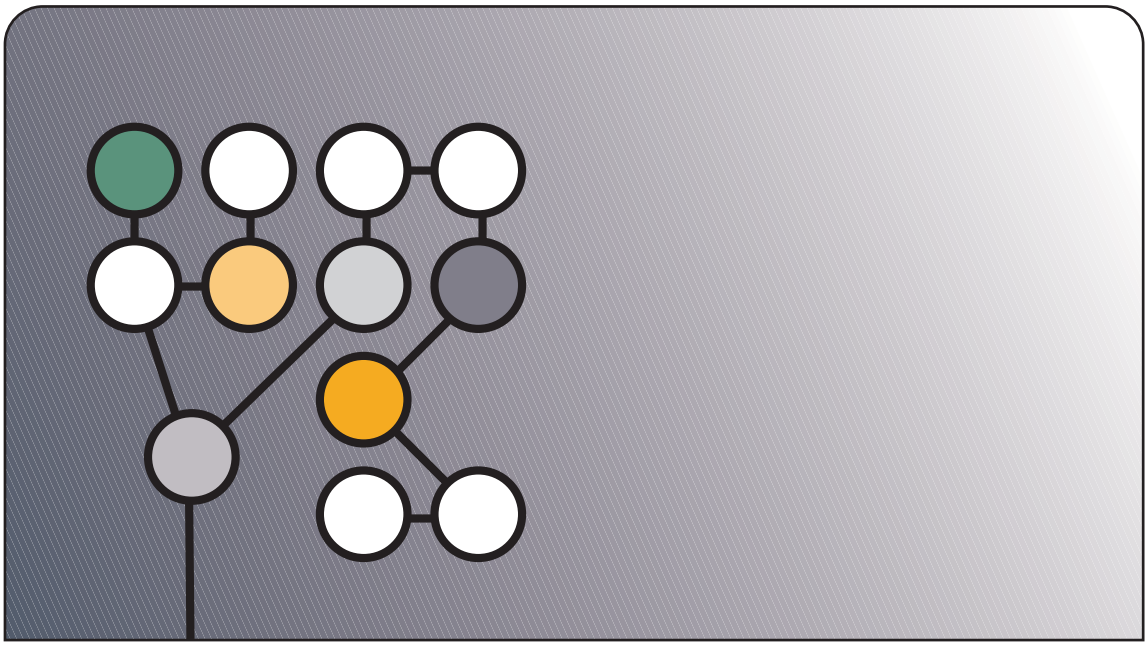

Fridolin Wild, Marco Kalz, Jan van Bruggen, Rob Koper (Eds.)

Mini-Proceedings of the 1st European Workshop on Latent Semantic Analysis in Technology-Enhanced Learning

March 29–30, 2007, Heerlen, NL

iCamp
PROLEARN
cooper
TEN Competence
Building The European Network for Lifelong Competence Development

Fridolin Wild, Marco Kalz, Jan van Bruggen, Rob Koper (Eds.)

Mini-Proceedings of the
1st European Workshop on

Latent Semantic Analysis in Technology-Enhanced Learning

March 29–30, 2007, Heerlen, NL

Scientific Committee

Benoit Lemaire (Laboratoire Leibniz-IMAG, Grenoble, France)

Fridolin Wild (Vienna University of Economics and Business Administration, Vienna, Austria)

Gustaf Neumann (Vienna University of Economics and Business Administration, Vienna, Austria)

Jan van Bruggen (Open University of the Netherlands, Heerlen, The Netherlands)

Michael Berry (University of Tennessee, Knoxville, US)

Rob Koper (Open University of the Netherlands, Heerlen, The Netherlands)

Organising Committee

Bas Giesbers (Open University of the Netherlands, Heerlen, The Netherlands)

Marco Kalz (Open University of the Netherlands, Heerlen, The Netherlands)

Fridolin Wild (Vienna University of Economics and Business Administration, Vienna, Austria)

Index

Question-Answering – Connecting and Supporting the Learner

Peter Van Rosmalen, Peter Sloep, Francis Brouns, Liesbeth Kester, Malik Koné, and Rob Koper, *Open University of the Netherlands*

***Abstract* — Tutors have only limited time to support the learning process. In this paper, we introduce a model that helps answering the questions of students. The model invokes the knowledge and skills of fellow students by bringing them together based on the combination of question posed and their study progress and supports them with text fragments selected from the material studied. We will explain how we used LSA to select and support these peers; examine the calibration of the LSA-parameters and conclude with a small practical simulation to show that the results of our model are fit for use in experiments with students.**

I. Introduction – The Model

THE prototype (see also Table 1) of the model (for a detailed description see [1]) consists of five modules: a Moodle learning environment; a wiki; GTP an LSA implementation [2]; GUP to ease the calibration of LSA (GTP Usability Prototype [3]); and ATL (A Tutor Locator [4]) for the selection of the peers who will assist, based on the topic involved and the users' background and performance.

Pre	A course with a set of topics and users with progress profiles.
Main	1. *Anne* poses a question. 2. The *system* determines: - the most relevant text fragments (LSA); - the appropriate topics (LSA); - the most suitable users (LSA + user profiles). 3. The *system* sets up a wiki with the question, the text fragments and guidelines. 4. The selected *users* receive an invitation to assist. 5. *Anne and the users* discuss and phrase an answer in the wiki. 6. If answered (or after a given period of time) *Anne* closes the discussion and rates the answer.
Post	The answer is stored

Table 1: The main steps of the model.

II. Calibration

The domain of the course we used is 'Internet Basics', a collection of texts, links and tasks that aim to instigate a basic understanding of the Internet [5]. It contains 11 topics, each of which introduces a different aspect of the Internet.. The topics consist of an introduction, exercises, references to external web pages for further study and an assessment. The corpus was extracted manually. It contained the Moodle pages and external web pages; the assessment questions were left out, however. These questions were used to calibrate the model. The documents were used as raw input; this means that no further corrections were applied such as removing irrelevant documents, diacritical signs or misspellings. The final corpus was relatively small. It consisted of 327 documents ranging in size from 50 to 23534 bytes (41 documents smaller than 250 bytes, 50 documents above 3000 bytes). The corpus contained a total of 82986 words divided over 10601 terms, 4440 of which occur in at least two documents.

In addition to the calibration, we investigated if it was possible to define the parameters with a predefined, limited number of steps that can be repeated and automated at a later stage. An overview of applications with LSA [6] revealed that there is no straightforward procedure to determine the LSA parameters. The parameters are influenced by the corpus and the way LSA is applied. We selected the five steps [2], [7] that should be the most important: the definition of a correlation measure and method, corpus preprocessing, normalisation, weighting and dimensionality. We did not carry out, however, an exhaustive test with all possible combinations of parameters. Instead, we started with an initial combination of parameters based on results reported [7]-[8], and in each step, we tested one parameter and continued to the next step using only used the best result(s) (Figure 1).

