American Journal of Educational Research, 2015, Vol. 3, No. 12, 1592-1596
Available online at http://pubs.sciepub.com/education/3/12/16

DOI:10.12691/education-3-12-16

Automatic Extraction of Nonlinguistic Representations of Texts to Support Writing

Eliseo Reategui*, Daniel Epstein

PGIE, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
*Corresponding author: eliseo.reategui@ufrgs.br

Abstract Graphic organizers may be a helpful instrument in assisting students to structure their text productions. This paper presents a mining tool that is able to extract relevant terms and relationships from texts and shows how the tool may be used to help students in writing summaries. A particular tool is presented, explaining the technique used to analyze the texts, to extract relevant terms and build graphs from them. These graphs are then used by students as graphic organizers, helping them to reflect about the main ideas of the text before the actual writing task. Two experiments are presented in this paper, in which a total of 29 students were asked to read and summarize a short text with the assistance of the text mining tool. Results demonstrated that the tool helped students reflect about the main ideas of the text and supported the actual writing of the summaries.

Keywords: *summarization, writing, text mining, graphic organizers*

Cite This Article: Eliseo Reategui, and Daniel Epstein, "Automatic Extraction of Nonlinguistic Representations of Texts to Support Writing." *American Journal of Educational Research*, vol. 3, no. 12 (2015): 1592-1596. doi: 10.12691/education-3-12-16.

1. Introduction

For many years researchers have been investigating how nonlinguistic representations could help students in reading and writing tasks. Marzano, Pickering & Pollock [1], for example, discussed the importance of using nonlinguistic representations to help students enhance their understanding of written material. Hyerle [2] 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 [3].

This research focuses on the use of a particular graphical representation to support writing tasks. More specifically, the paper is centered on text summarization, a learning task that is often proposed with the purpose of reviewing previous learning or preparing students for more conceptual demanding activities [4]. Winograd [5] has shown 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. Some of these are "higher order" difficulties that could be related to a number of possible sources, including different language experience, lack of prior knowledge or lack of strategic skills [6]. We focus here mostly on helping students develop strategic skills to identify the main topics in a text, to organize their ideas and later to place these ideas in their own writing.

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 employs a mining technique originally designed by Schenker [7] 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.

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 [1]. 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 [8]. Based on the Assimilation Theory, Novak [9] proposed a tool to build concept maps representing propositions about events or objects. The tool provided many features that allowed students to work in the representation of concepts and relationships in a variety of activities, including in collaborative learning tasks.

A different approach was described by Chang, Sung and Chen [10], who created a computer-based concept mapping system that enabled the construction of concept maps in a 'construct-on-scaffold' approach. The method proposed was based on the presentation of an incomplete concept map to the students, in which some nodes and

links were set as blank spaces. The students then had to complete the map according to their understanding of the subject studied.

Regarding the use of graphic organizers to support writing, Rudell [11] 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, Ricker and Sasak [12] showed that the visualization of information graphically can improve students' organization skills during the writing process.

These works propose systems and methods that are based on the use of simple graphic tools to allow the users to create their own representation of propositions about facts. None of these systems keep an internal knowledge representation that can be used by students as a starting point for the development of their own graphical representation and subsequent writing. The next section presents Sobek, a tool which has been developed to build graphs from texts, using the technique of text mining.

2. The Text Mining Tool Sobek

The text mining tool Sobek has been developed as part of this research, using a particular mining algorithm based on the n-simple distance graph model, in which nodes represent the main terms found in the text, and the edges used to link nodes represent adjacency information [7]. Previous research has shown promising results regarding the use of Sobek in educational applications, as in the evaluation of students' essays [13] and discussion forums [14]. Two examples of graphs obtained with Sobek are presented in Figure 1 and Figure 2.

