Using text mining to support text summarization

Eliseo Reategui and Daniel Epstein

[1]Programa de Pós-Graduação em Informática na Educação
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brasil

eliseoreategui@gmail.com, daepstein@gmail.com

Abstract

Reading and writing texts are foundational abilities required for daily activities, human interaction and almost all school disciplines. However, it is not uncommon to find students with reading and writing difficulties. Those difficulties interfere on their learning process and their ability to comprehend more complex ideas. Therefore, failing to read and write leads not only to academic failure but it may also hinder occupational success. Several studies have presented graphic organizers as a way to assist students in reading comprehension tasks and to help them structuring their own text production. Here we present a text mining tool capable of extracting the main concepts and relationships from texts present them in a graphical way. These visual representations of a text may be used by students as graphic organizers, helping them to reflect about the text's main ideas before the actual writing task. The results of an experiment are presented, in which a total of 20 students were asked to read and summarize a short text with the assistance of the text mining tool. The results show that Sobek helped students reflect about the main ideas of the text and supported the actual writing of the summaries.

Keywords: text mining, summarization, writing, reading, literacy, graphic representation.

1. Introduction

Knowing how to read and write a small sentence about one's own life was once considered enough to say that a person is literate. Nowadays, literacy is a much more complex concept than simply to be able to write or read sparse sentences. [Warschauer 2006] defines literacy as the ability to participate in the meanings of text, to use texts functionally and to critically analyze and transform texts. Individuals are expected to use oral and written language to demonstrate an understanding of the world, communicate, participate in problem solving and decision making [Jenner 2003].

It is clear that literacy is an important part of a student's development and several countries has been trying to improve theirs tools and methods to promote literacy. In Brazil, since 1980's, literacy has become the central debate problem in education and Portuguese learning. A lot of investments have been made and yet the country holds one of the last positions among those evaluated by PISA - Program for International Student Assessment [OECD 2013]. [Meneghetti et al. 2006] shows that students' performance in subjects such as Science and Mathematics tends to be directly connected with their ability to read and comprehend written material. Although to some degree this is a problem that affects most of the world, it is noticeably more evident in developing countries, such as

DOI: 10.5753/cbie.sbie.2015.1217

in Latin America and the Caribbean [SERCE 2008]. Therefore, it becomes important to find new ways to help students overcome such problems.

Different strategies and tools have been proposed along the years to support reading and writing, abilities that are essential for the construction of meaning, the understanding of new facts and concepts [Jacobs 2002]. Students with reading difficulties may have problems organizing the information read and comprehending it. Among the different approaches used to help solving this problem there are tools that organize the text information in graph [Rello et al. 2012] and that highlight the text main concepts [Nandhini and Balasundaram 2013]. [Hyerle 2008] tried to demonstrate how different types of visual tools, called graphic organizers, could help students and teachers represent information and communicate with others. These graphical representations have been applied across a large range of subject areas, demonstrating their benefits in different activities such as mapping cause and effect, note taking, comparing and contrasting concepts, organizing problems and solutions, and relating information to main ideas or themes [Hall and Strangman 2002].

This paper focuses on the use of a graphical representation tool to support reading and writing tasks. This tool intents to assist with text summarization, a learning task that is often proposed with the purpose of reviewing previous learning or preparing students for more conceptual demanding activities [MacArthur et al. 2008]. It is a widely used task that is part of the teachers' collection of activities targeting reading, writing and literacy. [Winograd 1983] showed that students' difficulties in text summarization often happen because of problems in identifying what is important in a text, what should be included in the summaries and how the original text should be transformed.

Our goal is to use text mining techniques in order to extract a representation of the domain knowledge from texts, so that it could be used by students as a starting point for the development of their own graphical representation. By employing a text mining tool to assist students identify and visualize relevant concepts from a text, a higher level of interactivity is introduced in the initial phases of the writing process. The tool used in this work employs a mining technique to identify the most frequent terms and relationships in a text, representing them in the form of a graph. These graphs can then be used as a starting point for the development of the students' own representation of relevant concepts and facts found in the text, elements that are later transformed into a written summary.