Correlation measure and method. For our correlation measure, we used cosine similarity. Our method directly follows from our model. First, we used LSA to identify to which topic(s) the question posed fits best. This information is used to identify peers that are competent in the pertinent topic. Second, we wanted to select the three documents that were most suited to assist the peers in answering the question. We combined the two by selecting the three best correlating documents. We used the result of the mapping on the topics to select the parameter combination with which to continue. The questions, 16 in total, were chosen from the original topic assessment questions. Therefore each question should map to one known topic. *Preprocessing* (run 1-3). Because we did not have access to a stemming application for Dutch, we only considered stopping. Moreover, given the size of our corpus, we created our own stop lists based on the term frequency in the corpus [8]. The stop list consisted of the terms that covered 33% (22 terms) and 50% (91 terms) respectively of the overall term frequencies with the exception of corpus specific terms. By way of comparison, we also used a 'general' Dutch stop list (Oracle Text Reference: Release 9.2). For our corpus, this resulted in a reduction of 188 terms. Finally, in each run (until the actual dimensionality step), we

chose to limit the number of singular values to 40% of the sum of the singular values (Wild et al, 2005). *Normalisation* (run 4-5). Next, with a limited number of documents per topic and quite a spread in document lengths we tested the use of normalisation. This has the effect that documents with the same semantic content are ranked equal in the question query. *Weighting* (run 6-8). Subsequently, we applied the three available types of Global Weighting. *Dimensionality* (run 9-10). In the last step, we determined the best value for the dimensionality by comparing the initial value of 40% of the sum of the singular values to 30 and 50%. Finally, in this step (run 11) we did one additional test i.e. we used the 50% stop list in order to check if this would improve our results. The other parameters followed the settings of Run 9. The result was good (15 out of 16) but not an improvement.

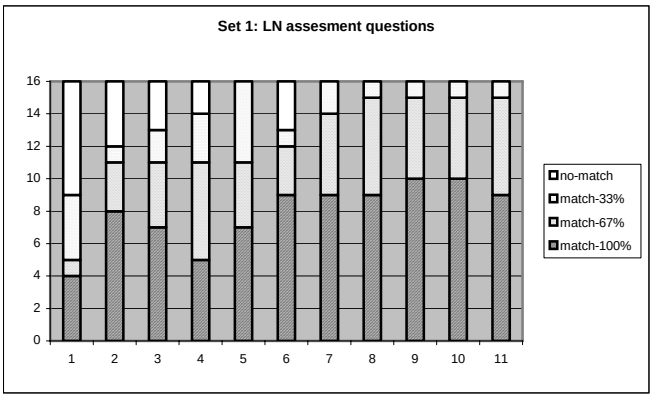

Figure 1: The mapping of the questions on the topics in the calibration runs.

III. A SIMULATION OF THE MODEL

For a final check, we formulated a new set of 16 questions, each connected to one topic. The questions were once again mapped on the topics, and the results were compared with their known topics. The parameter combination of the calibration run 9 and 10 were applied. The model identified the topic correct for 12 out of 16 questions. Case one (the settings of Run 9) did slightly better in the 100% recognition category. For this case we asked two of the designers of the course to rate, on a 5-point scale, the suitability of the text fragments selected through the application of LSA. The suitability of the text fragments is far less accurate; approximately 40% of the questions received one or more fragments rated 3 (5-point scale) or above. The designers of the course, however, indicated that approximately 35% of the questions posed were beyond the scope of the contents of the topic studied; as a consequence the topic did not contain any useful fragments at all. Together the results are promising. The corpus used is rather small, so the chances to find an answer are limited. Also the results may be stepwise improved by making use of successfully resolved questions and their answers. Finally, it is not merely the answer that matters. It is an important aid, but the first concern is to identify the appropriate topic so the right peers can be selected. With 75% recognition we think we are in a good position to achieve this.

IV. CONCLUSION

We introduced a model that intends to help the learner and alleviate the support task of tutors. We described how we calibrated LSA for an existing course. Subsequently, and for the same course, we checked with a simulation whether the model is fit for experimentation with students. In our opinion, the results shown are promising. Moreover, we were able to arrive at our results in a systematic way. The same steps can be followed for a new corpus or if the changes to an exiting corpus are relatively small, the known settings can be reapplied in just one additional run. Obviously, one should be open to retrace one's steps, in particular, if the results are very close (as in the normalisation step) and improvements develop insufficiently.

Clearly, there are still a number of issues to be considered. First, the model has only been applied once and to questions that exactly match one topic. It is fair to expect that, in real practice, part of the questions will cover not just one but more topics. This may complicate the recognition and thus dilute the results. Next, as shown by some of the results, the approach is sensitive to the size (and content) of the available corpus.

ACKNOWLEDGMENT

The authors' efforts were partly funded by the European Commission in TENCompetence (IST-2004-02787).

REFERENCES

[1] Van Rosmalen, P., Sloep, P., Brouns, F., Kester, L., Kone, M. and Koper, R. (2006) Knowledge matchmaking in courses: Alleviating the tutor load by mutually connecting course users. British Journal of Educational Technology, Vol. 37 (6), 881-895.

[2] Giles, J. T., Wo, L., & Berry, M. W. (2001). *GTP (General Text Parser) Software for Text Mining*. Retrieved online January, 2005 at: http://www.cs.utk.edu/~berry/papers02/GTPchap.pdf.

[3] De Jong, A., Brouns, F., Van Rosmalen, P., Sloep, P., Kester, L., & Koper, R. (2006). GUP GTP Usability Prototype. URL: http://sourceforge.net/projects/gup.

[4] Brouwers, M., Brouns, F., Van Rosmalen, P., Sloep, P. B., Kester, L. & Koper, R. (2006). ASA Tutor Locator. URL: http://sourceforge.net/projects/asa-atl.

[5] Janssen, J., Tattersall, C., Waterink, W.,Van den Berg, B., Van Es, R., Bolman, C., & Koper, R. (in press). Self-organising navigational support in lifelong learning: how predecessors can lead the way. *Computers & Education*.

[6] Haley, D. T., Thomas, P., De Roeck, A. & Petre, M. (2005). A research taxonomy for latent semantic analysis-based educational applications. Technical report no. 2005/09. Retrieved April 10, 2006, from http://computing-reports.open.ac.uk/index.php/content/download/187/1136/file/TR2005_09.pdf.

[7] Wild, F., Stahl, C., Stermsek, G. & Neumann, G. (2005) Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings of the 9th International Computer Assisted Assessment (CAA) Conference*, Loughborough, UK, July 5, 2005 (pp. 485–494). Retrieved March 6, 2006, from http://nm.wu-wien.ac.at/research/publications/b497.pdf.

[8] Van Bruggen, J. M., Rusman, E., Giesbers, B. & Koper, R. (submitted). Latent semantic analysis of small-scale corpora for positioning in learning networks. Retrieved February 24, 2006, from http://hdl.handle.net/1820/561.

Using Latent Semantic Analysis to Assess Social Competence

Fridolin Wild and Christina Stahl, *Vienna University of Economics and Business Administration, Austria*

***Abstract* — Assessing social competence is a time consuming and complex endeavor, which has prompted several institutions to implement automated assessment systems. Although other attempts have been made, the majority of automated assessments are still based on multiple-choice formats. Within this contribution, the authors therefore present a new method of how to assess social competence using LSA. The contribution first reports on the research design, then presents the findings which are subsequently discussed in brief.**

I. Introduction

STRUCTURAL changes in the qualification requirements for employees have led many countries to adopt competence-based education schemes. In particular, the promotion of social competencies is considered an important aspect of such education (e.g., [1], [2]). For the purpose of this research, we define 'social competence' as involving abilities that facilitate communicative and cooperative action and that aim at identifying, managing and mastering conflicts.

