What is behind a summary-evaluation decision?

IRAIDE ZIPITRIA, PEDRO LARRAÑAGA, RUBEN ARMAÑANZAS,
ANA ARRUARTE, AND JON A. ELORRIAGA
University of the Basque Country, Donostia, Spain

Research in psychology has reported that, among the variety of possibilities for assessment methodologies, summary evaluation offers a particularly adequate context for inferring text comprehension and topic understanding. However, grades obtained in this methodology are hard to quantify objectively. Therefore, we carried out an empirical study to analyze the decisions underlying human summary-grading behavior. The task consisted of expert evaluation of summaries produced in critically relevant contexts of summarization development, and the resulting data were modeled by means of Bayesian networks using an application called Elvira, which allows for graphically observing the predictive power (if any) of the resultant variables. Thus, in this article, we analyzed summary-evaluation decision making in a computational framework.

Among the different educational assessment methodologies, giving learners an open-ended writing task offers them the freedom to write anything they know, as well as the implicit responsibility to construct their own responses. However, data obtained when grading such texts are hard to quantify with precision, and such evaluation is therefore considered more subjective than evaluation via other methodologies. But, is there any common ground for human summary-evaluation decision making? What should we take into consideration to model human summary-evaluation decisions?

Experts in text-grading practice report a need for focusing on the identification and implementation of features of effective performance-based assessment and standards when trying to understand the impact of scoring (Goldberg & Roswell, 1999). In addition, according to Chung and Baker's (2003) study of automatic grading, there is a need to produce evidence that the scores produced through this system faithfully reflect the intended use of those scores. A method to do this is to base the dimensions of scoring rubrics on the performance of experts who possess the desired knowledge, not only in terms of content, but also in terms of cognition.

Most research into automatic grading is very much focused on obtaining single measures for text coverage, cohesion, coherence, and so forth. But, in addition to this identification of discourse parameters, a decision-making step has also been found to be necessary in conventional grading (Genesee & Upshur, 1996).

In this article, we specifically analyze what is involved during the decision-making process in summary grading. Our aim is to further define a performance-based model of the considerations taken by expert human graders when determining a global summary score. Evaluation decisions are modeled by means of a type of Bayesian network (Pearl, 1988) that has already been shown to be capable of explaining psychological phenomena (Glymour, 2001). Further work utilizing Bayesian classifiers to deal with text grading can be found in Burstein and Marcu (2003). The Bayesian approach allows for observing relations among the summary-evaluation variables and can account for human uncertainty. As a result, it allows for the classification process to go on, even when some of the variables are not present. This model has been observed and studied with the aid of a graphical application called Elvira (Elvira Consortium, 2002).

The article is organized as follows. In Section 1, we survey some previous work on summarization and summary assessment. Section 2 is devoted to a theoretical analysis of Bayesian networks. In Section 3, we draw conclusions from interviews with experts. Section 4 contains human summary-assessment performance data. Finally, in Section 5, we present the conclusions from our research and an analysis of possibilities for future work.

1. SUMMARY EVALUATION

Little work has been done on summary evaluation as such, but summarization has already been widely studied.

Human summarization is a learning strategy that is commonly used to measure text comprehension in educational practice. It involves a variety of different abilities: understanding, abstraction, organization, and the reproduction of information. Summaries can be either spoken or written, and summary maturity varies with age, abstraction ability, language ability, and learning ability, among other factors. Summarization takes place in a variety of forms, such as headlines, outlines, minutes, previews, digests, biographies, abridgments, bulletins, histories, and so on (Mani & Maybury, 1999).

A successful learning strategy helps with content comprehension. The information contained in a summary gives a broad idea of what has been understood by learners—that is, what has been retrieved from a text and what has not (Garner, 1982). Because of this, summarization is one of the most popular methods to evaluate text comprehension and content understanding.

Much has been written about what a good summary should and should not contain. However, human summarizers do not always produce what would be expected from a prototypical summary. Many authors have studied maturity as a factor to describe the similarities and differences among summarizers. Acquisition of the ability to summarize is part of the more general acquisition of writing ability. Therefore, it shares common features with other types of expression, such as essay writing.

Thus far, a great amount of work has been done on text structure, summarization, and text comprehension modeling. Bartlett (1932) was one of the first psychologists to analyze text retrieval and the processes that underlie the quality of the information retrieved. Rumelhart (1975) produced one of the first models to explain narrative summarization procedure. Thorndyke (1977) explored text retrieval on the basis of text plot complexity. Kintsch and van Dijk (1978) proposed a holistic model for text representation and structure; this model describes the mental representation of summaries and the major difficulties on the way toward mature summarization. Schank, Lebowitz, and Birnbaum (1980) took a different approach, producing a parser able to search for the most relevant information in text, according to the premise that some information is more relevant than other information. Finally, Lehnert (1981) produced a model that explains narrative summarization by taking into account emotional and affective variables.

In addition, plenty of empirical studies have tested and widened the conclusions from these models. For instance, Garner (1982, 1987) studied differences in maturity. According to her data, highly efficient summarizers not only summarized more efficiently, but also stored information in memory more efficiently. She argued that effective summarization implies effective memorization. Similarly, Brown and Day (1983) tested some conclusions provided by Kintsch and van Dijk's (1978) model and confirmed that young learners have difficulties with critical reading and effective studying. Later, Manelis and Yekovich (1984) analyzed expository text in relation to comprehension and learning processes, using key concepts to detect learning. Moreover, Sherrard (1989) provided an overall view of summarization by combining analyses of summarizer strategies and performance improvement with assessments of summary quality. Finally, Bransford, Vye, Kinzer, and Risko (1990) explained what students learn from text and how they are guided through the learning process itself.

The key issue on which research work has concentrated, however, is the difference between a poor and a good summary. Many researchers in this area have made a clear distinction between immature and mature summarizers. Yet how do we identify a mature versus an immature summary? When evaluating summaries, a major focus has been placed on use of a topic sentence and main-idea identification, and the right use of those skills has been taken as indicative of mature summarization (Manelis & Yekovich, 1984). Topic sentences are normally placed at the beginning of the text, although there is no fixed criterion in this respect. Elosúa, García-Madruga, Gutiérrez, Luque, and Gárate (2002) reported that students in secondary education tend to identify the main idea more easily when it appears at the beginning of the text. Those authors argued that the reason behind this pattern is that many readers expect to find the main idea in the first paragraphs of the text.

In a similar way, Garner (1982) argued that highly efficient summarizers recognize true information that does not appear in the source text a higher proportion of the time than do less efficient summarizers. Therefore, immature summarizers' difficulties are mainly related to comprehension and remembering. For instance, such summarizers have great difficulties differentiating superordinate from subordinate information (Taylor, 1982). Another tendency observed in poor summarizers is the *knowledge-telling strategy* (Brown & Day, 1983), which involves writing everything they know or remember about a reading text, resulting in a huge amount of irrelevant information and lack of abstraction. In conclusion, it seems clear that some signs of a poor summary are large amounts of irrelevant information, copying parts of the text, and comprehension failures. Consequently, a good summary should not show such performance markers.

Language proficiency can also make the difference between a mature and an immature summary. For instance, second language (L2) summarizers are faced with comprehension failures and lack of grammar and lexicon knowledge that they would not necessarily experience in their first language. Although the final result of their efforts might look similar to that of a monolingual but immature summarizer, the reasons behind the problems have been shown to be different (Kozminsky & Graetz, 1986; Long & Harding-Esch, 1978). Therefore, specific training and evaluation appear to be necessary. Overall, L2 learners' summaries can suffer from a number of deficits: information less well selected for relevance, less efficient language processing, and poorer use of language in summarization and recall.

Prior knowledge has also been found to be a factor related to producing a successful summary; it is related to category selection, facilitates information extraction, and reduces working memory demands (Symons & Pressley, 1993). Therefore, previous familiarity with content can determine text comprehension and main-idea identification. This is why teachers pay special attention when selecting reading texts, trying to find those that better match their students' backgrounds. This idea applies not only to content knowledge, but also to language proficiency.

Sherrard (1989) focused attention on text comprehension, arguing that expert summarizers make decisions

on the basis of the whole text, whereas poor readers and youngsters mainly look at sentences and details. She concluded that a mature summary should include three principal components: content, structure, and style features.

The absence or presence of a reading text is another relevant feature to bear in mind when evaluating summaries. It has been argued that, whereas text-present summarization encourages shallow processing in memory, text-absent summarization leads to deep processing (Kirby & Pedwell, 1991). Depending on the learning goals, one or the other of these methods might be chosen. Kirby and Pedwell stated that deep learners might prefer the text-absent mode, whereas surface learners should learn more with text-present summarization. Other issues to take into consideration when planning instruction are text readability and learners' approaches to learning. Consequently, the use of a text-present approach is recommended for unskilled learners.

What does human summary evaluation involve, though? This task involves the ability to produce an adequate mental summary, on one hand, and the ability to detect summarizers' success and difficulties, on the other. Hence, the task of summary evaluation should be assigned to a person who has mastered the ability to summarize and is able to evaluate all of its components: the summarization expert.

Little work has been done on summary assessment as such. Thus far, however, Sherrard (1989) reported poor interrater agreement on summary evaluation. The present study takes a Bayesian-network-based approach in order to observe the predictive power of the resultant variables.