2. Graphic Organizers to Support Reading and Writing

Research in Education has shown benefits of using graphic organizers in learning tasks that involve a variety of patterns, such as time/sequence, cause-effect, episodic information, descriptive information, generalization, concept description [Marzano et al. 2001]. The use of graphic organizers and other prewriting activities have also demonstrated to be an effective aid for writing, enabling learners to segment the topic they have to consider, and helping them to structure their writing [Beissner et al. 1994]. Although most of these works usually relate to the use of specific paper-based models, computer-based tools for building some types of graphic organizers have also been proposed along the years.

Based on the Assimilation Theory [Ausubel 1963], [Novak and Cañas 2006] proposed a tool to build concept maps representing propositions about events or objects. A concept map can be seen as a graphic organizer in which labeled nodes repre-

sent concepts and links denote relationships between them [Nesbit and Adesope 2006]. The tool presented by Novak provides many features that made it possible for teachers to use the maps in a variety of activities, including in collaborative learning tasks. [Reader and Hammond 1994] proposes a different approach to concept map building based on the use of hypertext techniques, showing that subjects who used their system obtained better results than those who didn't. A yet different approach was described by [Chang et al. 2001], who created a computer-based concept mapping system that enabled the construction of concept maps in a 'construct-on-scaffold' approach.

Regarding the use of graphic organizers for *reading* comprehension, some learning strategies have been proposed to encourage students to brainstorm prior knowledge by identifying relevant terms related to a certain topic, representing them graphically and then connecting them accordingly. The visualization of the terms and their relationships enable students to have in mind a more concrete representation of the information that could be important to consider in their discussion. In an investigation carried out in a particular American school, students claimed that the organizers were one of the most helpful strategies employed in their learning activities, providing them with visual information which could help them to better understand the subjects at hand [Fisher et al. 2002].

Regarding the use of graphic organizers to support *writing*, [Ruddell 1997] stresses the importance of providing tools that allow students to illustrate their constructions and organization of knowledge, enabling them to express visually which ideas are the most meaningful, and how these ideas are connected. [Capretz et al. 2003] showed that the visualization of information graphically can improve students' organization skills during the writing process.

As for text summarization, the use of graphic organizers (in particularly concept maps) has shown to be an effective method closely related to text comprehension [Chang et al. 2001]. The authors attribute the reason for this to the fact that concept mapping emphasizes the selection of major ideas, connecting and organizing these concepts with links, then using this information for writing. As noted by [Brown and Day 1983], there is an intersection between this sequence of tasks and the process of text summarization.

3. The Text Mining Tool

The text mining tool Sobek has been developed using a particular mining algorithm in which nodes represent the main terms found in the text and the edges used to link nodes represent adjacency information. Therefore, nodes and edges represent how the terms appear together in the text. Previous research has shown promising results regarding the use of Sobek in educational applications, as in the evaluation of students' essays [Macedo et al. 2009] and discussion forums [Azevedo et al. 2014]. Figure 1 shows a graph extracted from a short text about the atomic bomb. In our graphical representation of the graph, nodes which are more relevant are presented in a larger rectangle and in darker color (e.g. the terms "Nuclear", "weapon", "atomic bomb").

Sobek's operation can be divided into three stages. The first one consists in identifying the relevant concepts in the text and summarizes them; the second create relations between those concepts; the last one displays a graph representation of those concepts and their relations.