The increasing adoption of social competence education also poses the challenge of assessment. Assessing learners' social competence is a time-consuming endeavor, which has prompted several institutions to implement automated assessment systems. Although attempts to assess social competence by means of graph-based approaches [3], simulations [4], and natural language processing approaches [5] have been made, the majority of automated assessments is still based on multiple-choice formats.

We therefore present a new method of social competence assessment using Latent Semantic Analysis (LSA). The research design, results and a brief discussion are presented in the following sections.

II. Research Design

For our research we used a corpus of 337 textual contributions produced by students in an online discussion forum in the course of a university seminar. The contributions were split into sentences resulting in 1,012 individual messages. These messages were manually coded along ten dimensions of social competence (politeness, ability to motivate others, phatic communication, ability to express own opinion, cooperation competence, team competence, feedback competence, networking competence, ability to take initiative, and readiness to take on responsibility), assigning the code '1' when evidence of a dimension was found in a message, and'0' when no such evidence was identified. The dimension 'ability to motivate others' was found in only 37 messages and was therefore omitted from further analysis.

The goal of our analysis was to mimic the human coding with LSA. For this purpose the corpus coded along the remaining nine dimensions of social competence was split into a training corpus consisting of 490 messages and a test corpus of 522 messages. The training corpus was used to calculate the LSA space using the share-method proposed in [6] for space reduction. No document pre-processing (e.g., stemming) was performed.

From the 522 test messages, 16 had to be omitted from the LSA analysis as they did not include any terms from the LSA space. Each of the remaining 506 test messages was folded into the LSA space and Pearson's r was calculated to compare the test message with all training messages. For each test message the ten most highly correlated training messages were identified, which – for each dimension separately – were grouped according to their human coding ('1' or '0'). Each dimension of the test document was assigned the code that received the higher weighted sum within this comparison set (see [7] for a similar method).

III. Results

For seven of the nine remaining dimensions of social competence we received good to excellent results. For the dimension 'politeness', LSA was able to correctly predict human coding for 452 test messages (89.33%) out of 506. For the other dimensions, the respective results were 437 (86.36%) correctly coded messages for the dimension 'team competence', 430 (84.98%) correctly coded messages for the dimension 'cooperation competence', 432 (85.38%) correctly coded messages for the dimension 'networking competence', 393 (77.67%) correctly coded messages for the dimension 'ability to take initiative', 392 (77.47%) correctly coded

F. Wild and C. Stahl are with the Institute for Information Systems and New Media of the Vienna University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria (fridolin.wild@wu-wien.ac.at and christina.stahl@wu-wien.ac.at).

messages for the dimension ‘readiness to take on responsibility’, and 320 (63.24%) correctly coded messages for the dimension ‘feedback competence’. The other two dimensions – ‘emphatic communication’ and ‘ability to express own opinion’ – produced inferior results of 229 (45.26%) and 164 (32.41%) correctly coded messages, respectively.

We also compared our results to the plain vector-space model, and surprisingly received similar results. In particular, the vector-space model performed better for three of the nine dimensions, although the results differed only by a maximum of 0.59%.

IV. Discussion

We have been able to successfully apply LSA to assess social competence in contributions to online discussion forums. LSA was able to mimic the human coding process for seven of the nine analyzed dimensions of social competence. Concerning the other two dimensions, we assume that the language used to express these dimensions is too diverse to be captured. Surprisingly we found similar results when comparing LSA to the Vector-Space-Model. Further research and discussion will have to investigate this issue, as well as the applicability of LSA to the assessment of social competence in particular, and the automation of human coding in general.

References

[1] EUROPEAN COMMISSION (2002): The Key Competencies in a Knowledge-Based Economy: A First Step Towards Selection, Definition and Description. A proposal by the Working Group on key competencies, set up by the European Commission in the framework of the ‘Objectives Report’.

[2] RYCHEN, D. S. & SALGANIK, L. H. (2003): Introduction. In: D. S. RYCHEN & L. H. SALGANIK (Eds.): Key Competencies for a Successful Life and a Well-Functioning Society. Hogrefe & Huber, Cambridge, pp. 1-12.

[3] FISHER, D., SMITH, M. & WELSER; H. T. (2006): You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In: R. H. SPRAGUE (Ed.): Proceedings of the 39th Annual Hawaii International Conference on System Sciences. IEEE Computer Society, Los Alamitos.

[4] JOHNSON, W. & BEAL, C. (2005): Iterative Evaluation of a Large-Scale, Intelligent Game for Language Learning. In: C.-K. LOOI, G. MCCALLA, B. BREDEWEG & J. BREUKER (Eds.): Artificial Intelligence in Education. IOS Press, Amsterdam, pp. 290-297.

[5] SHAW, E. (2005): Assessing and Scaffolding Collaborative Learning in Online Discussions. In: C.-K. LOOI, G. MCCALLA, B. BREDEWEG & J. BREUKER (Eds.): Artificial Intelligence in Education. IOS Press, Amsterdam, pp. 587-594.

[6] WILD, F., STAHL, C., STERMSEK, G. & NEUMANN, G. (2005): Parameters Driving Effectiveness of Automated Essay Scoring with LSA. In: Proceedings of the 9th International Computer Assisted Assessment Conference (CAA). Loughborough, pp. 485-494.

[7] LANDAUER, T. K., LAHAM, D. & FOLTZ, P. W. (2003): Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In: M. D. SHERMIS & J. C. BURSTEIN (Eds.): Automated Essay Scoring. A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, pp. 87-112.

Similarity Measurement Applied to Information Research and Indexing

Y. V. Hoareau, F. Gandon, A. Giboin, G. Denhière, S. Jhean-Larose, W. Lenhard, H. Baier

I. Introduction

THE main goal of the work reported here is to compare two methods for simulating of similarity measures and to identify opportunities to combine them. The first method is based on distance measures between concepts as captured in ontologies ; for this part we rely on « Corese » a semantic web search engine [1]. The second method is based on a similarity measure in a word/document vector space; for this we rely on Latent Semantic Analysis [2] to compare different effects of corpus modification [3] on performances in research-indexing task [4]. We have created different semantic spaces from corpora composed of titles and abstracts of research reports of INRIA[1]

II. EXPERIENCES

A. *Description of semantic spaces*

We built five vector spaces: one space in French with a corpus which contains stop-words and four spaces in English : one with a lemmatised corpus, a second with a corpus which contains stop-words, a third which do not contains stop-words and the last one where the corpus has been augmented by adding broader terms from the ACM98 thesaurus to simulate the inclusion of ontological knowledge in the corpus. All these spaces have been evaluated with three tests corresponding to a particular situation of retrieval of information.

B. *Description of tests*

In the first test procedure, we compare the similarity between an abstract and a set of keywords list. The set of keyword lists is composed of lists of one hundred keywords each. In every set there is a target keywords list that actually corresponds to the abstract tested. We calculate the similarity between the vector corresponding to the abstract and the vectors corresponding to the keyword lists, assuming that the target keywords list should be more similar to the abstract tested than the other keyword lists.

In the second test procedure, we calculate the similarity between a keywords list and all the abstracts composing the data-base (more than 4000). We still have a target abstract that actually corresponds to the keyword list tested, and we postulate that the keywords list tested should be more similar to the target abstract than the other abstracts.