2. BAYESIAN NETWORKS

The obtained summary-evaluation decisions are modeled by means of a type of Bayesian network (Pearl, 1988) that has already been shown capable of explaining psychological phenomena (Glymour, 2001). This approach allows for the observation of relations among variables and can account for uncertainty. It can deal with missing values, and the generalization ability of Bayesian networks is also less sensitive to overfitting. Therefore, a Bayesian approach tends to be stronger when analyzing small samples. Moreover, using a graphical application called Elvira (Elvira Consortium, 2002) allowed us to graphically observe and study the variable relations. From this approach, it was possible to perform the classification process, even when all of the variable values were not known.

2.1. Introduction

For those who are not familiar with Bayesian networks, the following section introduces this type of probabilistic graphical model (Jensen, 2001; Neapolitan, 2003; Pearl, 1988). Such models have been used for the last decade for analysis in domains in which uncertainty is intrinsic.

Probabilistic graphical models represent multivariate joint probability distributions via a product of terms, each of which involves only a few variables. The structure of the product is represented by a graph that relates the variables that appear in a common term. This graph specifies the product form of the distribution and also provides tools for reasoning about the properties entailed by the product (Lauritzen & Spiegelhalter, 1988). For a sparse graph, the representations in these models are compact and in many cases allow for effective inference and learning.

2.2. Bayesian Networks

In Bayesian networks, the joint distribution over a set $X = (X_1, \ldots, X_n)$ of random variables is represented as a product of conditional probabilities. A Bayesian network associates with each variable X_i a conditional probability distribution $p(X_i = x_i \mid \mathrm{Pa}_i = \mathrm{pa}_i)$, where $\mathrm{Pa}_i \subset X$ is the set of variables that are called the *parents* of X_i. Intuitively, the values of the parents directly influence the choice of the values of X_i. Hence, the resulting product is of the form

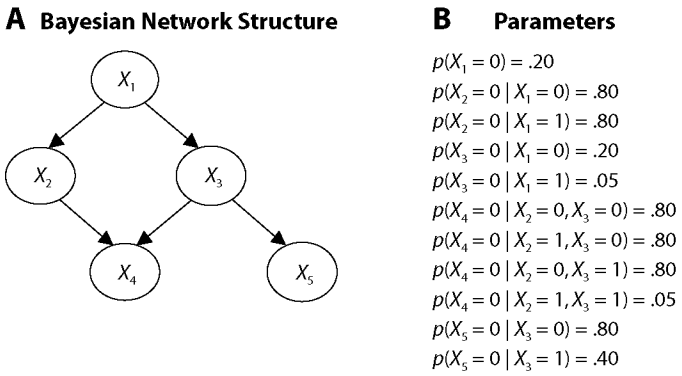

Figure 1. (A) Structure of a sample Bayesian network. (B) An achieved joint probability factorization for this network: $p(X_1, X_2, X_3, X_4, X_5) = p(X_1) \cdot p(X_2 \mid X_1) \cdot p(X_3 \mid X_1) \cdot p(X_4 \mid X_2, X_3) \cdot p(X_5 \mid X_3)$.

$$p\left(X_1 = x_1, \ldots, X_n = x_n\right) = \prod_{i=1}^{n} p\left(X_i = x_i \mid \mathrm{Pa}_i = \mathrm{pa}_i\right).$$

A graphical representation of such a network is given by a directed graph, in which we put lines from X_i's parents (Pa_i) to X_i—see Figure 1. As shown in this figure, the approach reduces the number of variables needed to obtain the joint distribution over the five nodes. More specifically, the use of the Bayesian network in Figure 1 reduces the number of variables from 31 to 11.

To assess a Bayesian network, it is necessary to specify two things. (1) A structure for the network must be specified by means of a directed, acyclic graph that reflects the set of conditional (in)dependencies among the variables. Thus, the concept of conditional independence between triplets of variables is the foundation for understanding and interpreting the Bayesian network framework. Subsequently, the structure constitutes the *qualitative* part of the model. (2) The unconditional probabilities for all *root nodes*—that is, nodes with no predecessors—must be specified as well as the conditional probabilities for all other nodes, given all possible combinations of each node's direct predecessors. These unconditional and conditional probabilities constitute the *quantitative* part of the model.

Once the Bayesian network is built, it constitutes an efficient framework for performing probabilistic inference (Lauritzen & Spiegelhalter, 1988). It allows us to assess a probability distribution over some variables of interest, given evidence of the values of some other variables in the net. Nevertheless, the initial problem of building the Bayesian network remains. The structure and conditional probabilities necessary for characterizing the network can be provided either externally, by experts—a process that is both time consuming and subject to mistakes—or by automatic learning from a database of cases. However, the learning task can be separated into two subtasks: structure learning—that is, identifying the topology of the Bayesian network—and parametric learning. The second subtask is related to the estimation of the numerical variables (conditional probabilities) for a given Bayesian network topology.

It is common to classify the different approaches to Bayesian network model induction according to the nature of the modeling. These approaches are based either on the detection of conditional (in)dependencies between triplets of variables or on score + search methods.

In the first of these approaches, the output of the algorithms is a directed, acyclic graph that represents a large percentage—or all, if possible—of these relations. Once the structure has been learned, the conditional probability distributions required to completely specify the model are estimated from the database. See Spirtes, Glymour, and Scheines (1993) for more details about this approach to Bayesian network modeling from data.

Although an approach to model building by detecting conditional (in)dependencies is quite appealing, because of its closeness to the foundation of Bayesian networks, a large percentage of the structure-learning algorithms developed belong to the category of score + search methods. To use this approach to learning, we need to define a *score* metric that measures the goodness of fit of every candidate Bayesian network with respect to a data file of different cases. In addition, we need a *search* procedure that will allow us to move in an intelligent way through the space of possible directed, acyclic graphs. The most usual score metrics are *penalized maximum likelihood*, a Bayesian score known as *marginal likelihood*, and scores based on *information theory*. With respect to the search procedure, many different alternatives exist in the literature: *greedy search*, *simulated annealing*, *genetic algorithms*, *taboo search*, and so on. For a review of score + search methods for learning Bayesian networks from data, see Heckerman, Geiger, and Chickering (1995).

3. CONCLUSIONS ABOUT HUMAN ASSESSMENT FROM INTERVIEWS WITH EXPERTS

Both the readability and composition literatures agree that reading ability, prior knowledge, interest, and motivation have influence over text comprehension and composition. Hence, we have chosen a sample that includes participants from those groups that have specific difficulties and achievements in summary evaluation. As part of the experiment, experts working in the target contexts were interviewed. The goal was to identify the most relevant issues in each learner group and to observe the ways these issues are assessed in practice. In a preliminary approach, primary and secondary education was chosen as representative of immature writing skill, and L2 learners were chosen as an immature group that tends to show lack of language ability and comprehension failures (see Section 1). Finally, the university context was chosen as representative of mature skills. Different requirements were reported by experts on the said groups.

3.1. Primary and Secondary Education

These learners use *scaffolding*, or a stepwise methodology, to acquire summarization strategies. Our experts here follow a teaching strategy similar to the one defined by Cassany (1993), Fitzgerald (1987), and Inoue (2005). The main goal of this strategy is to learn how to abstract main concepts from text and, at the same time, to produce the required language to create a good summary. Evaluation is performed as one of the steps in the process, together with summarization instruction. Therefore, evaluation is gradual, and the evaluation mode would be considered *assessment*, as defined by Inoue.

During the summary production process itself, learners use several tools that support summarization. Some of the important tools are concept maps and schema that allow selection and organization of main ideas from text. The use of these tools is considered good training in learning to identify relevant ideas, which has been found to be a prototypical difficulty in immature summarization (Manelis & Yekovich, 1984; Taylor, 1982). In addition, these tools are aided by theoretical materials on connectors; by feedback

to prevent reiteration; by input on avoiding cut-and-paste strategies utilizing actual pieces from the text (Brown & Day, 1983); and by skills at using coherence, cohesion, adequacy, grammar, and so on. Teaching is mainly instructive, although collaborative learning, peer evaluation, and self-evaluation are also integrated in the learning process. Assessment is produced stepwise, with teachers defining the evaluation criteria at the beginning and learners working toward these evaluation goals by trial and error. In the end, learners produce a portfolio about the whole process. This portfolio contains all the steps learners have gone through during the summarization–instruction process, such as concept maps, schema, use of connectors, use of prepositions, and so on. In this way, primary education students learn text comprehension strategies, main-idea identification, use of connectors, text transformation, and other skills. Generally speaking, they gain competence with using discourse and abstraction.

A few more considerations also need to be borne in mind. For instance, different text types derive from diverse writing requirements. In general, young summarizers find narrative text easier to summarize than explicative text. According to experts, the reason behind this might be that in narrative text, the information normally follows a sequential order, whereas explicative text will not necessarily follow this pattern. Our experts stated that because of their age, young summarizers have mainly been exposed to storytelling and reading. This idea is supported by Taylor (1982), who stated that while learning to summarize, young students tend to be more familiar with narrative text. As a result, habituation and the general lack of a chronological structure in expository text make it difficult for them to summarize (Garner, 1987).