The operation of Sobek can be divided into three steps, in which the first is to separate text into words, using spaces and punctuation marks as divisors. These words are then mapped into concepts that may consist of a single word (called a "simple concept") or many words or sentences (called "compound concept"). This mapping is a statistical process, which assesses the frequency with which 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 described by a single word and a compound concept is formed (e.g. "greenhouse effect"). During the process of identification of concepts, a set of words called "stop words" is used to remove those that do not add relevant information (as articles or prepositions). After identifying all different concepts, a stemming method is performed to reduce redundancy and remove concepts with the same meaning.

The second step of the mining process is to identify relationships between concepts. A new analysis of the text relates two concepts when they are distant not more than 'z' words of each other and when there is no punctuation mark between them. To reduce the number of connections among concepts and display only the most frequent ones, a maximum of 'r' links is allowed for each concept, and this value is proportional to the frequency of that concept. Thus, each concept has no more than r connections multiplied by its frequency, divided by the higher frequency of all terms. There is no lower boundary to the number of times a relationship between concepts must occur in the text to be considered a link in the graph. Sobek's default setting uses 'z = 5' and 'r = 7'. More about the operation of the tool can be found in [15].

The last step of the mining process is to produce a visual representation of the results in a graph, in which the concepts are presented as nodes and the relationships between them are represented as edges. To improve the visualization of the graph, each node has a different size and color. The larger and darker the node is, the higher is the frequency of the corresponding term.

The graphs built by Sobek may be edited by adding/removing their concepts, or by adding/removing relationships between them. While other text mining approaches rely on the analysis of relevant morph syntactic patterns to generate compound terms for the mining process, here we used a simpler method that computes the frequency with which these compound terms appeared in the text. The next section presents the method for using Sobek to help students in text summarization tasks.

3. The Summary Writing 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, identifying or producing topic sentences [16]. The method proposed here is based on a different approach where the student interacts with Sobek 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.

It has been argued that in a writing activity most of the time spent is dedicated to planning [17]. Previous to that, other authors proposed a partitioning of the writing process in three stages: pre-writing, writing and re-writing [18]. The use of the software Sobek focuses on the first two steps of this process. In a Pre-writing stage, 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. After reading the text, the student uses Sobek to extract automatically 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. The student then 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.

In a writing phase, the student uses the edited graph as a graphic organizer to start 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. 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, defined as the last phase in the writing process [18], 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 [19].

4. Evaluation and Results

Two studies were conducted in order to evaluate whether Sobek could effectively support students in summarization tasks. The first experiment was carried out with 9 undergraduate Pedagogy students of a large public university in Brazil, with ages ranging from 18 to 24 years old. In a first moment, Sobek has been presented to the students in order to make them familiar with the mining tool. In a second stage, the students were asked to read an introductory text on knowledge and scientific method. Then, by following the same summarization method explained in section 4, 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 graph obtained from the mining of the text can be seen in Figure 1.

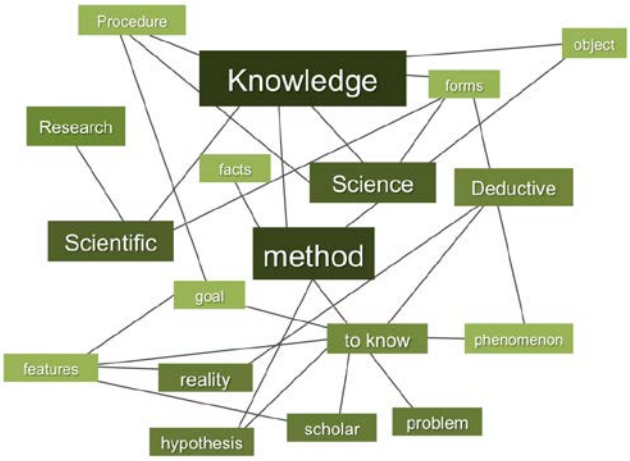

Figure 1. Graph extracted from text about scientific knowledge

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

Terms	# occurrences	Terms	# occurrences
research	26	facts	5
knowledge	25	problem	5
method	23	goal	5
scientific	22	features	5
deductive	12	to know	5
object	11	form	5
hypothesis	10	science	3
reality	8	procedure	1
inductive	6	scholar	0
phenomenon	6	-	-

After the graph editing phase, students used their version of the graph to make the summary of the text. As a first validation step, the summaries produced by the students were analysed to verify whether the terms of the graph were present in the students' writings. It was observed that the large majority of the terms that appeared in the graph were used by the students, according to the distribution shown in Table 1. Most of the nodes highlighted in the graph also showed a higher frequency of use.