The third test procedure is close to the first experience but instead of calculating the similarity between an abstract and a set of keyword lists, we calculate the similarity between an abstract and a set of five hundred individual keywords. We calculate the similarity measure between the vector corresponding to the abstract and all the vectors corresponding to each individual keyword. Some of these keywords actually are the keywords attached to the abstract by its author. We assume that the keywords that actually corresponds to the abstract being tested should be more similar to the abstract than the others.

C. *Description of experimental material*

We used two sets of research reports to do the test. The first set of research reports has been randomly extracted from the librarian database. The second set of texts has been chosen from the research reports of a research team with a precise research domain (networks). The comparison of the results for the different spaces allows (i) to evaluate the effect of lemmatization, (ii) to evaluate to effect of the subsumption information added in the corpora and (iii) to optimize the parameters of the spaces in both languages to make it possible to build a French-English semantic space.

As an illustration, we present the results of the third experience applied on the first set of research reports, that compare the performances of LSA using with different corpora to find the keywords given by the author for his research report.

The corpus tested are English with stop-words (“English Classic”) *vs* English lemmatised (“English Lemma”) *vs* English with ontological knowledge (“OntoLSA”).

[1] https://hal.inria.fr/INRIA-SOPHIA/fr/

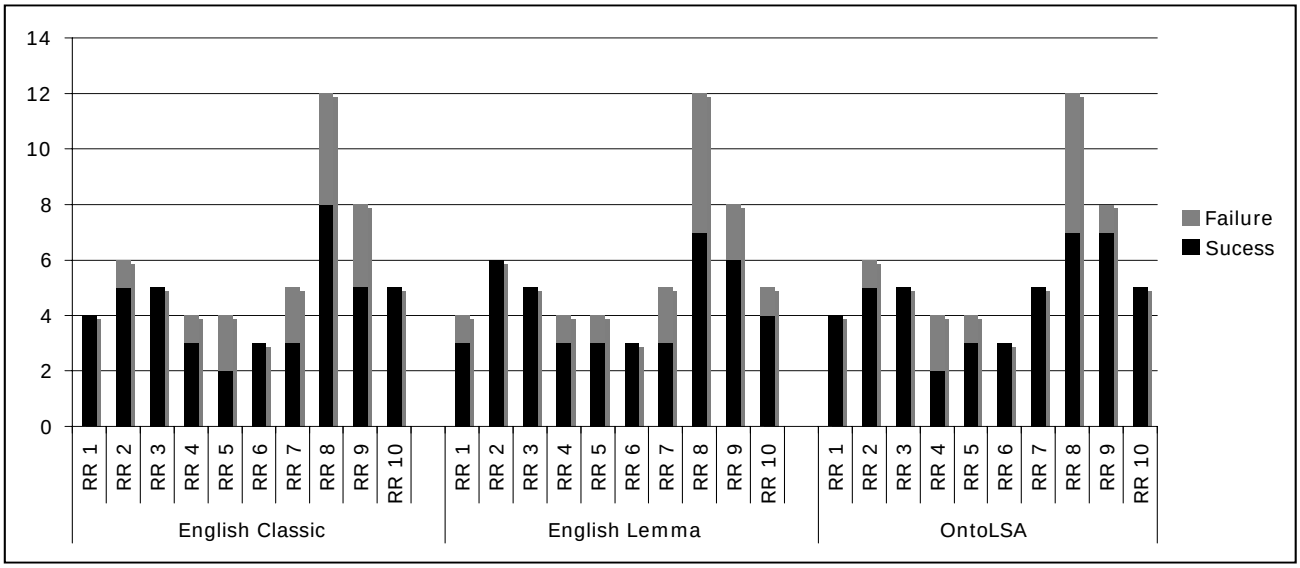

Fig. 1. *Numbers of successes and failures in finding author's keywords for ten research reports using different LSA spaces.*

The results show that the three systems are able to give relatively good performances to find the author's keywords. Lemmatisation gives the best performances in this specific task. The addition of ontological knowledge does not obviously gives the expected better performances for this task.

The experiment that has been developed raises natural queries and makes it possible to compare the performances of the different systems and to evaluate the interest of hybrid systems in the context of document retrieval.

REFERENCES

[1] O. Corby, R. Dieng-Kuntz, C. Faron –Zucker, F. Gandon "Searching the Semantic Web: Approximate Query Processing based on Ontologie", *IEEE Intelligent Systems Journal*, 21 (1), 2006.

[2] Landauer, T. K., Dumais, S. T. "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge", *Psychological Review*, 104(2) , 211-240, 1997.

[3] G. Denhière, B. Lemaire, C. Bellissens, S. Jhean-Larose, "A semantic space for modeling children's semantic memory" In D. McNamara, T. Landauer, S. Dennis, W. Kintsch (Eds). *The handbook of Latent Semantic Analysis.* Mahwah: Lawrence Erlbaum Associates, 2007.

[4] S. Deerwester, S. Dumais, G. Furnas, T. K. Landauer, R. Harshman, "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

Combining LSI with other Classifiers to Improve Accuracy of Single-label Text Categorization

Ana Cardoso-Cachopo
IST — TULisbon / INESC-ID; Av. Rovisco Pais, 1;
1049-001 Lisboa — Portugal; acardoso@ist.utl.pt

Arlindo L. Oliveira
IST — TULisbon / INESC-ID; Rua Alves Redol, 9;
1000-029 Lisboa — Portugal; aml@inesc-id.pt

***Abstract*—This paper describes the combination of k-NN and SVM with LSI to improve their performance in single-label text categorization tasks, and the experiments performed with six datasets to show that both k-NN-LSI (the combination of k-NN with LSI) and SVM-LSI (the combination of SVM with LSI) outperform the original methods for a significant fraction of the datasets. Overall, both combinations present an average Accuracy over the six datasets used in this work that is higher than the average Accuracy of each original method. Having in mind that SVM is usually considered the best performing classification method, it is particularly interesting that the combinations perform even better for some datasets.**

I. Introduction and Experimental Setting

The main goal of *text categorization* (TC) is to derive methods for the categorization of natural language text. The objective is to derive methods that, given a set of training documents with known categories, and a new document, which is usually called the *query*, will predict the query´s category. In this paper, we are interested in the case where the query belongs to a single category, also called *single-label text categorization*. In our work, we present the results obtained using some well known classification methods, namely the Vector method [8], k-NN [7], [9], LSI [4], and SVM [6], and compare them with the results obtained using a combination of k-NN and SVM with LSI.

To allow the comparison of our work with previously published results, we used three standard TC collections in our evaluation, namely the 20-Newsgroups, Reuters-21578 and Webkb. We also used two other collections, Bank which is a collection of messages sent to a bank's help-desk along with their respective answers, and Cade which is a collection of webpages from a Brazilian web directory. To be consistent with 20-Newsgroups and Reuters-21578, for Webkb, Bank, and Cade, we randomly split the documents into two thirds for training and the remaining for testing. Table I contains relevant information regarding the sizes and document distributions for the six datasets.

Regarding algorithm implementation and the parameters that were used, for the Vector method we used a Sourceforge project called IGLU [5]. For k-NN we implemented a "voting strategy", where the possible classes of a document are voted on by the documents that belong to that class; we used cosine similarity and considered only the 10 nearest documents. For LSI we used FAQO [2], and considered a reduced matrix with 200 dimensions. For SVM we used LIBSVM [3]and used a linear kernel in our experiments. We implemented the combinations between methods as described bellow.