Moreover, secondary education lecturers reported that students had difficulties when they were expected to change temporal or person references in text production. In addition, prior familiarity with the content can determine text comprehension and main-idea identification (Symons & Pressley, 1993). For these reasons, secondary-level teachers and lecturers tend to pay special attention to reading-text selection, trying to find those that better match learners' previous knowledge and background.

3.2. L2 Learning

L2 summarizers are often mature summarizers in their first language but lack ability with the L2 (Kozminsky & Graetz, 1986; Long & Harding-Esch, 1978). Therefore, their problems are different, and their learning and evaluation strategies vary from those of the previous learners.

When evaluating, the L2 teachers we interviewed affirmed that they first look at main-idea identification, and then at language competence. They distinguish the relevance of these variables according to learners' abilities and language levels.

Moreover, the support they provide learners is based on the use of dictionaries, theory of grammar, and in some cases concept maps. Whether or not they use concept maps depends on an L2 learner's personal criteria; whereas some L2 students tend to find such support helpful, others prefer to rely on their working memory capacity. Furthermore, the use of aid tools varies depending on L2 ability, so that the more proficient learners are, the closer their needs are to those of native mature summarizers. In short, those with lower levels of L2 ability focus mainly on grammar, whereas those with higher levels focus more on comprehension and style.

However, the L2 group's literacy level tends to be more heterogeneous than that of the primary and secondary group. Their summarization abilities depend on the one hand on their previous literacy, and on the other on their proficiency in the L2. However, the L2-learning group also requires specific training that does not necessarily match the needs of those with greater proficiency in the language.

3.3. University

According to our experts, the university group does not obtain any specific instructive training in summarization at all. Aid tools are used by learners according to their own criteria. Their work is graded for summarization as well as for other abilities, but there is no formal training on summarization. It is assumed that these students have proficient language abilities and are mature summarizers.

Thus, the three groups we have described show different contextual needs when producing summaries. A summary of the aid tools used by each of the three groups is shown in Table 1.

Relevant variables identified thus far in a summarization environment are *text related* (text type, text present or absent, theme, and text length), *aid tools* (dictionaries, spelling and grammar check, theory on summarization strategies, concept maps and schema, etc.), *summary related* (adequacy, coherence, cohesion, use of language, and comprehension), or *learner related* (learner level, learner's prior knowledge, etc.).

4. HUMAN SUMMARY-EVALUATION PERFORMANCE

Using the information gathered from the rater interviews described in the previous section, we designed a summary-evaluation experiment. We concentrated on modeling a single reading text in order to observe the evaluation decision making performance of experts with different backgrounds. Hence, the goal was to observe underlying evaluation patterns using a Bayesian graphical interface (Elvira Consortium, 2002).

4.1. Method

4.1.1. Participants. Most researchers agree on two primary target groups of interest: mature and immature summarizers (Brown & Day, 1983; Garner, 1982, 1987; Taylor, 1982). In addition, evidence supports the idea that the L2 group has specific characteristics (Kozminsky & Graetz, 1986; Long & Harding-Esch, 1978). The goal when choosing a varied sample of raters was to gain a fairly wide representation of expertise, to cover the observed different target views in evaluation. The premise was that, if a common criterion was found among varied target disciplines or contexts, evidence on summary-evaluation agreement would be

Table 1
Aid Tools in Summarization

Aid Tool	Evaluation Context		
	Primary & Secondary Education	L2 Learning	University
Dictionaries	Sometimes; students often ask about unknown terms	Often; students also ask about unknown terms	Often
Concept maps & schema	Often used as part of the training	Depends on the user's criteria	Depends on the user's criteria on main-idea identification
Theory of grammar	Rarely; students tend to ask when confused	Often, in low levels	When language rules change
Summarization theory	Teachers offer handouts and lecture notes to support learners	Sometimes; handouts and lecture notes	None
Sample summaries	Good reading; sometimes students practice and observe good and bad summarization strategies	Sometimes offered as a baseline	None

stronger, and that such agreement would be even more likely to be maintained within the same discipline.

Thus, we chose a sample of 15 participants, all of them experts on summary grading. In this sample, 5 were secondary school teachers, 5 were L2 teachers, and the other 5 were university lecturers. Apart from the university lecturers, the raters taught and evaluated summarization skills on a regular basis and had worked in different educational contexts for more than a decade. The participants did not have contact with each other.

4.1.2. Materials, Procedure, and Analysis. We maintained the same diversity of population in both summary collection (see the details in this section) and participant selection. The experiment included a booklet containing some experimental instruction, the reading text, five summaries, evaluation templates, and a definition of the evaluation variables. The data were analyzed using the Bayesian approach described at the end of this section.

Experimental instructions. This text contained the basic guidelines for the task. After being informed of confidentiality issues and the purpose of the experiment, the participants evaluated five summaries and responded to a questionnaire about their criteria and the methods they followed.

Reading text. All of the summaries were written on the basis of the same reading text. It was 1,203 words in length and concerned the influence of doping in cycling races. The summarizers were simply asked to write a summary of the given text.

Five summaries. A large sample of summaries was gathered from primary and secondary, L2, and university students. Five of these summaries were selected for experimental purposes.

1. The first summary (S1) was written by a first-course secondary education student who wrote the summary by copying several sentences from the reading text. In other words, it was produced using the cut-and-paste strategy described in Section 3.1. It contained 131 words.

2. The second summary (S2) was written by a mature high-intermediate L2 student of Basque and had 145 words.

3. The third summary (S3) was created by another first-course secondary education student and was 107 words long.

4. The fourth summary (S4) contained 222 words and was produced by a university student.

5. Finally, the fifth summary (S5) was written by another mature high-intermediate L2 student of Basque and was 167 words long.

The raters did not have any information on the summary writers' backgrounds and identities. All of the summaries were typed in the booklet in order to hide this information. The typed summary contents were identical to those of the originals, including any orthography and grammar errors. The only difference from the originals was that the booklet summaries were typewritten in order to avoid any possible inferences on ability based on handwriting.

Evaluation templates. The quantitative data were obtained by rating summaries on a 0–10 scale, producing a global overall score and partial scores in cohesion, coherence, language, adequacy, and comprehension. These criteria are explained in Section 3. Moreover, the raters were requested to write comments on the text and/or to give any further information that they thought relevant. Hence, these qualitative data were included to detect information that was not acquired numerically. Both the quantitative and qualitative information were obtained using a template that was common to all of the summaries. The raters were expected to fill in numeric evaluations, and space was left for free evaluation of the summary.

Definition of each of the rating variables. All the raters had access to definitions of the evaluation variables on the last page of the experimental booklet. The partial-rating variables, primarily identified in primary and secondary education, were chosen as experimental variables because they represented all of the rating possibilities identified in the study. In other words, all of the reported possibilities were contained in these criteria. These variables also included ratings proposed by other authors, mainly from detailed studies focusing on content coverage (Garner, 1982; Long & Harding-Esch, 1978; Winograd, 1984), content coverage and coherence (Kozminsky & Graetz, 1986; Taylor, 1982), and finally content coverage, adequacy, and coherence (Manelis & Yekovich, 1984).

Probabilistic analysis. As has been stated, the goal was to predict the global scores of summaries by using the partial-grading and context variables identified in the previous subsection.

The variables for this experiment were the *rating variables*—adequacy, coherence, use of language, cohesion, comprehension, and global score—and the additional *context variables*—rater, summary (summary typology in certain rating range), and summary origin (the background of the summarizer). Other variables were kept constant here: text presence or absence (in this experiment, summaries were produced in text-present mode), text type (which refers to the type of text, in this case explicative), text theme (cycling), text length (1,203 words), and aid tools (no aid tools were used while the five experimental summaries were produced).

So, how do these partial variables relate to global score? The fields of statistics and machine learning have developed different approaches to solve this supervised classification problem: classification trees (Breiman, Friedman, Olshen, & Stone, 1984), classifier systems (Holland, 1975), discriminant analysis (Fisher, 1936), K-NN classifiers (Cover & Hart, 1967), logistic regression (Hosmer & Lemeshow, 1989), neural networks (McCulloch & Pitts, 1943),

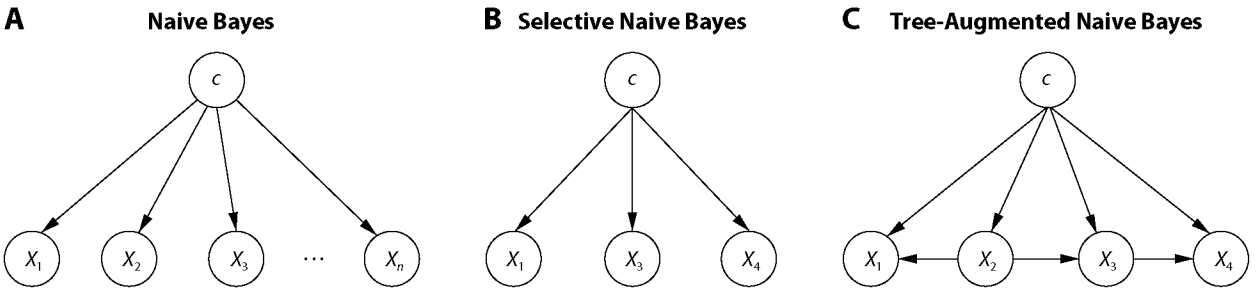

Figure 2. Three Bayesian classifier structures.

rule induction (Clark & Niblett, 1989), and support vector machines (Cristianini & Shawe-Taylor, 2000), among others. Within these approaches, Bayesian networks are models that require some of the least effort in order to interpret the results. They have a straight graphical representation, allowing observation and understanding of the underlying probabilistic classification process. Moreover, it is possible to classify even when the values of all of the variables are not known (e.g., missing values). Considering these factors, we chose Bayesian classifiers as the paradigm for carrying out our supervised classification task.

