TEXT ANALYTIC TOOLS FOR THE COGNITVE DIAGNOSIS OF STUDENT WRITINGS

Tjaart Imbos and Ton Ambergen
Maastricht University, The Netherlands
Tjaart.imbos@stat.unimaas.nl

Students can be stimulated to become active learners using a tool for active writing. In our university, we developed such a tool: POLARIS. Active writings of students about statistical concepts are valuable for the students and the teacher. In their writings, students show their understanding of statistical topics. The problem then is how to interpret and score the writings of students in relation to their proficiency in statistics. In this paper text analytic tools are used to cluster and score sample papers of students. Two approaches are compared: a statistics-based approach, Latent Semantic Analysis (LSA) and a linguistic-based approach, known as natural language processing (NLP). The key features of both approaches are discussed, as well as the usability, reliability and validity.

BACKGROUND

Students can be stimulated to become active learners using a tool to support collaborative learning, working on statistical problems and tasks in small groups. In our university we developed such a tool: POLARIS, an acronym for Problem Oriented Leaning and Retrieving Information System. Experiences in using this program in statistics courses have been reported at ICOTS7. Active writings of students about statistical concepts are valuable for the students but also for the teacher. In their writings students show their understanding of statistical topics. The problem then is how to interpret the writings of students in relation to their proficiency of statistics.

The problem can be rephrased as how to make sense of complex data like the writings of students in discussion boards as they are used in our learning tool, POLARIS. Advances in cognitive psychology have extended our understanding of students' learning and broadened the range of performances that can be used to acquire evidence about the developing abilities of the students. Furthermore advanced technology has made it possible to capture students' complex performances in assessment settings.

In this paper, two technological advances are explored as tools for solving the problem to understand and diagnose the knowledge base demonstrated in the students' writings (Ericsson & Smith, 1991; Ericsson, 1980). The first one is called Latent Semantic Analysis (LSA) (Landauer, Foltz & Laham, 1998). LSA is a model that induces representations of meaning of words by analyzing the relations between words and passages in texts. The method used by LSA to capture the essence of semantic information is dimension reduction.

The second tool is called a natural language processing approach (NLP). This is a bottom up approach revealing the concepts and themes contained in a body of documents. Both approaches are applied to a small learning sample of student documents which are compared to each other and to a standard statistical text from a text book.

They can be used as automatic scoring of texts. Based on this scoring of a text a group of students or an individual student could be diagnosed on a certain level of statistical proficiency.
In the paper we will elaborate on these topics and propose a system for ongoing assessment of student writings in an e-learning environment. Because it's work in progress, empirical results are preliminary.

LATENT SEMANTIC ANALYSIS (LSA)

Researchers who use quite a number of variables frequently use data reduction techniques as factor analysis and cluster analysis. In the case of discussion forums in POLARIS, quite a lot of information becomes available in the form of words and sentences related to statistical topics. Here also data reduction is needed. For qualitative data, as students writings, a technique comparable to factor analysis is available: Latent Semantic Analysis (LSA) (Landauer et al., 1998).

Suppose two students write about their understanding of statistical testing. Even if they both have a good understanding of the topic, compared to that of an expert, their writings will still differ. They can be seen as different but also comparable with both having a good understanding of

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

the subject matter. Therefore the objective is how to explain that the first writing is similar to the second, or how two parts of the texts can possibly be compared quantitatively, indicating reasoning processes of the two students. It seems possible to introduce a theory and methodology for writing tasks based on LSA. This theory addresses issues related to the induction, representation and application of knowledge. Actually, LSA infers knowledge from many weak constraints about statistical topics, that are present in the writings of students while they are learning. LSA does not represent a whole *knowledge space* but only the paths students have chosen to find their way in that space. This offers a nice view of how the knowledge space is understood by the students. LSA is a *computational* theory on how students learn to find their way in the knowledge space of statistics and how that space can be described. The features of LSA are: (1) it does not assume independence of writing actions, instead it uses dependencies to infer the structure in the writings; (2) LSA reduces the dimensionality of the space; and (3) there are no *a-priori* assumptions about the knowledge space. LSA is self-organising.

DESCRIPTION OF LSA

LSA is a machine-learning model that induces representations of the meaning of words by analysing the relation between words and passages in large bodies of text. LSA is both a method for educational applications as well as a theory of knowledge representation to model comprehension of statistical topics. The method used by LSA captures the essential information in text passages, while ignoring accidental and inessential word usages. The method selects the most important dimensions from a co-occurrence matrix using *single value decomposition*. As a result LSA can be used to assess semantic similarity between samples of text in an automatic way. That is what we need for the problem described in this paper. Using LSA, student writings can be compared with the expert knowledge base. LSA has been used in applied settings with a high degree of success like essay grading, automatic tutoring and in human language acquisition simulations and in modelling comprehension phenomena (Landauer & Dumas, 1997). More information of how the LSA procedure works in **R** can be found at: http://cran.r-project.org/web/packages/lsa/index.html. In this study we used a LSA procedure of the open source program **R**.