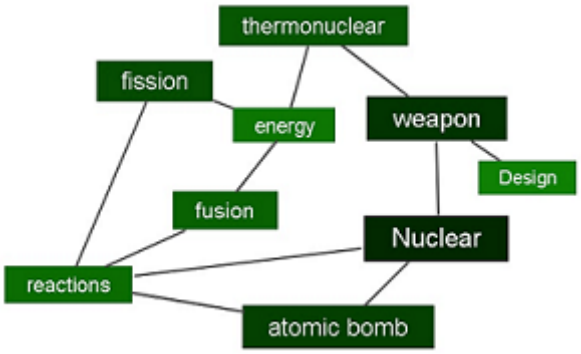

Figure 1. Graph representing relevant terms extracted from a text about the atomic bomb

The first step of Sobek is to split the text into words, using spaces and punctuation as divisors. These words are then mapped into concepts that may consist of a single word (called *"simple concept"*) or many words or sentences (called *"compound concept"*). This mapping is a statistical process, which assesses the frequency that each word is found in the text. When a set of words constantly appear in sequence, the idea associated with this set of words may not be describe by a single concept and a compound concept is formed (e.g. "Global Warming"). The combinations of words that form a compound concept are removed from the word list and the words remaining are considered single concept.

To identify whether a concept is simple or compound, each word is combined a n number of subsequent words, creating strings of words of size 1 to n. For instance, the sequence of words 'AA BB CC' in a scenario where $n = 3$ will create the follow set of strings: $\{$*'AA', 'AA BB', 'AA BB, CC', 'BB', 'BB CC', 'CC'*$\}$. Once this process is completed, the strings with higher frequency are selected and the rest is discarded. A set of words called *"stop words"* is used to remove those that do not add information or should not be displayed in the resulting graph (mainly articles and prepositions).

The last part of the first step consists in summarizing the concepts. The concepts that have a greater number of occurrences are identified, excluding all others. Those concepts will be displayed to the user in the representation graph. The number of concepts can be determined by the user but, according to [Novak and Cañas 2006], no more than 25 concepts should be necessary to identify the focus centered text.

Sobek's second stage is to identify relationships between concepts. The relationship between two concepts implies that there is a connection between them or that they are close related in the text. This connection could represent several information, such as an effect of cause and consequence, a time sequence event or that the concepts are related by their meaning. A new analysis of the text relates two concepts when they are distant not more than a number ϵ of words from each other and when there is no end punctuation between them. Usually a concept will be related to many others and that could produce a graph where the connections would not have a meaning, such as an all connected graph. To solve this problem and display only those connections that are more relevant, a maximum of γ links is permitted for each concept. The exact number of link will be proportional to the frequency of that concept, allowing more frequent concepts (and, therefore,

more important ones) to have a higher number of connections than others.

Sobek's final step is to construct a graph from the extracted concepts. In this graph, the concepts are presented as nodes, and the relation between them, as links between those nodes. To enhance visualization, each node has a different size and color based on its frequency. The larger and darker the node is (varying in hues of green), the higher it relative frequency is when compared to the other nodes frequency. The relative frequency of a node also affects the number of connections that a concept may have, thus making the most frequent concepts even more important and highlighted in the graph.

The graph has a set of functionality that allow the user to personalize it, promoting critical thinking, autonomy and further reflection about the text main ideas. It is possible to add and remove nodes and connections, as well as interact with the graph by changing a node's position.

3.1. Using text mining with summarization method

Summary writing techniques either follow a more intuitive approach without step by step instruction, or follow a rule-governed approach which may focus on tasks such as identifying macro level ideas, deleting unnecessary or redundant information and identifying or producing topic sentences [Bean and Steenwyk 1984]. Here, the method proposed is based on a different approach where the student interacts with a text mining tool in order to grasp the main ideas of the text and to build a visual representation in which these ideas are expressed. Only in a second moment the student moves to the actual writing of the summary.

[Ellis 2003] states that in a writing activity, most of the time spent is dedicated to planning. Aligned with this theory, [Hayes and Flower 1980] divide the writing process in three stages: *pre-writing*, *writing* and *re-writing*.

The use of the software Sobek, as proposed here, focuses on the first two steps of this process. The complete method for text summarization is depicted in Figure 2 and described below.

