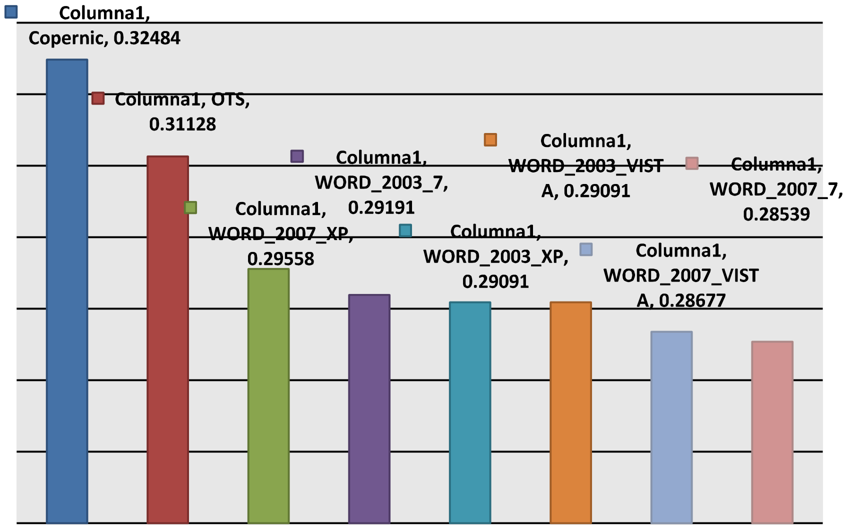

Figure 3. The evaluation results of all installed commercial tools testing using the operating system Windows XP, Vista, and 7.

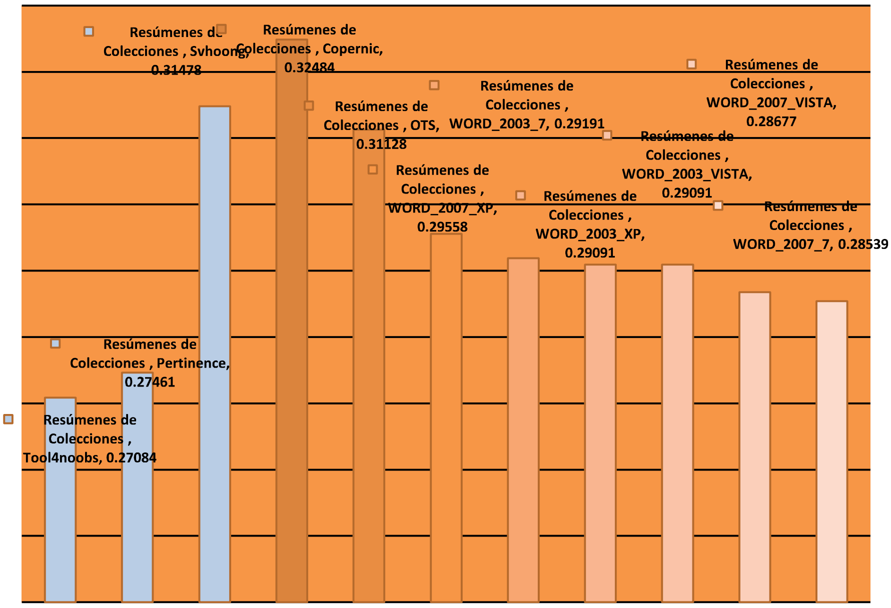

Figure 4. The evaluation results of all installed and online commercial tools testing using the operating system Windows XP, Vista, and 7.

4.4 Quality of all commercial tools

Copernic Summarizer obtained the best results of all installed commercial tools (see Figure 3).

Microsoft Office Word 2007 y Microsoft Office Word 2003 obtained less quality then online tools Copernic, OTS y a Shvoong. The installed commercial tools are marked with orange color (see Figure 4).

4.5 Quality of Commercial Tools and State-of-the-Art Methods

In order to see the quality of previous results compared to those obtained with the state-of-the-art methods, in Figure 5 are shown the best results obtained by commercial tools and the results reported by the state-of-the-art methods.