Assuming a 0/1 loss function, a Bayesian classifier $\gamma(x)$ assigns the most a posteriori probable class to a given instance—that is,

$$\gamma(x) = \arg\max_{c} p(c \mid x_1, \ldots, x_n),$$

where c represents the class variable to be predicted and $x_1, \ldots, x_n$ are the values of the predictor variables.

Using the Bayes formula (Bayes, 1764), we can express the a posteriori probability of the class in this way:

$$p(c \mid x_1, \ldots, x_n) \propto p(c) \cdot p(x_1, \ldots, x_n \mid c).$$

Now, assuming different factorizations for $p(x_1, \ldots, x_n \mid c)$, we can obtain a model hierarchy of increasing complexity between Bayesian classifiers. In this article, we consider three paradigms for this hierarchy: naive Bayes, selective naive Bayes, and tree-augmented naive Bayes.

Naive Bayes (Minsky, 1961) is a Bayesian supervised classification algorithm built from the assumption of conditional independence of the predictive variables, given the class. From this assumption, we have

$$p(x_1, \ldots, x_n \mid c) = \prod_{i=1}^{n} p(x_i \mid c),$$

and the naive Bayes classifier uses the following formula:

$$\gamma_{nB}(x) = \arg\max_{c} p(c) \prod_{i=1}^{n} p(x_i \mid c).$$

See Figure 2A for a graphical representation of a naive Bayes structure.

The introduction of all of the predictive variables into a model can degrade the predictive accuracy of the naive Bayes classifier. In fact, the naive Bayes paradigm is robust with respect to irrelevant variables, but very sensitive to redundant or correlated variables. Therefore, a variable-selection process is required. This combination of feature subset selection and naive Bayes is known as a *selective naive Bayes* paradigm (Langley & Sage, 1994). It is similar to naive Bayes, but in this case not all of the predictive variables are used by the classifier. For its construction, a greedy search process is performed, looking for the subset of variables that maximizes the model classification power. Figure 2B presents a selective naive Bayes structure.

Naive Bayes and selective naive Bayes are both unable, however, to deal with dependencies between the predictive variables. In domains in which the conditional independence between predictor variables (given the class variable) is violated, the performance

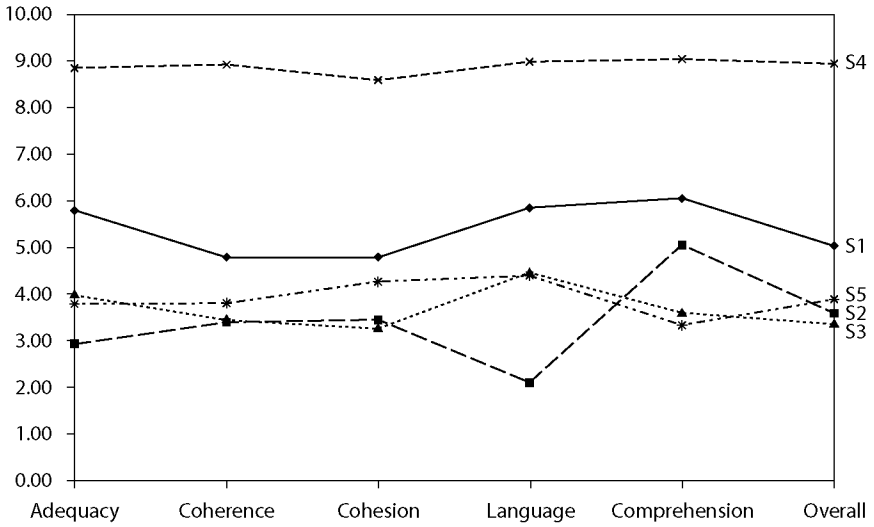

Figure 3. Graph of partial and overall scores for each summary.

of both paradigms can be very limited. One additional paradigm, the tree-augmented naive Bayes (TAN; Friedman, Geiger, & Goldszmidt, 1997), builds a classifier in which a probabilistic tree-like structure, built among the predictive variables, is extended with a naive Bayes structure. The method proposed by Friedman et al. is based on the computation of conditional mutual information between pairs of variables,

$$I\left(X_i, X_j \mid C\right) = \sum_{c}\sum_{x_i}\sum_{x_j} p\left(x_i, x_j, c\right) \log \frac{p\left(x_i, x_j \mid c\right)}{p\left(x_i \mid c\right) \cdot p\left(x_j \mid c\right)},$$

and forces us to construct a connected tree structure including all of the variables in the problem domain. However, as is explained in Section 4.2, our approach does not take all of the variables into account when building the classification model. See Figure 2C for a TAN model.

We utilize an extension of the TAN model, known as the *TAN filter approach*. The original TAN algorithm requires construction of a connected tree structure with all of the variables in the problem domain. Our proposed filter approach (Blanco, Inza, Merino, Quiroga, & Larrañaga, 2005), however, allows us to not necessarily take all of the variables into account when building the classification model; instead, we use the subset of domain variables that overcome the $\chi^2_{(ri-1)(r0-1);1-\alpha}$ test (Whittaker, 1990) to perform the classic TAN algorithm.

4.2. Results and Discussion

Our descriptive results show that S4 was rated highest overall, and S2 lowest. S2, S3, and S5 showed very similar overall grades, and S1 was rated highest among the nonmature summarizers. A graphic representation of the overall and partial-score means can be seen in Figure 3. The lowest scores in language are assigned to the two L2 student summaries (S2 and S5). S2 had the lowest partial score in language but a relatively much higher one in comprehension.

Overall, S1 had a mean score of 5.4, S2 of 3.4, S3 of 3.7, S4 of 8.9, and S5 of 3.9. The lowest mean score was produced by S2, with the lowest data point on the curve in language and the highest in comprehension. This result was followed by S3, with a low point in cohesion and a high in language. Then, S5 had its peaks in cohesion and language and a low point in comprehension. Next came S1, with its peaks in comprehension and language and low points in coherence and cohesion. Finally, the highest-scored summary was S4, with a very homogeneous evaluation: high scores in adequacy, coherence, language, comprehension, and overall score, and only a slightly lower score in cohesion.

From the supervised point of view, the first aspects to take into account are the type of the class variable and the values that it takes. This variable distinguishes between the different classes found in a problem. In the present study, global score is our class variable. The second aspect to consider is that Bayesian networks deal only with discrete (categorical) data; six of the variables in our problem, including the global score, are continuous. Hence, a categorization process is necessary. This discretization process is supported by real educational context practice, in which numeric scores are categorized in standard marks.

Since we are dealing with academic results, the chosen cutoff approach was to split them into three categories—

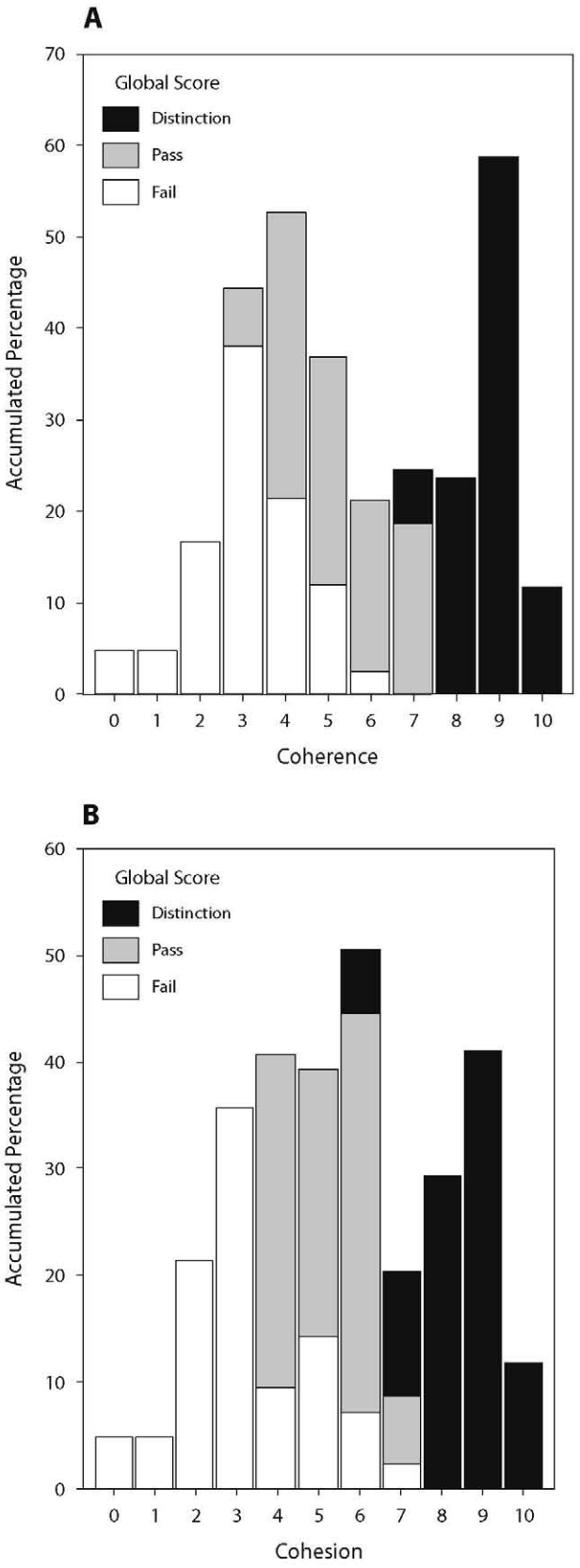

Figure 4. Histograms of levels of the global score variable as a function of the coherence (A) and cohesion (B) variables.

distinction, *pass*, and *fail*. In this way, we can distinguish maximum, medium, and minimum performance levels.