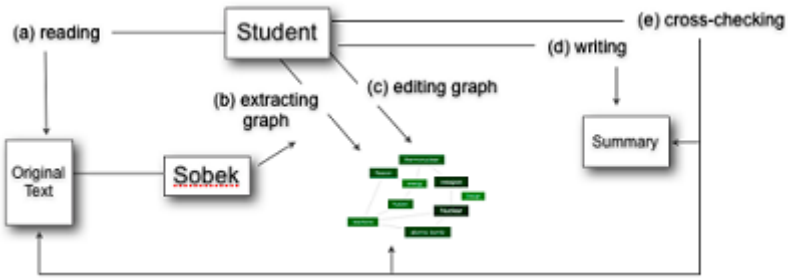

Figure 2. Summarization method using Sobek

Pre-writing:

(a) The student reads a text to be summarized. In this step the student learns about the topic he/she has to write about and identifies macro level ideas.

(b) After reading the text, the student uses Sobek to automatically extract relevant terms and relationships from the text, representing them in a graph. This graph is used as

a first draft of a graphic organizer to help them organize their ideas.

(c) The student reviews the terms and relationships identified by the tool, editing the graph according to what he/she believes to be appropriate. This is a very important step, as it leads the student to reflect about the text and reread it (or portions of it), leading to a deeper understanding of the text.

Writing:

(d) Using the edited graph as a graphic organizer, the student starts the actual writing of the summary. From time to time during the writing process, the student contrasts the graph with the original text, as to make sure that the summary written is faithful to the ideas of the text.

(e) The cross-checking that happens in this phase makes the writing process a cycle, which may involve previous steps in the process, including the re-reading of the text, the re-editing of the graphs extracted by the mining tool, and so on.

The rewriting step, placed by [Hayes and Flower 1980] as the last phase in the writing process, is seen here as a subsequent phase in which the main goal is the revision of the text already structured and written. In this phase, form and style become the most relevant aspects. Our option to focus here in the steps of pre-writing and writing is justified as these are the moments in which the student has to reflect more about the ideas to be considered in the summary, and to structure its main outline. In this sense, the tool may operate as a support to the logical organization of information, a process which relates reading and writing as steps of the same cognitive process [MacArthur et al. 2008].

4. Evaluation and Results

In the summarization experiment, a group of 20 high school students was asked to summarize a short text about the topic "*Realism*". In a first moment, Sobek was presented to the students in order to make them familiar with the mining tool. In a second stage, the students were asked to read the text. Then, by following the same summarization method detailed in section 3.1, students used Sobek to generate their graphs and summarize the text read. It was suggested to students to look at the graphs carefully, observing if the terms and relationships identified by Sobek were in accordance with their reading, eliminating and adding concepts and relationships as appropriate. The students were between 15 and 18 years old. The graph obtained from the text given to the students is presented in Figure 3.

Figure 3. Graph extracted from text about *realism*

Allowing the students to modify the graphs to make them closer to their understanding of the text is similar to the approach proposed by [Chang et al. 2001], where a map-correction strategy was used. In their method, the students used a concept map provided by an expert where many of the nodes and relationships were incorrect, with the goal of letting the learners identify the problems and correct them. Here, however, the graph with the visual representation of the topic to be summarized was not provided by an expert, but by the text mining tool.

The summaries produced by the students were analyzed to verify whether the terms of the graph were in fact present in the students' writings (Table 1).

Table 1. Occurrence of terms in the students' summaries of the text

Terms	*# occurrences*	*Term*	*# occurrences*
realism	100	romance	18
literature.	42	playwright	15
author	34	emphasize	12
theater	34	social	10
naturalism	24	russian	9
romanticism	24	france	5
screen	23	write	5
theme	23	-	-

The results showed that the students used all of the words present in the original graph, composed of 15 terms. Most of the nodes highlighted in the graph also showed a higher frequency of use. It was also noticeable that the students made changes in their graphs while reflecting about the accuracy of the terms and relationships represented.