Figure 5 shows clearly that the results of Copernic Summarizer is the highest score, just below the proposed method MFSs (1Best+First) and below of Sentence Clustering with MFSs, which confirms this software is one of the best of its kind.

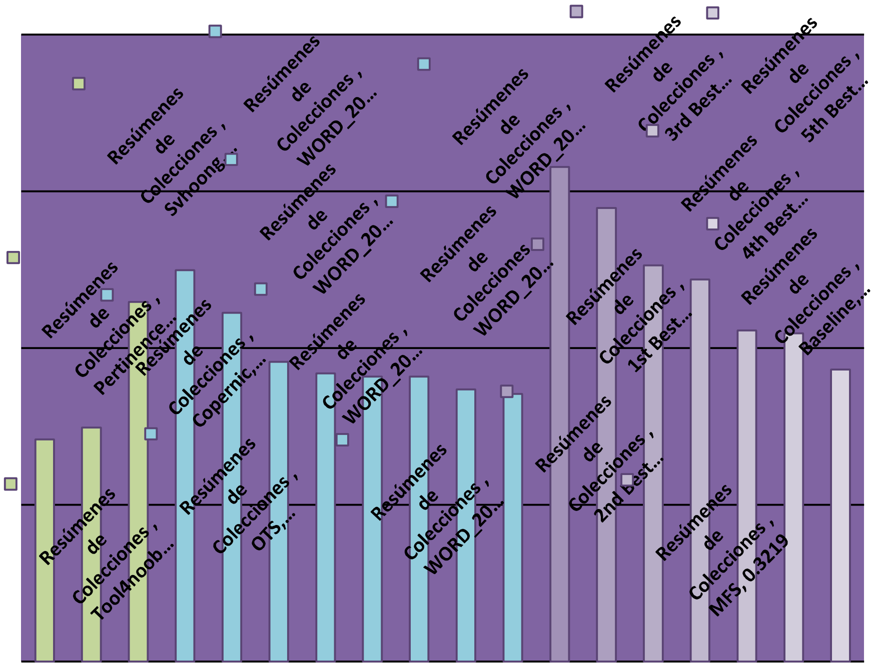

Figure 5. Results for the collection of documents obtained for commercial tools and the state-of-the-art methods.

Table 2. The results of Commercial Tools are ordered by the quality of Text Summaries.

Commercial Tools	F-measure	Characteristic
Copernic	0.32484	Installed
Svhoong	0.31478	Online
OTS	0.31128	Installed
WORD_2007_XP	0.29558	Installed
WORD_2003_7	0.29191	Installed
WORD_2003_XP	0.29091	Installed
WORD_2003_VISTA	0.29091	Installed
WORD_2007_VISTA	0.28677	Installed
WORD_2007_7	0.28539	Installed
Pertinence	0.27461	Online
Tool4noobs	0.27084	Online

According to Figure 1 Copernic Summarizer outperforms the two versions of Microsoft OfficeWord, although Microsoft Office Word 2003 was slightly better than its 2007 version. However, during the experimentation an inconsistency in Microsoft Office Word was observed because the generated summaries change depending on the operating system. In order to verify this fact the same setup package of Microsoft Office Word 2003 was installed on Windows XP Professional SP2, which also was done with Microsoft Office Word 2007. The resulting abstracts were assessed with the same version of ROUGE and obtained the results shown in Figure 2.

In Figure 2, we can observe the slight difference between the tools of auto summary of Microsoft Office Word. In contrast, Copernic Summarizer tool show the same results in both operating systems. We can conclude that Copernic Summarizer is independent of the operating system).

Table 3. The results of the State-of-the-Art Methods (SAMs) and Commercial Tools are ordered by the quality of Text Summaries.