In this procedure the following steps can be discerned: creation of a latent semantic space to allow comparisons between terms and documents; find close terms in a text matrix; essay scoring for grading students en comparing to docent grading; calculate cosine measures to produce a similarity matrix between column (documents, normally) vectors; calculate the dimensionality based on a 'good' number of singular values; fold-in the text matrixes into a latent semantic space; create a vector space with LSA, this space is assumed to reflect the semantic structure in de documents, new documents can be folded-in later in the analysis; print a text matrix; create a query in the format of a given matrix; create a random sample of files; return a summary with statistical information about a given text matrix; create a document-term matrix from text files in a given directory (text pre-processing); bind SPO-triples (subject, predicate, object) to a text matrix; calculate a weighting document-term matrix according to the chosen weighting scheme. A combination of these procedures are used in this paper.

NATURAL LANGUAGE PROCESSING (NLP)

The world has seen an explosion of information the last decade. It is easy to predict that this will only continue the coming decades. Lots of information is stored as text. More than 40 years ago visionary researchers began to seek ways to enrich knowledge in different fields of applications. These applications tried to uncover connections in textual documents by using computer technologies. This gave rise to the birth of what now a days is known as *computational linguistics*. Initially the focus was on categorization and exploration of concepts found in books and other publications. Recently efforts expanded to include ways to *mine* the amount of digitally published information. There is a growing recognition of the importance of analyzing text in various types of scientific research and that this added a significant value to other forms of data analysis. This added value has been recognized by distributors of data analysis computer programs as SAS, and SPSS to mention a few. See for example: http://cran.stat.ucla.edu/web/views/NaturalLanguageProcessing.html.

TEXT ANALYTICS AND HOW IT IS USED IN SPSS' CLEMENTINE

First of all text analytics is not the same as searching texts. The latter is a top down process while the first is bottom up. User's content knowledge of the text is not strictly necessary, but is helpful in building dictionaries for specific fields of application as in our case statistics. Text analytics reveals the concepts and themes in a body of documents and maps the relationship between them. It is a method to extract usable knowledge from unstructured by identifying core concepts and by using the emerging knowledge for decision making purposes. Connections and relationships between core knowledge concepts can be discovered in a large collection of documents. The difference between NLP and LSA is primarily a difference in data processing during the analysis phase. NLP starts with defining important concepts and linguistic links between concepts as defined by grammatical rules. LSA starts with data analysis and reducing the dimensionality of the data matrix of words and documents and ends up with an interpretation of the dimensions and related concepts. SPSS' Clementine uses language based text mining technologies.

LSA in **R** instead is based on data reduction methods, e.g., single value decomposition, a two-mode factor analysis of a given document-term matrix.

Seven steps can be discerned in text analytics processes in **SPSS**: preparing text for analysis; extracting concepts; uncovering relationships through text link analysis; building categories; building text analytics models; merging text analytics models with other data models and using the results to predictive models.

USING LSA IN R AND NLP IN SPSS TO ANALYZE DOCUMENTS ON STATISTICAL TESTING

Learning material and learning task and student documents

First year Health Science students studied an introductory chapter on statistical testing and were stimulated to write about their understanding of important concepts related to this topic. Students worked together in small learning groups of about 10 to 12 students. They were allowed to produce and discuss their writings collaboratively.

In total 10 documents were collected for this study. These documents were completed with the text of the chapter studied by the students. The latter document was added to the collection of 10 student documents. All documents were transformed into text documents and analysed with LSA in **R** and NLP in **SPSS**. Both analyses were exploratory addressing the following global questions: are the two procedures able to discriminate between the students' documents and the text book chapter? Are the two procedures able to discriminate between students' documents with clearly different quality? Are they able to score the student documents; have both procedures equal accuracy? Have both procedures graphical modalities for displaying the data. And finally are there noteworthy differences in data handling and user friendliness?

CONCLUSION

The project reported here is work in progress. Detailed results will be presented at ICOTS8. At this moment we already know that the R LSA procedure clearly discriminates between the book chapter text and the documents of the students. It is less clear whether it is able to discriminate between de student documents themselves.

It's an open question whether SPSS' Clementine can discriminate between the book chapter and de students' documents. Building the dictionary and defining relevant statistics concepts related to the topic of statistical testing is very time consuming, but may be during the final analysis this will pay off.

Both procedures will be discussed, using detailed results that become available during analysis.

REFERENCES

Ericsson, K. A., & Smith, J. (Eds.). (1991). *Towards a general Theory of Expertise* (1st ed.). Cambridge: Cambridge University Press.

Ericsson, K. A. e. S., H. A. (1980). Verbal Reports as Data. *Psychological Review, 87*(3), 215-251.

Landauer, T. K., & Dumas, S. T. (1997). A solution to Plato's problem: The latent semantic theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.