Once the three main categories were defined, the next step was to categorize the rest of the continuous variables. Bear in mind that the remaining variables had constant values in this study, and so could not be included. See Catlett (1991); Dougherty, Kohavi, and Sahami (1995); and Kerber (1992) for classical discrimination methods for continuous variables. All of these studies tried to fit the continuous plot of the function with the areas of disjoint intervals. This process has one primary inconvenience: It does not take into account any semantic information about the original variable.

Because all of our continuous quantitative variables are scores between 0 and 10, we need to perform a more sophisticated discretization policy. One of the most extended discretization procedures in the machine learning domain is the *entropy discretization* (Fayyad & Irani, 1993). This procedure uses the information provided by the class variable of the supervised problem to identify the best adequate disjoint discretization bins. On the basis of the variable marginal distributions given the class, a search process is performed to look for the points at which the marginal continuous functions change their behavior. This process could be done automatically, using a measure of entropy, or by hand, for each variable individually. In this work, considering that only five variables must be discretized, we chose to use a manual procedure by means of conditional histogram plots.

For each continuous variable, histograms divided by the three possible states of global score have been generated—see Figure 4, in which the plots for coherence and cohesion are displayed. We observed the points at which the values' behavior switched between classes; as Figure 4 shows, three behavior changes are easily detected for both variables. On the basis of those trend-changing points, we could set the intervals for each individual categorization. Note that, similar to the examples in Figures 4A and 4B, three disjoint intervals were also detected for all the other continuous variables.

The intervals used in the categorization process of each continuous variable are shown in Table 2. The first row shows the marks for the global score category. The rows below then list the disjoint intervals detected for the rest of the continuous variables, on the basis of the three categories defined for global score.

4.2.1. Learning selective Bayesian classifiers. Once the data are ready for statistical treatment, the next step is to determine which of the available Bayesian classifiers is the most suitable for the current problem. For this purpose, we have used the Elvira framework (Elvira Consortium, 2002). Elvira is a software platform built to deal with graphical probabilistic models that allows a user to produce Bayesian models from raw data.

As we mentioned in Section 2.2, there are two main issues when learning Bayesian networks: the structure of the graph and induction of the values associated with each of the probabilistic variables. In the structures presented in Section 4.1.2, we described four different Bayesian models for the data. The final goal was to observe the one that most accurately reflects human behavior.

In order to analyze in depth the underlying process reflected by these data, it is necessary to select a model to work with. The chosen model must be the one that shows the best accuracy in predicting the class variable. A robust cross-validation analysis was performed considering sample size in order to better estimate the accuracy of each paradigm at predicting the class variable. Then, to determine a model's accuracy, a process of *leaving-one-out* (loo) validation (Stone, 1974) was performed. A loo validation consists of learning a classification model with $N-1$ cases from the original data—where N is the number of instances for the problem—in order to discover the class of the excluded Nth case. The excluded case is different in each loo iteration, and the model is learned using the rest of the data, which always include $N-1$ cases. The process is repeated N times. In particular, for this work there were a total of 75 instances. Thus, for each classification model, 75 intermediate classifiers were learned and tested. Across the four different classification models, a total of 300 validation models were learned. Finally, an average and a standard deviation of the predictions were calculated in order to establish each model's accuracy.

The accuracy measures are shown in Table 3. Of the four models, the naive Bayes and selective TAN obtained the best accuracy. The election between these two models was therefore made according to the semantic representations that each model could handle. As shown in the previous section, selective TAN is able to detect dependencies among the variables, an ability the naive Bayes paradigm does not have. Since we were mostly interested in variable relationships, the selective model was chosen as the model for the knowledge inference process.

4.2.2. (In)Dependencies found by the selective TAN model. One of the most critical analyses in a discussion of the results obtained from a model is the presence or

Table 2
Categorization of the Quantitative Variables

	Fail	Pass	Distinction
Global score	[0, 5)	[5, 8)	[8, 10]
	1st Interval	2nd Interval	3rd Interval
Adequacy	[0, 5)	[5, 8)	[8, 10]
Coherence	[0, 4)	[4, 8)	[8, 10]
Cohesion	[0, 4)	[4, 7)	[7, 10]
Language	[0, 6)	[6, 8)	[8, 10]
Comprehension	[0, 6)	[6, 9)	[9, 10]

Table 3
Means and Standard Deviations of Leaving-One-Out Validations of the Different Bayesian Models

	Accuracy (%)	
Model	*M*	*SD*
Naive Bayes	86.67	3.95
Selective naive Bayes	80.00	4.65
Tree-augmented naive Bayes	82.67	4.40
Selective tree-augmented naive Bayes	86.67	3.95

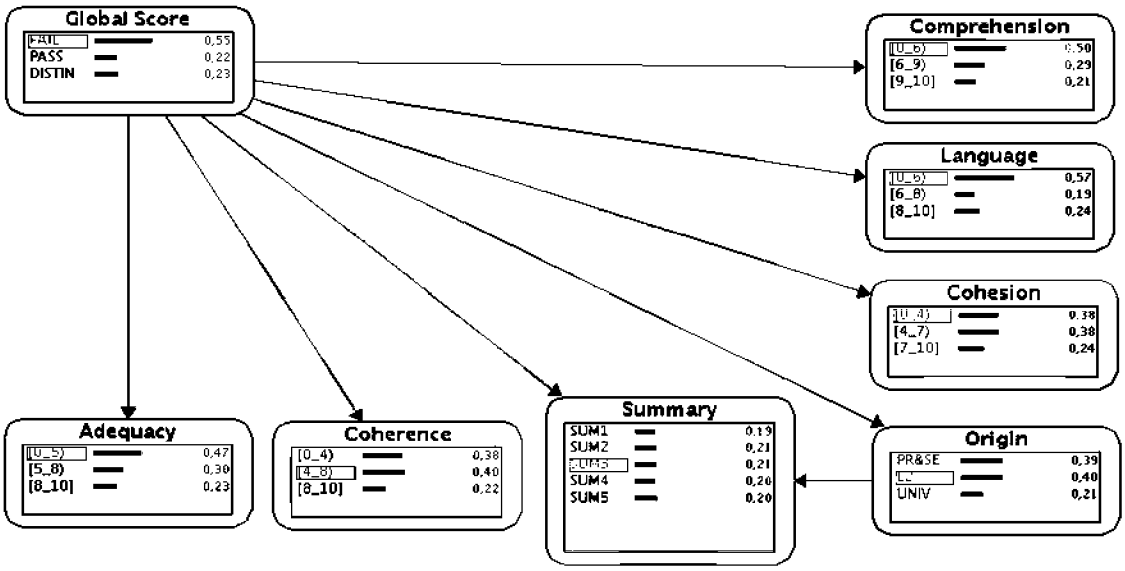

Figure 5. Selective tree-augmented naive Bayes knowledge inference model with the a priori distribution for each variable.

absence of dependencies between the variables involved in the problem. In this case, selective TAN correctly predicted 86.67% of the categorical marks. The structure of this model, with the entire marginal probability distribution for the variables selected by the learning process, is shown in Figure 5.