II. Combinations Between Methods

This section describes the rationale behind the combination of k-NN and SVM with LSI, that ideally will perform better than the original methods.

The difference to the original approaches is that now, instead of applying their transformations to the usual term/document matrix used in the vector method, the combinations apply their transformations to the concept space that was previously obtained using Singular Value Decomposition. Then, given another p-dimensional vector representing the query document, choose one of the options:

- Apply to the query document the same transformation as the one applied to the initial term/document matrix; apply cosine similarity to the transformed query and to each of the transformed train documents, select the k most similar documents, apply a voting strategy, where each transformed document "votes" for its class, weighted by its similarity to the transformed query; the class of the query is the most voted class — k-NN-LSI method. This method was already proposed in [1].
- Apply a kernel function to the transformed concept matrix, so that concepts are represented in a high dimensional feature space, where each class is linearly separable from the others; apply to the query document the same transformation applied to the initial term/document matrix; apply a voting strategy, where possible classes are ranked according to the number of votes that they had in a one-against-one classification approach; the class of the query is the class which got more votes — SVM-LSI method.

III. Comparing the combinations with the original methods

Table II shows Accuracy values obtained by each method for each of the six datasets, and average Accuracy over all the datasets for each method. The Dumb classifier ignores the query and always predicts the most frequent class in the training set, and is included here to provide a baseline Accuracy value.

When comparing k-NN-LSI with k-NN and LSI, it is important to note that, from the two original methods, there

Dataset	Collection	Classes	Train Docs	Test Docs	Total Docs	Smallest Class	Largest Class
Bank37	Bank	37	928	463	1391	5	346
20Ng	20-Newsgroups	20	11293	7528	18821	628	999
R8	Reuters-21578	8	5485	2189	7674	51	3923
R52	Reuters-21578	52	6532	2568	9100	3	3923
Web4	Webkb	4	2803	1396	4199	504	1641
Cade12	Cade	12	27322	13661	40983	625	8473

TABLE I
DESCRIPTION OF THE DATASETS: COLLECTION FROM WHICH IT IS DERIVED, NUMBER OF CLASSES, NUMBER OF TRAIN DOCUMENTS, NUMBER OF TEST DOCUMENTS, TOTAL NUMBER OF DOCUMENTS, NUMBER OF DOCUMENTS IN THE SMALLEST CLASS, NUMBER OF DOCUMENTS IN THE LARGEST CLASS.

Dataset	Dumb	Vector	k-NN	SVM	LSI	k-NN-LSI	SVM-LSI
Bank37	0.2505	0.8359	0.8423	0.9071	0.8531	0.8488	0.9179
20Ng	0.0530	0.7240	0.7593	0.8284	0.7491	0.7557	0.7775
R8	0.4947	0.7889	0.8524	0.9698	0.9411	0.9488	0.9680
R52	0.4217	0.7687	0.8322	0.9377	0.9093	0.9100	0.9311
Web4	0.3897	0.6447	0.7256	0.8582	0.7357	0.7908	0.8897
Cade12	0.2083	0.4142	0.5120	0.5284	0.4329	0.4880	0.5465
Average	0.3030	0.6961	0.7540	0.8383	0.7702	0.7904	0.8385

TABLE II
ACCURACY OBTAINED BY EACH METHOD FOR EACH OF THE SIX DATASETS, AND AVERAGE ACCURACY OVER ALL THE DATASETS FOR EACH METHOD.

is none that always outperforms the other. For Bank37, R8, R52, and Web4, LSI performs better than k-NN, whereas for 20Ng and Cade12 it is k-NN that provides better results. This is probably because LSI is very effective at finding the "concepts" in the first datasets, but for the other datasets, which consist of newsgroup messages (that can quote others) and web pages (that can be copies of others), finding the most similar document is more effective. For R8, R52, and Web4, the combination of k-NN with LSI is the best performing method. For the other datasets, k-NN-LSI is second best, independently of which of the original methods shows a better performance, and its Accuracy is closer to the one achieved by the best method. As can be seen in Table II, if one considers average Accuracy over the six datasets, k-NN-LSI is the best performing method, when compared to k-NN and LSI.

When comparing SVM-LSI with SVM and LSI, we can see that, for all datasets, the worst performing method is LSI. SVM-LSI is the best for the datasets in Portuguese, Bank37 and Cade12 and also for Web4. SVM is the best for the datasets in English, R8, R52 and 20Ng. Having in mind that SVM was the best performing method in several comparisons of classification methods [9], [1], it is particularly interesting that its combination with LSI performs even better for some datasets. As can be seen in Table II, when one considers average Accuracy over the six datasets, SVM-LSI slightly outperforms SVM, even if this difference is not significant.

Given the results obtained, it is important to note that the best performing method depends on the dataset that is used. As such, it is important to test different methods and combinations to decide which one to use in each situation.

IV. CONCLUSIONS AND FUTURE WORK

We described the combination of k-NN and SVM with LSI and showed that SVM-LSI is the method that presents the best performance for some of the datasets that were used in our experiments.

We think that, given the present results, this kind of method combination represents an interesting line of research, and that more tests need to be done, namely regarding the combinations between the number of dimensions that are considered in the LSI method and the kernel function that is used by the SVM method.

REFERENCES

[1] A. Cardoso-Cachopo and A. Oliveira. An empirical comparison of text categorization methods. In *Proceedings of SPIRE-03*, pages 183–196. Springer Verlag, 2003.

[2] J. Caron. Experiments with LSA scoring: Optimal rank and basis. Presented at SIAM Computational Information Retrieval Workshop, 2000.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] IGLU Java—Java Library for Information Retrieval Research, 2002.

[6] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142. Springer Verlag, 1998.

[7] B. Masand, G. Linoff, and D. Waltz. Classifying news stories using memory-based reasoning. In *Proceedings of SIGIR-92*, pages 59–65. ACM Press, 1992.

[8] G. Salton. *The SMART Retrieval System*. Prentice Hall, 1971.

[9] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of SIGIR-99*, pages 42–49. ACM Press, 1999.

Using LSA for word prediction: A comparison of integration techniques

Tonio Wandmacher and Jean-Yves Antoine, *LI, Université François-Rabelais de Tours, France*

***Abstract* — Many word prediction systems make use of n-gram based statistical language models (LM) to estimate the probability of the following word in a phrase. In the past years there have been many attempts to enrich such language models with further syntactic or semantic information. We want to explore the predictive powers of Latent Semantic Analysis, which has shown to provide reliable information on long-distance semantic dependencies between words in a context. We present and evaluate here several methods that integrate LSA-based semantic information with a standard language model: a semantic cache, a reranking approach, and different forms of interpolation. We found that all methods show significant improvements, compared to the 4-gram baseline, and most of them to a simple cache model as well.**

I. Introduction

AUGMENTED and Alternative Communication (AAC) aims at restoring the communicative abilities of disabled people with severe speech and motion impairments. These people have to communicate via an AAC system that enables them to enter messages with single switch devices (triggered by eye-glimpses, breath or head movements). Since this mode of communicating is extremely slow (1 to 5 words per minute) and also very tiring, AAC systems usually incorporate a word predictor, i.e. a module that estimates themost likely words to appear and proposes them to the user, who can then select among them. Whereas most commercial systems only offer a lexicon-based prediction, AAC systems such as *FASTY* [1] or *SIBYLLE* [2] normally use n-gram language models (LM) to estimate the probability of occurrence for a word.