The student interaction with the computer was monitored by the use of a *screen capture software*. The films obtained from the monitoring of the students interacting with Sobek and using a word processor, also provided subsidies to validate the approach proposed here for summary writing. Two important pieces of evidence were identified in the films, showing how Sobek contributed both to the process of understanding the original text and to the production of the final summary. Concerning the understanding of the text, it was clear that after viewing the graphs produced by Sobek, the students always went back to the text to re-read it. Such behavior implies that the students began by questioning themselves whether a certain term and/or relationship represented in the graph was in fact accurate. Having Sobek to instigate the students to further explore the original text is a positive finding, considering that re-reading leads to a better understanding of the material read and may improve accuracy [Rawson et al. 2000].

As for the use of the graphs in the production of the summaries, the films brought other evidence confirming that the students referred to the graphs in the writing of their texts. Besides the fact that most of the terms represented in the graphs were also found in the students' writings, as shown in Table 1, the films demonstrated that learners went back and forth to their graphs several times while producing their summaries. Such behavior confirms that the students referred to their graphs while writing, which is positive if one considers that the structuring of ideas in graphic organizers may facilitate the more complex task of writing [Ruddell 1997].

According to their teacher, most of the students identified accurately the central theme of the text provided to them. Some of the students' testimonials reinforced this idea:

- "... based on the graph I identified what was important in the text..."
- "... I realized that the words selected by the graph were important, relevant..."
- "... I used the graph, as I wanted to include all of its terms in my text"
- "... I used the graph many times - I had a look at it whenever I did not want to get lost in the text and I wanted things to make sense..."

The testimonials of the language teacher who worked with the students in text production confirmed that the methodology for summary writing using the mining tool was very productive. The teacher stated that normally the students would get worse marks in their essays, and that she was impressed with the level of engagement of all students in the activity proposed.

5. Conclusion

This paper presented a text mining tool and proposed a methodology for using it as a support in summary writing. Other research has shown in the past that diagrams such as concept maps may help students in learning activities in domains as varied as science, statistics and nursing [Nesbit and Adesope 2006]. Our goal here has been different in that we did not want to investigate whether such maps could improve learning, but we wanted to evaluate whether such tools could be used in pre-writing phases of writing activities as a way to help students organize their writing process. Results in different studies demonstrated that the tool was able to produce graphs that were close to what was considered to be important about a text read by the students, but not too perfect as not to give them room to express their own ideas about the most relevant information.

Gao et al. [Gao et al. 2005] also proposed a method for extracting terms from texts automatically, focusing mainly in business applications. Our approach to text mining differs considerably from this method mainly for its representation mechanism based on graphs, and the consequent specificity of its algorithms.

As for the presentation of the mining results, other tools present relevant terms extracted from texts by highlighting these terms in the actual document [Frantzi et al. 2000], or by simply ranking terms through a frequency count. Our solution is based on a visual representation, following the idea of working with graphic organizers. From an educational perspective, presenting the mining results in the form of a graph is interesting as it takes learners to focus on concepts and their relationships, and to reflect about them.

We are currently carrying out further research to define how other types of graphic organizers, such as concept maps, spider maps and affinity diagrams, may be extracted from texts and how they can be used to support text comprehension and text production. Sobek is also being integrated to a virtual learning environment, which will make it available to a large number of students. The observation of how students will use it should give us further insight about possible methods and applications of the text mining technology in educational settings.

Acknowledgement

This project has been partially supported by the following institutions: CNPq, FAPERGS and SEAD/UFRGS.

References

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning.* Grune & Stratton.

Azevedo, B. F., Reategui, E., and Behar, P. A. (2014). Analysis of the relevance of posts in asynchronous discussions. *Interdisciplinary Journal of Knowledge and Learning Objects*, 10:107–121.

Bean, T. W. and Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Literacy Research*, 16(4):297–306.

Beissner, K. L., Jonassen, D. H., and Grabowski, B. L. (1994). Using and selecting graphic techniques to acquire structural knowledge. *Performance Improvement Quarterly*, 7(4):20–38.

Brown, A. L. and Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, pages 1–14.