As a selective paradigm model, selective TAN tries to distinguish between relevant and irrelevant features of a problem and to include only the relevant ones. One of the studied variables, rater, is not present in the model, which means that this variable did not provide significant information to the domain. This relevance result was verified using the nonpaired k-independent-samples Kruskal–Wallis (1952) hypothesis test. The null hypothesis to test was that all the rest of the quantitative variables were independent of the rater variable. The obtained p values were .254 for adequacy, .155 for coherence, .433 for cohesion, .239 for language, .522 for comprehension, and .288 for global score. These p values clearly show that the tested hypothesis could not be rejected. Bear in mind that this hypothesis contrast was performed using the original continuous values of the quantitative variables, and that all the p values make sense of the categorization and modeling approach. Hence, the professional

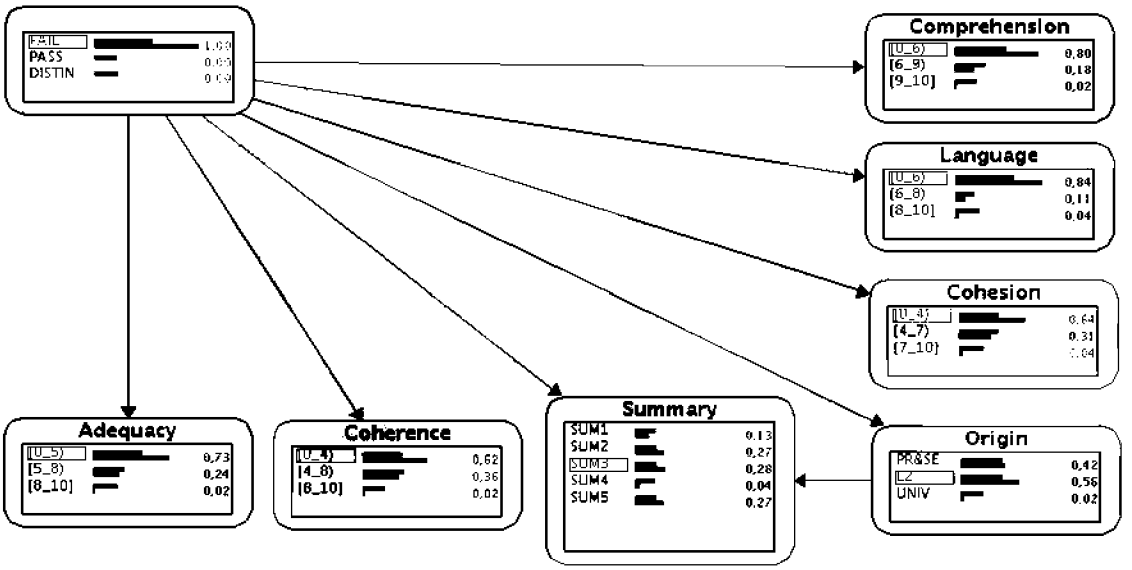

Figure 6. Propagation of the *fail* mark over the rest of the variables.

background of the raters was not relevant when taking evaluation decisions in this model. Therefore, whether raters came from primary and secondary education, L2 teaching, or a university, they all had similar evaluation criteria.

The next consequences derived from this modeling experiment are the independence of the predicting variables when global score is known and the conditional dependency between origin and summary (see Figure 5).

As a tree-augmented paradigm, selective TAN tries to set up conditional dependencies between variables. For this purpose, it uses statistical tests. The fact that only one dependency was fixed shows that no other combination of variables was able to pass the test threshold. Bear in mind that the simple naive Bayes obtained the same prediction accuracy, and this model assumes conditional independence for each variable on the model. Thus, we can clearly state that adequacy, coherence, cohesion, language, origin, and comprehension were all conditional independent, given a global score.

In other words, on the basis of conditional independence, a sole evaluation of each independent variable is possible, given the class. An important consequence of this fact is the possibility of studying the changes between each of the predictive variables and global score in an independent way. Moreover, it is even possible to identify which of the variables were most relevant when deciding the final score provided to a learner summary.

4.2.3. Probabilistic distributions on the propagation when deciding learner evaluation. On the basis of the probability variables learned by the classifier from human rating behavior, it is possible to determine changes in the a posteriori probabilities for each variable as evidence is found. Once evidence is selected in the graph for one or more variables, it is possible to propagate that fact to the rest of the nodes of the Bayesian network (Lauritzen & Spiegelhalter, 1988). As an example, three different pieces of evidence can be selected in order to observe the changes in the rest of the variables considering this particular sample. In Figure 6, a *fail* mark is fixed for the global score variable, and the model shows in red the probabilities that change. In other words, joint probabilities lead to this grade for our group of participants dealing with this particular text.

Hence, in the example in Figure 6, the probability of achieving a language score between 0 and 5 rises to .84. At the same time, the probability of obtaining a score between 0 and 5 in comprehension goes up to .80. The rest of the variables do not show such significance for producing a *fail* mark in global score. Therefore, the high probabilities for comprehension and language we obtained illustrate the fact that—according to our participants' criteria—the borderline between a *pass* and a *fail* score is related to degree of understanding and use of language, which is consistent with previous work in L2 learning (Kozminsky & Graetz, 1986; Long & Harding-Esch, 1978) and immature summarization (Garner, 1982; Taylor, 1982).

The second example illustrates the changes produced by fixing the global score on a *pass* mark. In this case—see Figure 7—the two most significant changes are in coherence and cohesion. The probability of achieving a score in the intervals [4, 8), and [4, 7), respectively, rises to .84 in both cases.

Finally, when fixing a global score at the *distinction* level, three variables show significant changes—see Figure 8. Two of them are common to the evidence for *pass* grades: coherence and cohesion. However, for gaining a *distinction* mark, adequacy predominates. For all of the variables, the probability of achieving the highest rating interval rises to .85 when a *distinction* mark is observed. Thus, in this context, adequacy is the partial-

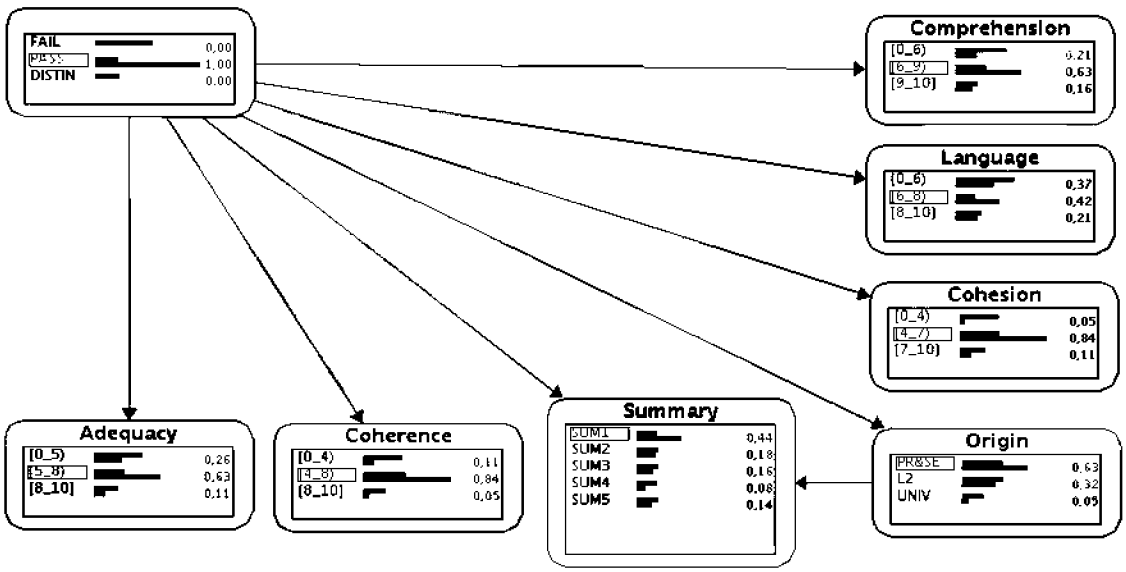

Figure 7. Propagation of the *pass* mark over the rest of the variables.

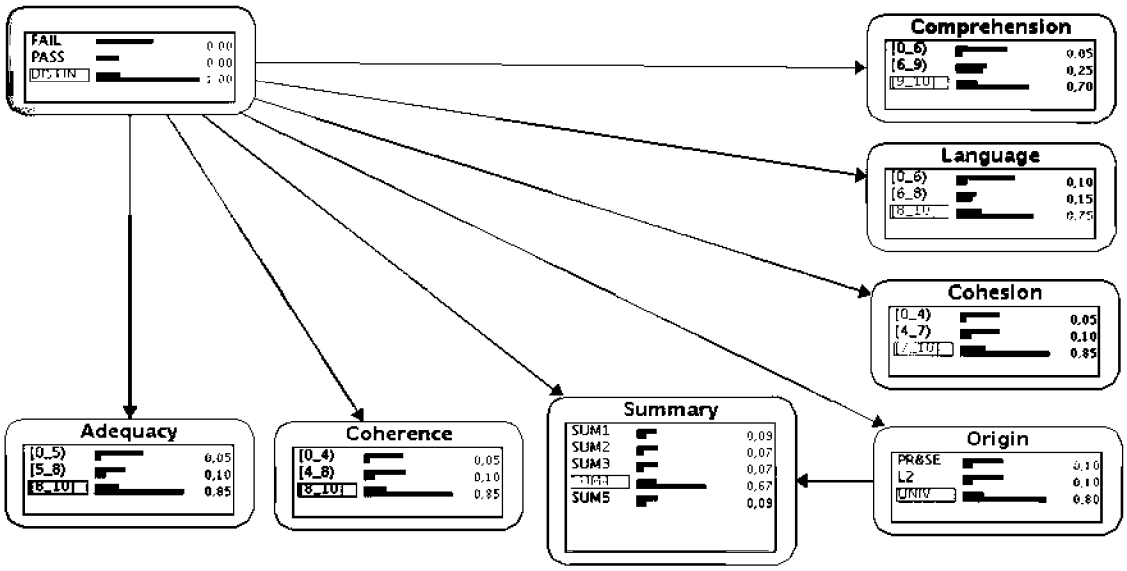

Figure 8. Propagation of the *distinction* mark over the rest of the variables.

rating variable that allows a mark to increase from *pass* to *distinction*, which is consistent with previous work that has argued that expert summarizers make decisions on the basis of the whole text, whereas poor readers and youngsters look mainly to sentences and details (Sherrard, 1989).

This probabilistic model reflects the conjoint criteria shown by our participants and offers a framework for observing the extent to which success or failure on partial scores can lead to success or failure on the global score.