N-gram based word prediction has shown to give very good results in terms of the *keystroke saving rate* (*ksr*), the evaluation metric normally used for word predictors. Based on a simple 4-gram model, our system *SIBYLLE* has a *ksr* of up to 54,4%, i. e. less than half of the characters of a message still have to be typed (or selected from a virtual keyboard).

In the past 15 years a variety of approaches has been presented that try to improve n-gram language models by incorporating syntactic or semantic information. In particular, Coccaro and Jurafsky [3] integrated LSA based similarity measures to a bigram model, by using different types of interpolation.We present and discuss here three different techniques for incorporating information from LSA to an LM: (a) a semantic cache model, (b) a reranking approach, and (c) (confidence-weighted) interpolation. We evaluate these techniques on our word predictor in terms of *ksr*, based on a prediction list of 5 words (ksr_5), and the often used perplexity measure (PP).

II. Integration methods

A. Semantic cache:

The underlying idea of cache models is that words having already occurred in the context are likely to occur another time. Their probability is therefore raised by a constant or an exponentially decaying factor, depending on the actual position of the element in the cache. This rather simple model has shown to bring slight but constant perplexity reductions [4], [5]. We extend this model by calculating for every cache element its *n* nearest LSA neighbours and add them to the cache as well.

B. Reranking:

A particular problem of LSA based probabilities is that the a priori probability of a word is not taken into account; rare words are then highly overestimated, whereas very common words (function words) are neglected. Our reranking approach therefore considers only the *n* (e.g. 1000) best candidates from the base LM for the LSA component. These *n* best candidates are then reranked, according to their LSA similarity with the context vector.

C. Standard and confidence-weighted Interpolation:

Interpolation is the traditional way to integrate information from heterogeneous resources. Whereas linear interpolation (LI) simply adds the (weighted) probabilities of two components, geometric interpolation (GI) multiplies them; weighting is then performed by exponential coefficients.

In standard approaches, the coefficients remain stable, Coccaro and Jurafsky [3] however used a dynamic weighting-scheme. They observed that LSA gives much more reliable estimates for words having a low entropy, they therefore controlled for each word the influence of the LSA component by applying an entropy-based confidence metric. However it seems that entropy does not correlate very much with the mapping quality of a word in the LSA vector space.

Especially words with a very low entropy are not reliable predictors. We therefore introduce a new confidence metric which relies on the average distance of the *n* nearest neighbours around a word vector, (*density* of the word cluster). The closer the neighbours are (i.e. the higher the density of a word is), the better a word can be predicted by the LSA component.

III. Results

Our baseline 4-gram model was calculated on a 44M word corpus from the French newspaper *Le Monde*, using the SRI toolkit ([6], http://www.speech.sri.com/projects/srilm/).

For calculating the semantic space we used the *Infomap* software (v. 0.8.6, http://infomap.stanford.edu/), which works on a term × term co-occurrence matrix. We generated a co-occurrence matrix (size = 80.000 rows × 3.000 cols; co-occurrence window = ±100) from a 100M corpus (again from *Le Monde*). After applying Singular value decomposition, the resulting matrix of term vectors was reduced to 150 dimensions. We calculated for all methods described above *ksr* and perplexity on a test corpus (58.743 words), which was split into 8 samples.

All methods had significant gains over the 4-gram baseline (ksr_5 = 54,4%; PP = 109,2), and most of them over the baseline + cache as well. Whereas the semantic cache model, the reranking method and simple forms of interpolation were rather close (ksr_5 between 55,1% and 55,3%; PP from 99,9 to 101,5), density-weighted geometric interpolation scored best (ksr_5 = 55,7%; PP= 98,6). The gains obtained by applying LSA are not very high, still they are significantly better than those of a simple 4-gram model + cache, which marks already a strong baseline. Moreover, what we have not considered yet, are the qualitative gains for persons using an AAC system, who can feel cognitively supported by semantically related predictions. A qualitative evaluation still remains to be done.

Figures 1 and 2 display the overall results in terms of *ksr* and perplexity:

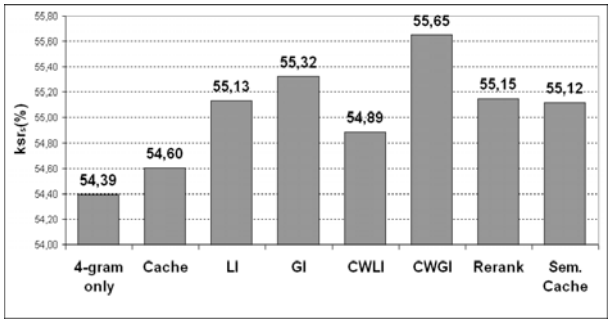

Fig. 1: Results (ksr_5) for all methods tested.

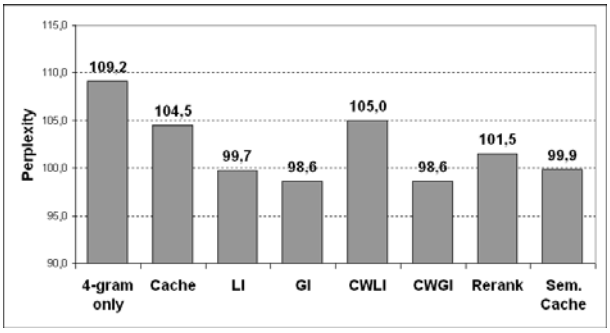

Fig. 2: Results (perplexity) for all methods tested.

Acknowledgements

This research is partially founded by the UFA (Université Franco-Allemande) and the French foundations APRETREIMC (ESAC_IMC project) and AFM (VOLTAIRE project). We also want to thank the developers of the *SRI* and the *Infomap* toolkits for making their programs available.

References

[1] Trost, H., M.J., Baroni, M.: "The Language Component of the FASTY Text Prediction System". *Applied Artificial Intelligence* 19(8), 2005, pp. 743–781.

[2] Schadle, I.: Sibylle: "Système d'aide à la communication pour les personnes handicapées". PhD thesis, Université de Bretagne Sud, 2003.

[3] Coccaro, N., Jurafsky, D.: Towards better integration of semantic predictors in statistical language modeling. In: *Proceedings of ICSLP-98*, Sydney, 1998.

[4] Kuhn, R., De Mori, R.: "A cache-based natural language model for speech reproduction". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(6), 1990, pp. 570–583.

[5] Clarkson, P.R., Robinson, A.: "Language model adaptation using mixtures and an exponentially decaying cache". In: In *Proceedings IEEE International Conference on Speech and Signal Processing*, München, 1997.

[6] Stolcke, A.: "SRILM - an extensible language modeling toolkit". In: *Proceedings of the Intl. Conf. on Spoken Language Processing*, Denver, Colorado, 2002.