Commercial Tools and State-of-the-Art Methods	**F-measure**	***Characteristic***
1st Best Method	0.3578	*SAMs*
2nd Best Method	0.3447	*SAMs*
3rd Best Method	0.3264	*SAMs*
Copernic	0.32484	*Installed*
MFS	0.3219	*SAMs*
Svhoong	0.31478	*Online*
OTS	0.31128	*Installed*
4th Best Method	0.3056	*SAMs*
5th Best Method	0.3047	*SAMs*
WORD_2007_XP	0.29558	*Installed*
Baseline	0.2932	*SAMs*
WORD_2003_7	0.29191	*Installed*
WORD_2003_XP	0.29091	*Installed*
WORD_2003_VISTA	0.29091	*Installed*
WORD_2007_VISTA	0.28677	*Installed*
WORD_2007_7	0.28539	*Installed*
Pertinence	0.27461	*Online*
Tool4noobs	0.27084	*Online*

Also, it is important to mention that Microsoft Office Word 2007 with Windows XP Professional SP2 obtained the worst result, among the versions of Microsoft Office Word.

Nevertheless, the best result, among the versions of Microsoft Office Word, was obtained with Microsoft Word 2003 with Windows Vista Home Premium SP1. This shows the dependence of these tools with respect to the operating system they are using.

5 Discussion and Conclusions

Only four commercial tools are better than baseline configuration: Svhoong, Copernic, OTS, WORD_2007_XP. Other six evaluated commercial tools are worst than the baseline configuration. All the state-of-the-art methods overcome the baseline configuration and quality of the commercial tools (see Figure 6).

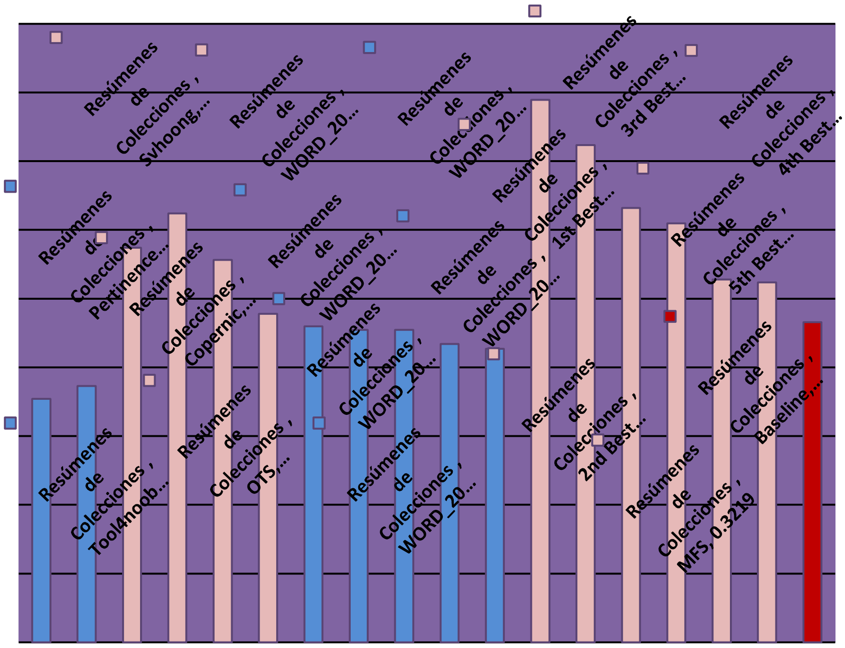

Figure 6. Results for the collection of documents obtained for commercial tools and the state-of-the-art methods.

In this paper, the evaluation of the automatic summaries generated by commercial tools (Copernic Summarizer, Microsoft Office Word Summarizer 2003 and Microsoft Office Word Summarizer 2007) was realized. The summaries were evaluated using the ROUGE system. The following conclusions can be given based on obtained results:

– 18 different automatic multi-document summarization state-of-the art methods and commercial tools were compared

- In most cases, the state-of-the-art methods are better that commercial tools
- Copernic Summarizer gets the best results of commercial tools.
- Copernic Summarizer is the best commercial tool
- Shvoong is the best online tool
- Microsoft Office Word is inconsistent because it generates different summaries depending on the operating system.
- The results obtained with Microsoft Office Word 2003 and Microsoft Office Word 2007 with Windows Vista were better than with Windows XP.
- Microsoft Office Word 2003 gets a better result than Microsoft Office Word 2007 with Windows Vista operating system.