4.2.4. Assessing the most probable explanation in the data relations. Magnani (2004) has claimed that abduction can be used to study cognitive activity in many areas of model-based reasoning. *Abduction*, or *abduction inference*, can be defined as "the process of obtaining the most plausible explanations for a sequence of observed facts." In the abduction process, the result is a hypothesis—that is, a possible explanation for the observed fact—and not a certain conclusion. The logical concept of abduction was first introduced by Aristotle and was borrowed into modern science by Peirce (1955). Therefore, abduction starts from consequences and looks for reasons.

Table 4
Most Probable Configurations of the Modeled Variables for a Given Mark

Variable	Fail	Pass	Distinction
Adequacy	[0, 5)	[5, 8)	[8, 10]
Coherence	[0, 4)	[4, 8)	[8, 10]
Cohesion	[0, 4)	[4, 7)	[7, 10]
Comprehension	[0, 6)	[6, 8)	[8, 10]
Language	[0, 6)	[6, 9)	[9, 10]
Origin	PR&SE	PR&SE	UNIV
Summary no.	S3	S1	S4
Joint probability	.051	.047	.206

Note—PR&SE, primary and secondary education; UNIV, university.

Within probabilistic systems, abduction focuses on a search for the configuration of values of the nonobserved variables that has the maximum probability (Pearl, 1987). The best explanation is the one that maximizes the probability p(configuration | evidence). Abduction in this case stands for the most plausible explanatory diagnostic criteria, in line with the model-based abduction concept defined by Magnani (2004).

In Bayesian networks, *abduction* can be defined as a search for the configuration of states that yields the highest probability for the nonobserved variables. Different names can be found in the literature for the abduction process, including *search for the most probable explanation* (Pearl, 1987) and *search for the maximum a posteriori configuration* (Shimony & Charniak, 1990). In this study, abductive inference is used to determine which were the common criteria in grader performance.

The results for the abduction process in this study are gathered in Table 4. In each column, the most probable joint variable configuration is shown for each value of the global score variable. In other words, each configuration reflects the common evaluation model of the participants. The last row reflects the joint probability distribution obtained for that configuration. The total number of possible combinations of variables in this study is higher than 10.4. Therefore, the joint probabilities obtained via abduction in Table 4 for the gathered configuration are considerably greater than would be expected purely by chance ($2.74 * 10^{-4}$). After considering the studied context and the given configuration, it can be said that in order to gain a *distinction* score in this particular task, the following joint configuration would be required: a score from 8 to 10 in adequacy, a coherence score from 8 to 10, a cohesion score from 7 to 10, a comprehension score from 8 to 10, and a score in use of language of either 9 or 10. The prototypical summary for this category would be S4, and the likeliest

level where this performance could occur would be at a university. Therefore, abduction analysis presents the joint criteria, or the criteria that are common to all graders.

5. CONCLUSIONS

The philosophy behind this study was to observe human evaluation first, and then to move toward a model of summary-evaluation decisions. Therefore, we found it relevant to analyze why we would use each grading measure and how each was related to potential users in an education setting.

As was mentioned in Section 1, one of the issues we wished to address was the common perception in summary assessment of the high subjectivity of evaluation decisions (Sherrard, 1989). Research in collaborative teaching, however, has reported the need to agree on grading criteria (Genesee & Upshur, 1996; Robinson & Schaible, 1995). In addition, Page (2003) defends the assumption that variety among judges allows a closer approximation to the truth behind a grade. In our case, this modeling approach allowed us to analyze how score variation can be attributed to a number of components or facets of measurement. In other words, this approach permits one to analyze quantified results and the relative contributions of a number of sources of variation (Cizek & Page, 2003). To this assumption, this work also added a developmentally relevant variety of contexts and judges. Grounded in this framework, Bayesian networks were then trained to analyze the procedure that raters followed toward each global evaluation grade.

From interviews and analysis of previous work, we concluded that the most complex evaluation criterion shown by our experimental participants contained elements of less complex evaluation criteria. Bearing in mind that the proposed Bayesian model can also estimate probabilities with a smaller number of variables, we chose a paradigm in which the most complex approach included simpler configurations.

In this Bayesian analysis, the irrelevance of the rater variable that we observed in the graphical representation supports the idea that the graders' decisions were based on common criteria (see Section 4.2.2). It shows that, underneath the apparent subjectivity of the decisions, graders from critically different professional contexts displayed common assessment tendencies.

From interviews with experts, it turned out that our raters took decisions on the basis of summarization prerequisites and what they remembered as being the most relevant features of the text. Decisions were based on comparisons of their mental representations of an expected summary text with the features found in each of the evaluated summaries. Thus, it could be assumed that their mental representations, mental summaries, or text macrostructures (Kintsch & van Dijk, 1978) had many features in common. The content of this interrater common view was gathered and represented in the abduction results shown in Table 4. Then, this abductive evaluation was analyzed to determine the specific criteria shown by our participants, or their common grading mental model. Magnani (2001) asserted that a considerable part of thinking is model based, and model-based reasoning acquires relevance by being embedded in abductive processes.

Empirical support for the variable configuration was inferred from the resultant variables' mutual independence. This independence showed that the grading variables we chose on the basis of expert experience and previous research also had empirical entity by themselves. As a result, evaluation of each individual variable on its own was possible. This implied the possibilities that we could identify which variables were most relevant to our participants, how each affected the final score, which differences could be found for certain grades, what the probability configurations were for certain grading levels, and so on. The Bayesian graphical framework allowed us to visually analyze the impact of performance on the global summary score.

Previous research reported that large amounts of irrelevant information in a summary, copying parts of the text, and comprehension failures would indicate a poor summary. This left open the question, however, of what graders expect from a mature summary. Our framework allowed us to visually observe this type of information. Thus, given the studied context and a particular configuration, based on our participants' criteria, the borderline between a *pass* and a *fail* score was related to degree of understanding and use of language, which is consistent with previous work on L2 learning (Kozminsky & Graetz, 1986; Long & Harding-Esch, 1978) and immature summarization (Garner, 1982; Taylor, 1982). To gain a *pass* mark, the two most significant changes were in coherence and cohesion. For a *distinction* score, three variables showed significant changes: Two of them were common to the *pass* mark (coherence and cohesion), but in order to gain a *distinction* mark, adequacy was of prime importance. Thus, adequacy was the partial-rating variable that increased a mark from *pass* to *distinction*. This is consistent with previous research, which has argued that expert summarizers make decisions on the basis of the whole text, whereas poor readers and youngsters look mainly at sentences and details (Sherrard, 1989).

A summary-evaluation framework can have relevant applications within both computer-aided educational environments and traditional educational contexts. The relevance of this capability in learning contexts is supported by the demands for instructional modeling resources, reported in the literature on intelligent tutoring systems (Virvou & Moundridou, 2001), and for mutual collaborative apprenticeship in education (Glazer & Hannafin, 2006). Teachers have reported that they saw their own teaching much more clearly as a consequence of a grading performance analysis, which in turn made them more critical and deliberate in their work (Goldberg & Roswell, 1999). In various learning contexts, the present framework provides a graphical environment for observing the evaluation performance of instructors or graders, beyond simple self-perception.

In education contexts, common grading experience has been observed as a valuable aid to reconsidering teaching practice, making professionals more thoughtful and

more focused on determining the goals of their teaching (Goldberg & Roswell, 1999; Robinson & Schaible, 1995). The present decision-making model can allow one to configure a scenario and analyze how the decisions made will affect global score. In this way, the perception of grading criteria goes beyond mere beliefs to actual performance analysis.

In this article, we have reported how teachers adapt their evaluation focus according to summarizers' learning stages (see Sections 1 and 3). Similarly, our modeling analysis makes it possible to adapt summary-evaluation decisions to the different educational scenarios under a common grading criteria. The Bayesian evaluation model serves as a diagnostic framework for adapting instruction to learner needs. For instance, by selecting the L2 option within the origin variable in the graph, it is possible to observe the most probable summary-evaluation configuration for this specific context. This knowledge can then be used to make evaluation decisions for a variety of criteria, not just on the basis of the provided individual grading variables, but also when selecting only some of the variables or groups of them, depending on the goals involved. For example, it is possible to provide global grade estimations only for comprehension; for comprehension and language; or for comprehension, cohesion, and coherence; the most probable configuration for a *distinction* global score, or for each of the categories studied within a variable, could also be determined.

Moreover, research has found that some graders are unfamiliar with the precise indicators of proficiency (Goldberg & Roswell, 1999). A performance-based approach allows not only for graders to improve their analysis of their own performance, but also for learners to clearly identify grading criteria. In a similar way, learners also gain information from open learner models (Bull & Pain, 1995; Cook & Kay, 1994; Dimitrova, 2003). This framework could be used as an open evaluation model to increase summary graders' awareness of evaluation impact or to show grading criteria to learners. Professional communication is difficult and time consuming (Glazer & Hannafin, 2006). The need for clear evaluative criteria that coincide precisely with description (Goldberg & Roswell, 1999) has been shown in the setting of writing evaluations. Such modeling can help in observing and discussing performance and in agreeing on criteria that could be handled objectively. It can be applied in collaborative instructional settings, as well as in other settings where professionals are required to coordinate criteria with peers. The relevance of showing agreement is high, since any differences that a learner sees between graders can be magnified in learners' eyes (Bandura, 1977; Bower & Hilgard, 1981).