Capretz, K., Ricker, B., and Sasak, A. (2003). *Improving Organizational Skills Through the Use of Graphic Organizers*. Research Project Saint Xavier. University and Skylight Professional Development.

Chang, K. E., Sung, Y. T., and Chen, S. F. (2001). Learning through computer-based concept mapping with scaffolding aid. *Journal of Computer Assisted Learning*, pages 21–33.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford applied linguistics. Oxford University Press.

Fisher, D., Frey, N., and D., W. (2002). Seven literary strategies that work. *Educational Leadership*, 60(3).

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Gao, X., Murugesan, S., and Lo, B. W. N. (2005). Extraction of keyterms by simple text mining for business information retrieval. In Lau, F. C. M., Lei, H., Meng, X., and Wang, M., editors, *ICEBE*, pages 332–339. IEEE Computer Society.

Hall, T. and Strangman, N. (2002). *Graphic organizers*. National Center on Accessing the General Curriculum, Wakefield.

Hayes, J. R. and Flower, L. S. (1980). Identifying the organization of writing processes.

Hyerle, D. (2008). *Visual Tools for Transforming Information Into Knowledge*. SAGE Publications.

Jacobs, V. A. (2002). Reading, writing, and understanding. *Educational Leadership*, 60(3):58–62.

Jenner, J. (2003). A bridge to reading and writing literacy: Developing oral language skills in young children.

MacArthur, C., Graham, S., and Fitzgerald, J. (2008). *Handbook of Writing Research*. Guilford Press.

Macedo, A. L., Reategui, E. B., Lorenzatti, A., and Behar, P. A. (2009). Using text-mining to support the evaluation of texts produced collaboratively. In *WCCE*, volume 302 of *IFIP Advances in Information and Communication Technology*, pages 368–377. Springer.

Marzano, R., Pickering, D., and Pollock, J. (2001). *Classroom Instruction that Works: Research-based Strategies for Increasing Student Achievement*. Gale virtual reference library. Association for Supervision and Curriculum Development.

Meneghetti, C., Carretti, B., and Beni, R. D. (2006). Components of reading comprehension and scholastic achievement. *Learning and Individual Differences*, 16(4):291 – 301.

Nandhini, K. and Balasundaram, S. (2013). Improving readability through extractive summarization for learners with reading difficulties. *Egyptian Informatics Journal*, 14(3):195 – 204.

Nesbit, J. C. and Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3):413–448.

Novak, J. D. and Cañas, A. J. (2006). The origins of the concept mapping tool and the continuing evolution of the tool. *Information Visualization*, 5(3):175–184.

OECD (2013). Pisa 2012 results: What students know and can do. student performance in mathematics, reading and science (volume i). 1.

Rawson, K. A., Dunlosky, J., and Thiede, K. W. (2000). The rereading effect: metacomprehension accuracy improves across reading trials. *Memory & Cognition*, 28(6):1004–1010.

Reader, W. and Hammond, N. (1994). Computer-based tools to support learning from hypertext: concept mapping tools and beyond. *Computers and Education*, 22.

Rello, L., Saggion, H., Baeza-Yates, R., and Graells, E. (2012). Graphical schemes may improve readability but not understandability for people with dyslexia. *NAACL-HLT*.

Ruddell, M. R. (1997). *Teaching Content Reading and Writing*. Wiley.

SERCE (2008). Second regional comparative and exploratory study: Student achievement in latin america and the caribbean. santiago: Orealc/unesco. 1.

Sticht, T. (1974). *Auding and Reading: A Developmental Model*. Human Resources Research Organization.

Vieira, M. C. T. (1981). Levantamento das dificuldades de alunos do 1º ano da universidade na compreensão de materiais escritos.

Warschauer, M. (2006). *Laptops and Literacy: Learning in the Wireless Classroom*. New York: Teachers College Press.

Winograd, P. N. (1983). *Strategic difficulties in summarizing texts*. University of Illinois at Urbana-Champaign ;Cambridge, Mass.