An LSA Package for R

Fridolin Wild, *Vienna University of Economics and Business Administration, Austria*

***Abstract* — Latent semantic analysis (LSA) is an algorithm applied to approximate the meaning of texts, thereby exposing semantic structure to computation. LSA combines the classical vector-space model − well known in computational linguistics − with a singular value decomposition (SVD), a two-mode factor analysis. Thus, bag-of-words representations of texts can be mapped into a modified vector space that is assumed to reflect semantic structure. In this contribution the author describes the *lsa* package for the statistical language and environment *R* and illustrates its proper use through an example from the area of automated essay scoring.**

I. Concepts used in latent semantic analysis

When applying a latent semantic analysis (Deerwester et al., 1990), a process is executed that typically involves several (optional) steps and involves various data-types created as an output of these steps. To clarify which entities and processes are involved when performing an LSA, the following concepts shall be defined.

Term: The 'word' as it is written down in a document.

Corpus: The collection of documents containing texts that consist out of terms separated by punctuation marks.

Textmatrix: A representation of the document collection in matrix format: the cells contain the frequency, how often a particular term appears in a specific document. Terms are the rows, documents the columns. By transforming a corpus to this representation format, documents are treated as so-called bag of words, where the term order is neglected.

Latent-Semantic Space: When applying a singular-value decomposition (SVD) to a textmatrix, the matrix is resolved into the term-vector matrix T (constituting the left singular vectors), the document-vector matrix D (constituting the right singular vectors) being both orthonormal and the diagonal matrix S (Berry et al., 1995). These partial matrices are then truncated in order to reflect strong associations, eliminate noise, etc. The set of these three, truncated partial matrices T_k, S_k, and D_k is called 'latent-semantic space'. A latent-semantics space can be converted back to textmatrix format.

Folding In: To keep additional documents from changing the structure of a latent-semantic space, documents can be folded into the previously calculated space. Thereby, T_k and S_k of the space are re-used and combined with a textmatrix constructed over the new documents. See Wild and Stahl (2006) for more details.

Dimension: When truncating the partial matrices from the SVD, a particular number of the highest singular values are retained. This is called the 'dimensionality' of the latent-semantic space.

Distance / Similarity: Within a textmatrix, various methods can be applied to measure the distance (or, the other way round, similarity) between terms, documents, or terms and documents. One method is, e.g., to use the measure the cosine of the angle between two column-vectors in order to calculate the similarity between two documents. A high cosine value is equal to a small angle between the vectors, thus indicating high similarity.

Vocabulary: All terms used within a corpus form the vocabulary of this corpus. The vocabulary has a certain order to ensure that additional text matrices can be constructed that can be appended to an existing textmatrix.

II. The Process

A typical LSA process using the R package looks similar to the one depicted in Figure 1. First, a textmatrix is constructed with textmatrix() from the input corpus. The textmatrix can (but does not need to be) weighted using one of the various weighting schemes provided (see Wild (2005) for more details). Then, the singular-value decomposition is executed over this textmatrix and the resulting partial matrices are truncated and returned by lsa(). The number of dimension to keep can be set using various recommender routines (e.g., dimcalc_kaiser()). The resulting latent-semantic space can be converted back to textmatrix format using as.textmatrix().

Manuscript received October 9, 2001. (Write the date on which you submitted your paper for review.) This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

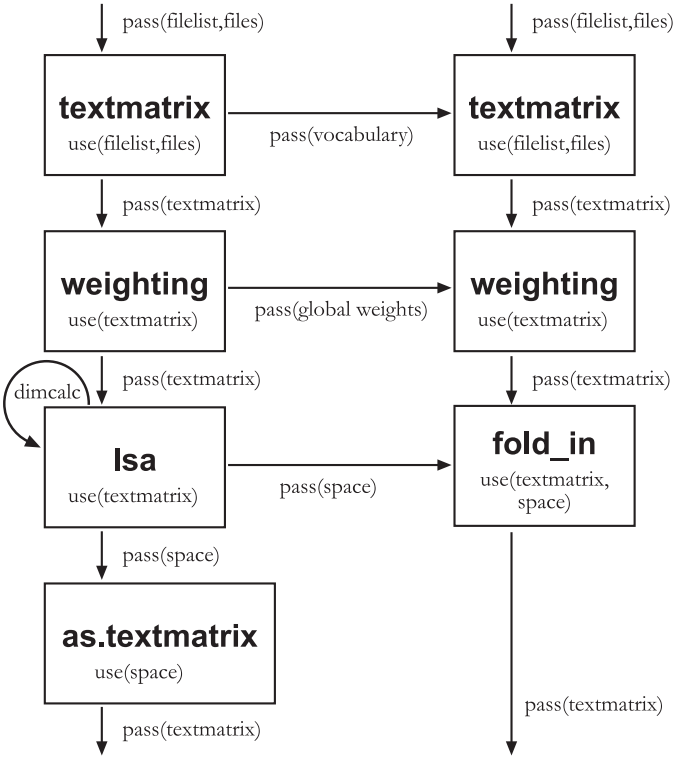

Figure 1. Overall Workflow.

In case that additional documents are to be folded into the existing latent-semantic space, again a new textmatrix is constructed using textmatrix() re-using the vocabulary of the first. Again the resulting textmatrix can be weighted (eventually re-using the global weights of the first textmatrix). Using fold_in(), the resulting textmatrix can be mapped into the existing latent-semantic space, thereby re-using the truncated left-sided and the diagonal partial matrices of the SVD. In this case, the result is directly a textmatrix.

III. Sanitizing Corpora with the Package

Looking more closely at the textmatrix routine, it can be seen that several text sanitizing and pre-processing steps are embedded in the textmatrix generation routines: the routine included means to convert the terms to lower case, simple routines for stripping XML tags, automatic removal of punctuation marks and some other special characters, and trimming of multiple white spaces. Furthermore, stop words can be filtered (by providing stop word lists) or a controlled vocabulary can be deployed. Furthermore, frequency filters can be applied to delete terms below or above a certain frequency threshold (within a document or within the corpus) or outside a certain term-length range. Terms consisting purely of numbers can be removed automatically. Also all terms can be reduced to their word stems (using Porter's snowball stemmer).

The package is open-source and available via CRAN, the Comprehensive R Archive Network.

IV. Demonstration

The following code example illustrates how LSA may be used to automatically score free-text essays in an educational assessment setting.

```
library("lsa")

# load training texts
trm = textmatrix("trainingtexts/")
trm = lw_bintf(trm) * gw_idf(trm) #
weighting
space = lsa(trm) # create LSA space

# fold-in test and gold standard essays
tem = textmatrix("essays/",
vocabulary=rownames(trm))
tem = lw_bintf(tem) * gw_idf(tem) #
weighting
tem_red = fold_in(tem, space)

# score essay against gold standard
cor(tem_red[,"gold.txt"],
tem_red[,"E1.txt"]) # 0.7
```

Listing 1. Essay Scoring Example.

References

[1] DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T., HARSHMAN, R. (1990): Indexing by Latent Semantic Analysis. JASIS, 41, 391--407.

[2] BERRY, M., DUMAIS, S. and O'BRIEN, G. (1995): Using Linear Algebra for Intelligent Infor-mation Retrieval. SIAM Review, 37, 573—595.

[3] WILD, F. (2005): lsa: Latent Semantic Analysis. R package version 0.57.