The common abduction model can also be used by learners to acquire awareness of summary-evaluation criteria in greater detail. Showing learners the open evaluation model in full, or just in certain aspects relevant to an instructor's evaluation criteria, can allow them to gain a deeper understanding of the graders' evaluation styles and a better awareness of the instructor's expectations. A common fear of learners in a grading context is the impact that subjectivity might have on their marks. This framework can provide a context in which decisions are influenced only by the previous grading performance of graders included in the model.

The model is also a user-friendly, intuitive utility that does not require any formal training in Bayesian networks or in any other discipline. Little effort is required in terms of the inputting of expert educational data. It also has the advantage of providing a framework that is easily handled; a simple Web application gathers data on summary-evaluation criteria to feed the Bayesian modeling framework.

The aim of this work was, first, to observe how this particular knowledge evaluation mode is dealt with in real educational contexts, and then to use this knowledge to offer tools adapted to common grading practice. This framework could be adapted to automatic grading systems that contain grades for the observed variables. Our following work is focused on producing automatic measures for individual variable grades using latent semantic analysis (Landauer & Dumais, 1997) and natural language processing tools. An example of this work is automatic modeling of cohesion and comprehension (Zipitria, Arruarte, & Elorriaga, 2006). Our future work plans include analyses of further educational contexts and of mathematical modeling.

AUTHOR NOTE

P.L. is now affiliated with the Technical University of Madrid. This work was partially supported by the University of the Basque Country (Grant UE06/19) and the Spanish Ministry of Education and Science (Grant TIN2006-14968-C02-01), as well as by the Gipuzkoa Council in collaboration with the European Union and by the Etortek, Saiotek, and Research Groups 2007–2012 (IT-242-07) programs (Basque Government), TIN2005-03824 and Consolider Ingenio 2010-CSD2007-00018 projects (Spanish Ministry of Education and Science), and COMBIOMED network in computational biomedicine (Carlos III Health Institute). R.A. is supported by Basque Government Grant AE-BFI-05/430. Correspondence relating to this article may be sent to I. Zipitria, Department of Social Psychology and Behavioral Science Methodology, University of the Basque Country, Tolosa etorbidea, 70, 20018 Donostia, Basque Country, Spain (e-mail: iraide.zipitria@ehu.es).

REFERENCES

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370-418.

Blanco, R., Inza, I., Merino, M., Quiroga, J., & Larrañaga, P. (2005). Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics*, **38**, 376-388.

Bower, G. H., & Hilgard, E. R. (1981). *Theories of learning* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Bransford, J. D., Vye, N., Kinzer, C. K., & Risko, V. (1990). Teaching thinking and content knowledge: Toward an integrated approach. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 381-413). Hillsdale, NJ: Erlbaum.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning & Verbal Behavior*, **22**, 1-14.

Bull, S., & Pain, H. (1995, August). *Did I say what I think I said, and do you agree with me? Inspecting and questioning the student model.* Paper presented at the Seventh World Conference on Artificial Intelligence in Education (AACE '95), Washington, DC.

Burstein, J., & Marcu, D. (2003). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 209-229). Mahwah, NJ: Erlbaum.

Cassany, D. (1993). *Reparar la escritura: Didáctica de la corrección de lo escrito*. Barcelona: Editorial Graó.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In Y. Kodratoff (Ed.), *Machine learning—EWSL-91: Proceedings of the European Working Session on Learning* (pp. 164-178). Berlin: Springer.

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23-40). Mahwah, NJ: Erlbaum.

Cizek, G. J., & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 125-145). Mahwah, NJ: Erlbaum.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, **3**, 261-283.

Cook, R., & Kay, J. (1994). The justified user model: A viewable, explained user model. In *Fourth International Conference on User Modeling* (pp. 145-150). Hyannis, MA: Mitre Corp.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21-27.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Dimitrova, V. (2003). STyLE-OLM: Interactive open learner modelling. *International Journal of Artificial Intelligence in Education*, **13**, 35-78.

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 194-202). Tahoe City, CA: Morgan Kaufmann.

Elosúa, M. R., García-Madruga, J. A., Gutiérrez, F., Luque, J. L., & Gárate, M. (2002). Effects of an intervention in active strategies for text comprehension and recall. *Spanish Journal of Psychology*, **5**, 90-101.

Elvira Consortium (2002). Elvira: An environment for creating and using probabilistic graphical models. In J. A. Gámez & A. Salmerón (Eds.), *Proceedings of the First European Workshop on Probabilistic Graphical Models* (pp. 222-230), Cuenca, Spain.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pp. 1022-1027). Tahoe City, CA: Morgan Kaufmann.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research*, **57**, 481-506.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, **29**, 131-163.

Garner, R. (1982). Efficient text summarization: Costs and benefits. *Journal of Educational Research*, **75**, 275-279.

Garner, R. (1987). Strategies for reading and studying expository text. *Educational Psychologist*, **22**, 299-312.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Glazer, E. M., & Hannafin, M. J. (2006). The collaborative apprenticeship model: Situated professional development within school settings. *Teaching & Teacher Education*, **22**, 179-193.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Goldberg, G. L., & Roswell, B. S. (1999). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment*, **6**, 257-290.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, **20**, 197-243.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor: University of Michigan Press.

Hosmer, D. W., Jr., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Inoue, A. B. (2005). Community-based assessment pedagogy. *Assessing Writing*, **9**, 208-238.

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. New York: Springer.

Kerber, R. (1992). ChiMerge: Discretization for numeric attributes. In P. Rosenbloom & P. Szolovits (Eds.), *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 123-128). Menlo Park, CA: AAAI Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363-394.

Kirby, J. R., & Pedwell, D. (1991). Students' approaches to summarisation. *Educational Psychology*, **11**, 297-307.

Kozminsky, E., & Graetz, N. (1986). First vs. second language comprehension: Some evidence from text summarizing. *Journal of Research in Reading*, **9**, 3-21.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, **47**, 583-621.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.

Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence* (pp. 399-406). San Francisco: Morgan Kaufmann.

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B*, **50**, 157-224.

Lehnert, W. G. (1981). Plots units and narrative summarization. *Cognitive Science*, **5**, 293-331.

Long, J., & Harding-Esch, E. (1978). Summary and recall of text in first and second languages: Some factors contributing to performance differences. In D. Gerver & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 273-288). New York: Plenum.

Magnani, L. (2001). *Abduction, reason, and science: Processes of discovery and explanation*. New York: Kluwer/Plenum.

Magnani, L. (2004). Model-based and manipulative abduction in science. *Foundations of Science*, **9**, 219-247.

Manelis, L., & Yekovich, F. R. (1984). Analysis of expository prose and its relation to learning. *Journal of Structural Learning*, **8**, 29-44.

Mani, I., & Maybury, M. T. (1999). *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115-133.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, **49**, 8-30.

Neapolitan, R. E. (2003). *Learning Bayesian networks*. Harlow, U.K.: Prentice Hall.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Erlbaum.

Pearl, J. (1987). Distributed revision of composite beliefs. *Artificial Intelligence*, **33**, 173-215.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Peirce, C. S. (1955). Abduction and induction. In J. Buchler (Ed.), *Philosophical writings of Peirce* (pp. 150-156). New York: Dover.

Robinson, B., & Schaible, R. M. (1995). Collaborative teaching: Reaping the benefits. *College Teaching*, **43**, 57-59.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding:*

Studies in cognitive science (pp. 185-210). New York: Academic Press.

SCHANK, R. C., LEBOWITZ, M., & BIRNBAUM, L. (1980). An integrated understander. *American Journal of Computational Linguistics*, **6**, 13-30.

SHERRARD, C. (1989). Teaching students to summarize: Applying text-linguistics. *System*, **17**, 1-11.

SHIMONY, S. E., & CHARNIAK, E. (1990). A new algorithm for finding MAP assignments to belief networks. In P. P. Bonissone, M. Henrion, L. N. Kanal, & J. F. Lemmer (Eds.), *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence* (pp. 185-196). New York: Elsevier.

SPIRTES, P., GLYMOUR, C., & SCHEINES, R. (1993). *Causation, prediction, and search*. New York: Springer.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, **36**, 111-147.

SYMONS, S., & PRESSLEY, M. (1993). Prior knowledge affects text search success and extraction of information. *Reading Research Quarterly*, **28**, 250-261.

TAYLOR, B. M. (1982). Text structure and children's comprehension and memory for expository material. *Journal of Educational Psychology*, **74**, 323-340.

THORNDYKE, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, **9**, 77-110.

VIRVOU, M., & MOUNDRIDOU, M. (2001). Adding an instructor modelling component to the architecture of ITS authoring tools. *International Journal of Artificial Intelligence in Education*, **12**, 185-211.

WHITTAKER, J. (1990). *Graphical models in applied multivariate statistics*. Chichester, U.K.: Wiley.

WINOGRAD, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, **19**, 404-425.

ZIPITRIA, I., ARRUARTE, A., & ELORRIAGA, J. A. (2006). Observing lemmatization effect in LSA coherence and comprehension grading of learner summaries. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)* (pp. 595-603). Berlin: Springer.

(Manuscript received July 14, 2007;
revision accepted for publication November 24, 2007.)