Modeling Summarization Assessment Strategies with LSA

S. Mandin, B. Lemaire and Ph. Dessus

***Abstract*—This paper presents a model based on LSA which attempts to simulate the way humans assess student summaries. It is based on the automatic detection of 5 cognitive operations that student may use in writing a summary. Comparisons with data from 33 human raters show the strengths and limits of this approach.**

I. Introduction

THERE is a large literature on how computers could help writing summaries : either by automatically performing summarization (e.g., Endres-Niggemeyer & Wansorra, 2004) or by assessing student summaries (e.g., Wade-Stein & Kintsch, 2004). However, computer models of the strategies used by teachers to assess students' summaries are yet lacking. This kind of model is more difficult to implement because it has several complex goals: it has first to represent the most important ideas of a text (i.e., sentences/propositions hierarchisation), then to implement a cognitive model of summarization skills (i.e., what kind of operations to perform on these sentences/propositions) and finally to model the teachers skills that lead to assess the summary as a result.

We claim that *Latent Semantic Analysis* (Landauer & Dumais, 1997) is an adequate way to perform all these tasks, since it has been successfully tested as a cognitive model of the representation of knowledge, both static (i.e., knowledge represented in a text) and transient (i.e., knowledge built by students in performing summaries or by teachers in assessing them). In a first experiment (Lemaire et al., 2005), we tested four models of summarization assessment, which were all tested on students' productions. However an actual validation of human assessment skills was lacking. This paper is devoted to such an aim.

II. Description of the Model

During reading, the macrostructure of the text is built and updated (Kintsch & van Dijk, 1978). Since this macrostructure can be considered as a summary, we used it for modeling purposes. Three macrorules, i.e. mental operations on the source text, were involved: the *deletion* of minor propositions, the *generalization* of several propositions into a superset idea and the *construction* of a new proposition denoting a global fact about events described by several propositions. Three summary-specific operations were added: the *copy* of a part of the text, the lexical or syntactic transformation of a sentence without modifying its meaning (*paraphrase*) and the production of *off-the-subject* sentences (Brown & Day, 1983).

These macrorules can either be used for automatic summarization purposes (e.g. Hutchins, 1987) or, in our case, for supporting the assessment of student summaries. We implemented these macrorules in the LSA framework in the following way:

--A *copy* is a summary sentence which is semantically very close to a source text sentence;

--A *paraphrase* is a summary sentence which is close to only one source text sentence;

--A *generalization* is a summary sentence which is close to several source text sentences;

--A *construction* is a summary sentence which is close to no source text sentences but is at least related to one of them;

--An *off-the-subject* sentence is a summary sentence which is not close to any source text sentences.

There is actually another mental operation which is not visible in the summary, namely the *deletion*, but we will not take it into account in this paper. Three similarity thresholds separate the different operations. Figure 1 gives an example of semantic distances (ranging from 0 to 1) between each summary sentence and the different source text sentences. Thresholds will be empirically determined by confronting our model to human data. We first assume that they are rater-independent.

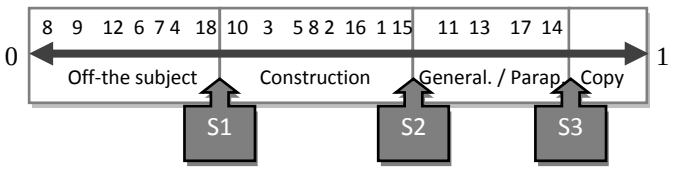

Fig. 1. Representation of the comparisons between a given summary sentence and each source text sentence (represented by numbers). In this example, the summary sentence is classified as a generalization since it is close to several source text sentences.

III. Validation of the Model

33 post-graduate students in educational science from our

Sonia Mandin is with the "Laboratoire des sciences de l'éducation" (EA 602) in the university of Grenoble, France. (e-mail: Sonia.Mandin@upmf-grenoble.fr).

Benoît Lemaire is with TIMC-IMAG (CNRS UMR 5525) in the university of Grenoble, France. (e-mail: Benoit.Lemaire@imag.fr).

Philippe Dessus is with the "Laboratoire des sciences de l'éducation" (EA 602) and the IUFM in the university of Grenoble, France. (e-mail: Philippe.Dessus@upmf-grenoble.fr).

university were given the following task. They were given two summaries of a same source text (either a narrative text for 15 raters, or an expository one for the 18 others) and had to guess what were the macrorules used by their authors (11^{th} grade students). In order to reduce the inter-rater variability, raters had to refer to a grid in which the different macrorules were described without any technical vocabulary. Data were processed as follows. First, raters' judgments about macrorules use were coded (ranging from 1, copy, to 5, off-the-subject). Second, all possible thresholds triplets (between 0 to 1, $s1 < s2 < s3$, with a .05 step) were computed, based on a 13 million-word corpus composed of a children corpus (3 million words), newspaper texts (5 million words) and novel (5 million words), using the Bellcore implementation. Finally, a rater-model agreement was computed (Spearman correlation), and the 3 thresholds leading to a maximum of the highest correlations beyond .60 were kept.