Some terms were either not used or used scarcely, such as "scholar" or "procedure". However, synonyms were used, as "researcher" for "scholar", or "steps" for "procedures". This shows the relevance of many of the terms selected by the tool, but also shows that the students did not follow the graph strictly. In the editing step, all students removed one or more terms they did not consider important, and only four participants added terms. The fact that the students made changes in their graphs is a positive finding, considering that such an action is the result of a reflection about the accuracy of the terms and relationships represented. Although some changes were made in the graphs, the teacher observed in the students summaries that all of them were able to identify and write about the main theme of the text provided.

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. [10], 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 graphs with the visual representation of the topic to be summarized was not provided by an expert, but by the text mining tool.

At the end of the experiment, students responded to an online questionnaire about their perception of Sobek as a tool to support summarization tasks. The questionnaire consisted of 6 statements, which aimed to observe 5 levels of agreement or disagreement based on a 5-point Likert scale (from strongly disagree to strongly agree). Out of the 9 respondents, all of them agreed or strongly agreed that the terms extracted by the graph led to a greater reflection on the text (9). Most of the respondents also thought that the use of the tool led them to reread the text more times (7), facilitated the summarization task (8) and helped them produce a better summary than if the tool were not used (8). In addition, the majority of students agreed or strongly agreed that they did use the graphs generated by the tool in the construction of the final text (8). All of the students agreed or strongly agreed that they would use Sobek again to help them summarize texts (9).

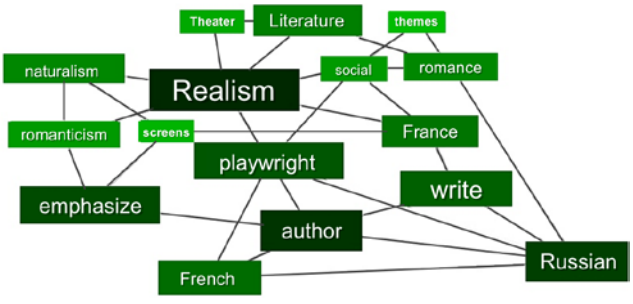

Figure 2. Graph extracted from text about realism

In the second experiment, another group of 20 high school students was asked to summarize a short text about the topic "realism". The students were between 15 and 18 years old, and they were asked to use Sobek to help them write a summary of the text, following the steps detailed in section 4. The graph obtained from the text given to the students is presented in Figure 2.

The summaries produced by the students were analysed to verify whether the terms of the graph were in fact present in the students' writings (Table 2).

Table 2. Occurrence of terms in the students' summaries of the text about scientific knowledge

Terms	# occurrences	Terms	# 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. Once again, most of the nodes highlighted in the graph also showed a higher frequency of use. But it is also noticeable that some of the terms did not appear in the texts of all 20 students. In this experiment it was again noticeable that the students made changes in their graphs while reflecting about the accuracy of the terms and relationships represented.

This time 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 [20].

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 [11].

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..."*
- *"... at the end of the activity I needed the graph to know what were the important parts that had to be included in the text..."*

The testimonials of the Portuguese 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 [21]. 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. [22] 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 [23], 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.

State of Competing Interests

The authors have no competing interests

Acknowledgement

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

References

[1] Marzano, R.J., Pickering, D.J. and Pollock, J.E, *Classroom Instruction that Works: Research-Based Strategies for Increasing Student Achievement*, Association for Supervision and Curriculum Development, Alexandria, 2001.

[2] Hyerle, D, *Visual Tools for Transforming Information into Knowledge*, Corwin Press, Thousand Oaks, 2009.