We consider that computer methods can perform better and quicker than human the task of multi-document text summarization, in particular, reducing the contents of a big collection of documents to a single short text, so that the user can judge about the contents of the whole collection upon reading only this short text. However, it is still a challenge for computer methods to improve the quality for multi-document text summarization and even bigger challenge coherence. Good news is that the state-of the-art methods perform better than commercial tools.

Acknowledgments. Work done under partial support of Mexican Government (CONACyT, SNI, SEP-PROMEP, UAEM, SIP-IPN 20111146, 20113267) and Mexico City Government (project ICYT-DF PICCO10-120). The authors thank Autonomous University of the State of Mexico for their assistance.

References

1. Lyman, Peter and Hal R. Varian. How Much Information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 (2003)
2. Yulia Ledeneva, Alexander Gelbukh, René A. García-Hernández. Terms Derived from Frequent Sequences for Extractive Text Summarization, 9th Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008), Lecture Notes in Computer Science, Springer-Verlag, Vol. 4919. pp. 593-604 (2008)
3. Yulia Ledeneva. Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization, National Polytechnic Institute, PhD. Thesis, Mexico (2009)
4. Yulia Ledeneva, René García Hernández, Alexander Gelbukh. Multi-document Summarization using Maximal Frequent Sequences. Research in Computer Science, pp.15-24, vol. 47, ISSN 1870-4069 (2010)
5. Yulia Ledeneva, René García Hernández, Anabel Vazquez Ferreyra, Nayely Osorio de Jesus. Experimenting with Maximal Frequent Sequences for Multi-Document Summarization. Research in Computing Science, pp.233-244, vol. 45, ISSN 1870-4069 (2010)
6. René Arnulfo García-Hernández, Romyna Montiel, Yulia Ledeneva, Eréndira Rendón, Alexander Gelbukh, Rafael Cruz. Text Summarization by Sentence Extraction Using Unsupervised Learning, 7th Mexican International Conference on Artificial Intelligence

(MICAI08), Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 5317 pp. 133-143 (2008)

7. Rada Mihalcea; Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization; Department of Computer Science; University of North Texas; Texas; EUA (2004)
8. Rada Mihalcea, Paul Tarau. A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP'2005 (2005)
9. DUC. Document Understanding Conference, www-nlpir.nist.gov/projects/duc.
10. Copernic Summarizer, Technologies White Paper (2003) http://www.copernic.com/data/pdf/summarization-whitepaper-eng.pdf
11. Online Tool Noobs Summarizer. http://www.tools4noobs.com/summarize/
12. Pertinence Summarizer. http://www.pertinence.net/index_en.html
13. Shvoong Summarizer. http://www.shvoong.com/summarizer/
14. Lin, C., y E. Hovy: Automatic evaluation of summaries using N-gram co-occurrence statistics. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Vol. 1, pp. 71-78 (2003)
15. Lin, C.: ROUGE: A package for automatic evaluation of summaries. Proceedings of the Association for Computational Linguistics 2004 Workshop, pp. 74-81. Spain (2004)
16. Garcia, R., F. Martinez, A. Carrasco, Finding maximal sequential patterns in text document collections and single documents, Informatica, International Journal of Computing and Informatics, ISSN: 1854-3871, No. 34, pp. 93-101 (2010)
17. Sidorov, G. Lemmatization in automatized system for compilation of personal style dictionaries of literary writers. In: "Word of Dostoyevsky", Russian Academy of Sciences, pp. 266-300 (1996)
18. Gelbukh, A. and Sidorov, G. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. Lecture Notes in Computer Science, N 2588, Springer-Verlag, pp. 215–220 (2003)
19. Sidorov, G., Barrón-Cedeño, A., and Rosso, P. English-Spanish Large Statistical Dictionary of Inflectional Forms. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA), pp. 277-281 (2010)
20. Esaú Villatoro-Tello, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, and David Pinto-Avendaño. Multi-Document Summarization Based on Locally Relevant Sentences. Mexican International Conference on Artificial Intelligence MICAI 2009. Guanajuato, Mexico, November 09-13, pp. 87-91, IEEE Computer Society (2009)