The results are mixed. First, the inter-rater agreement is low: 39% and 63% for expository texts ratings, and slightly better for narrative texts ratings: 80% and 53%. Second, our threshold-based model appears to be relevant only for expository texts ratings: 33% and 63% of raters correlate with the model at the same thresholds ($s1 = .05$; $s2 = .10$; $s3 \in [.80; .85]$). These percentages are not lower than those of inter-rater agreements (39% and 63%). However, the threshold values for the narrative texts for which the number of model-raters correlations is maximum are different for the two summaries: $s1 = .05$, $s2 = .10$, $s3 \in [.50; .70]$ for summary 1, and $s1 = .05$, $s2 \in [.55; .65]$, $s3 \in [.60; .65]$, or $s1 \in [.55; .60]$, $s2 \in [.60; .65]$ or $s3 \in [.65; .95]$, for summary 2. Besides, for both cases, the percentage of raters who correlate beyond .60 with the model is weak (27% for one of the summary and 20% for the other).

These results show that our model only fits with expository text data: its performance is close to human one. Since this kind of texts is often about a unique subject, each sentence is highly related to the whole source text. Therefore, our model adequately selects the category of the summary sentences. On the other side our model is inadequate to assess narrative texts because they deal with a lot of different themes throughout the story (Pinto Molina, 1995). Raters may likely assess the similarity of summary sentences inside a narrative sequence not based on the whole text. Two summary sentences that do not refer to the same sequence of the source text would be semantically distant for the raters whereas they would be linked for LSA as long as they would be composed of some similar words. These results have to be confirmed with the assessment of more summaries.

IV. Towards a Learning Environment

This model could be embodied in a learning environment that would help teachers assess summaries. Novice teachers often lack methods for achieving this task. The goal is to focus them to uncover cognitive processes that are likely performed by students rather than to help them deliver summative assessments. We designed a prototype interface hooked up to LSA to reach this goal. Our system teaches students to rely on the aforementioned five categories that are based on sound psycholinguistic theories. The system presents two adjacent panes: the source text and a summary. Summary sentences are colored according to the categories the model judges they belong to. The three thresholds that define the boundaries between categories are visualized and the user would be requested to adjust them according to her idea of what is a copy, an off-the subject sentence, etc. Sliding a boundary with the mouse would obviously change the category of some sentences and their color would immediately change on the screen. In case a sentence is not correctly classified by the system, the user would be able to force its category. The threshold values set by the user for different summaries would be highly valuable. They would tell us to what extent these values are user-dependent or summary-dependent.

The goal is not to indicate to the user the category of each summary sentence, but rather to engage them in the process of identifying categories. This learning environment could be viewed as an assistant to the task of categorizing summary sentences.

References

[1] A. L. Brown and J. D. Day, "Macrorules for summarizing texts: The development of expertise," *J. Verb. Learn. Verb. Behav.*, vol. 22, pp. 1–14, 1983.

[2] B. Endres-Niggemeyer and E. Wansorra, "Making cognitive summarization agents work in a real-world domain," presented at the Natural Language Understanding and Cognitive Science Conference (First NLUCS Workshop), Porto, 2004.

[3] J. Hutchins, "Summarization: Some problems and methods," in *Meaning: the frontier of informatics. Informatics 9*, K. P. Jones, Ed. London: Aslib, 1987, pp. 151–173.

[4] W. Kintsch and T. A. van Dijk, "Toward a model of text comprehension and production," *Psychol. Rev.*, vol. 85, pp. 363–394, 1978.

[5] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge*," Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.

[6] B. Lemaire, S. Mandin, Ph. Dessus and G. Denhière, "Computational cognitive models of summarization assessment skills," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society (CogSci' 2005)*, B. G. Bara, L. Barsalou and M. Bucciarelli, Ed. Mahwah: Erlbaum, 2005, pp. 1266–1271.

[7] M. Pinto Molina, "Documentary abstracting : toward a methodological model*," J. Am. Soc. Inform. Sci.*, vol. 46, pp. 225–234, 1995.

[8] D. Wade-Stein and E. Kintsch, " Summary Street: Interactive Computer Support for Writing*," Cognition and Instruction*, vol. 22, no. 3, pp. 333–362, 2004.