[3] Hall, T. and Strangman, N, *Graphic organizers*, National Center on Accessing the General Curriculum, Wakefield, 2002.

[4] Newell, G.E, "Writing to Learn: How Alternative Theories of School Writing Account for Student Performance", in MacArthur, C., Graham, S. and Fitzgerald, J. (Eds), *Handbook of Writing Research*, The Guilford Press, New York, 2008, 235-247.

[5] Winograd, P.N, *Strategic Difficulties in Summarizing Texts*, Technical Report No. 274, Center for the Study of Reading, University of Illinois, Urbana, 1983.

[6] Collins, A. and Haviland, S.E, *Children's reading problems*, Reading Education Report No. 8, University of Illinois, Urbana, 1979.

[7] Schenker, A, *Graph-Theoretic Techniques for Web Content Mining*, Unpublished PhD thesis, University of South Florida, Tampa, 2003.

[8] Beissner, K., Jonassen, D.H. and Grabowski, B.L, "Using and Selecting Graphic Techniques to Acquire Structural Knowledge", *Performance Improvement Quarterly*, 7(3-4). 20-38. 1994.

[9] Novak, J.D, "The Origins of the Concept Mapping Tool and the Continuing Evolution of the Tool", *Information Visualization Journal*, 5(3). 175-184. 2006.

[10] Chang, K.E., Sung, Y.T. and Chen, S.F, "Learning through computer-based concept mapping with scaffolding aid", *Journal of Computer Assisted Learning*, 17(1). 21-33. 2001.

[11] Ruddell, M.R, *Teaching content reading and writing*, John Wiley & Sons, New York, 2001.

[12] Capretz, K., Ricker, B. and Sasak, A, *Improving organizational skills through the use of graphic organizers*, Research Project, Saint Xavier University and Skylight Professional Development, Chicago, 2003.

[13] Macedo, A., Reategui, E., Lorenzatti, A. and Behar, P.A, "Using text-mining to support the evaluation of texts produced collaboratively", in Tatnall, A. and Jones, A. (Eds), *Education and Technology for a Better World*, Springer, 2009, 368-377.

[14] Azevedo, B.F.T., Reategui, E. and Behar, P.A, "Analysis of the relevance of posts in asynchronous discussions", *Interdisciplinary Journal of Knowledge and Learning Objects*, 17(1). 107-121. 2014.

[15] Reategui, E., Epstein, D, Lorenzatti, A. and Klemann, M, "Sobek: a Text mining tool for educational applications", in *International Conference on Data Mining*, CSREA Press, 2011, pp. 59-64.

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

[17] Ellis, R, *Task-based language learning and teaching*, Oxford University Press, New York, 2003.

[18] Hayes, J.R. and Flower, L.S, "Identifying the organization of the writing process", in Gregg, L.W. and Steinberg, E.R. (Eds), *Cognitive processes in writing*, Lawrence Erlbaum Associates, 1980, 3-30.

[19] Shanahan, T, "Relations among Oral Language, Reading, and Writing Development", in MacArthur, C., Graham, S. and Fitzgerald J. (Eds), *Handbook of Writing Research*, The Guilford Press, 2008, 171-183.

[20] Rawson, K.A. and Dunlosky, J, "The rereading effect: Metacomprehension accuracy improves across reading trials", *Memory & Cognition*, 28(6). 1004-1010. 2000.

[21] Nesbit, J.C. and Adescope, O.O, "Learning with Concept and Knowledge Maps: A Meta- Analysis", *Review of Educational Research*, 76(3). 413-448. 2006.

[22] Gao, X., Murugesan, S. and Lo, B, "Extraction of Keyterms by Simple Text Mining for Business Information Retrieval", in *IEEE International Conference on e-Business Engineering*, IEEE Press, 2005, 332-339.

[23] Frantzi, K., Ananiadou, S. and Mima, H, "Automatic recognition of multi-word terms", *International Journal of Digital Libraries*, 3(2). 117-132. 2000.