[4] WILD, F., STAHL, C. (2006): Investigating Unstructured Texts with Latent Semantic Analysis. In: Lenz, Decker (Eds.): Advances in Data Analysis, Springer, 2007

New LSA Education Applications at University of Colorado and Pearson Knowledge Technologies

Thomas K Landauer, Peter W. Foltz

I. Introduction

Pearson Knowledge Technologies (a subsidiary of Pearson, Plc.), in collaboration with the Institute of Cognitive Science at the University of Colorado has been steadily improving and extending its educational LSA [1] and related [2] applications since the first public availability of the Intelligent Essay Assessor in 1998 [2]. The Intelligent Essay Assessor itself has undergone many significant advances in the ways in which LSA is applied, and the aspects of essay scoring with which it deals. It has incorporated other computational language modeling techniques and added a variety of new features and scoring capabilities. Its software, data storage capacity, response time has been reduced while feedback detail has increased. And it has received extensive and varied confirmation of reliability and validity.

During the same period, LSA has been combined with other statistical analysis and machine learning methods to create a suite of complementary educational tools. These include:

- *Summary Street* presents texts, students summarize them in many fewer words and receive immediate feedback on how well they have understood and expressed the important aspects of the content of the reading.
- Summary Street has recently been combined with IEA in an integrated reading and writing literacy tutorial and assessment tool called *WriteToLearn*.
- *SuperManual* is an automatically produced digital instruction tool in which LSA makes learning objects easier to locate and understand by providing meaning-based search, summaries and optimum learning paths
- *Standard Seeker*_automatically aligns instructional texts and test items with compendia of learning standards
- *Career Map* automatically matches educational and work experience with job and training programs.
- PKT's automatic *metadata tagger*, annotates the content of learning object repositories in keywords that best express the central content of paragraphs or longer text using words that are not necessarily from the text itself, and with semantically most representative sentences and classifications into pre-specified categories
- *Open-Cloze* and *Meaningful Sentences* are new web-delivered and automatically scored constructed response reading and writing exercises with immediate feedback
- *Team Communications* and *Knowledge Post* are both LSA-based systems that "listen in" to communication generated during group training or learning activities and provide immediate and aggregated automatic mentoring, assessment and real-time moderator intervention.

In the current R&D pipeline are technologies to select the most important words for vocabulary instruction, ones to modify the reading difficulty of texts to suit individual students and techniques for choosing readings that will maximize growth of useful vocabulary, reading comprehension and writing ability. Fuller descriptions of a selection of these tools and evidence supporting their educational utility follows.

II. The new Intelligent Essay Assessor

IEA now uses LSA in several more ways than in its original instantiation. It still uses as a principle variable a near-neighbor algorithm to accomplish what we call "direct estimation", basing a score for an essay on the scores that humans have assigned to other essays that are highly similar. It also avoids using variables, such as number or rarity of words, that can be easily counterfeited or coached, demanding instead appropriate amounts of relevant semantic content and the use of words that express it well. It uses the same approaches to evaluate various writing traits of an essay as requested by testing agencies.

Recent versions also use LSA to evaluate coherence and to construct a monotonic alignment of all the essays in a testing program to produce a score prediction that does not use human scores at all. LSA is also used to detect unusual and outlier essays, including the highly creative ones that pundits often fear will be mistreated, ones that are off topic, and--with very great accuracy--ones that have been copied from others with attempts at disguise by word substitution and rearrangement.

In general, a different scoring model, that is, one in which the variables that go into its predictions and their importance, are selected and optimized by analysis of a human training set.

Most of IEA's mechanisms were originally designed to

Thomas K Landauer and Peter W. Foltz University of Colorado and Pearson Knowledge Technologies, 4940 Pearl East Circle, Suite 200, Boulder, CO, 80305, 303-545-9092; (e-mail: landauer@pearsonkt.com, pfoltz@pearsonkt.com)

evaluate the substantive content of expository essays. However, popular applications have often centered on writing quality as such rather than knowledge. Therefore, we have recently created a version that measures only essay characteristics that are desirable in essays on any topic, such as coherence and choice and order of words appropriate to whatever is being said. This move creates a scoring model that is prompt independent and can be applied with prior human scores for training.

A. *IEA reliability and validity.*

Over tens of thousands of essays to hundreds of prompts, IEA score agreement and correlation with human marks has been statistically equal to or greater than the agreement between two trained human raters. For example, in one-evaluation conducted by an independent testing organization, some 3,000 essays were written by 4th to 12th grade students to 12 different prompts in a specially constructed experimental design in which each student wrote on six essay prompts, one each on six testing days, each essay was scored independently by four different readers, and prompts, day, reader and day occurred in every possible counterbalanced combination.

This made it possible to evaluate fundamental scoring reliability extremely accurately by statistically holding constant all variables except whether an essay was scored by humans or IEA. IEA correlated with the *human* readers significantly, p<. 01, better than they correlated with each other.

In addition, in this study, half of each set of prompts were used with students in two different school grades, each pair of grades two years apart. This allowed us to compare how well human and IEA scores measured average progress over two years of schooling. The answer: a tie.

In another large study, we analyzed essays on 81 different reading-related topics answered by students in grades 6 through 12 online in a web-based companion to a Prentice Hall reading series, a very similar application as in WTL. The correlation between IEA and Humans was better than that between the two human readers for every grade level, by an average of .037, with probability less than one in a thousand. Exact agreements on the 6 point scale were nearly identical, 61.1% for IEA to human and 61.7% for human to human.

III. Summary Street

To date, more than 3,000 students have used Summary Street. In one field study, after four classroom uses, both blind teacher ratings and scores on some relevant reading and writing items from a government sponsored objective test were significantly better--effect size over 1 for the lower 75% of students--than for random controls. Students find Summary Street motivating, rewarding and fun, and teachers have greeted it with virtually unanimous enthusiasm.

IV. Team Communication:

Over a series of studies, LSA-based communications analysis methods have demonstrated practical ability to predict team performance. (see [3] for a review). Using human and ASR transcriptions of team missions, LSA predicted both objective team performance scores and subject-matter expert ratings with correlations ranging from r=0.5 to r=0.9 over 20 tasks.

For example, using human transcriptions of 67 team missions in a unmanned vehicle environment, LSA predicted objective team performance scores with r = 0.79. The Team Performance Score used as the criterion measure was a composite of objective measures including the amount of fuel and film used, the number and type of photographic errors, route deviations, time spent in warning and alarm states, unvisited waypoints and violations in route rules.

A. *Knowledge Post*

Knowledge Post supports the ability:

- To find material similar in meaning to a given posting in ongoing or prior discussions or in an electronic library.
- To have contributions automatically summarized by hovering the mouse over the subject of the note.
- To have expert comments or library articles interjected into the discussion in appropriate places by automatically monitoring the discussion board activity.
- To automatically notify the instructor when the discussion goes off track.
- To enhance the overall quality of the discussion and consequent learning level of the participants..

In a study at the U.S. Air Force Academy, cadets who received automatically selected expert (i.e., senior military officer) comments made contributions of significantly higher quality than those in control instructor led classes or electronic discussion groups. [4].

V. Conclusion

There have been other applications as well, for example an experimental LSA-based method for automatic indexing of books by the central meaning of pages rather than by single words and phrases, used for the index of the new *Handbook of Latent Semantic Analysis* [5]. Overall, our experience has been that LSA offers an enormous spectrum of valuable opportunities for educational tools.

References

[1] Landauer, T. K, Laham, D. and Foltz, P.W. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In, Shermis, M. D and Burstein, J. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum.

[2] Landauer, T.K., Laham, D., & Derr, M. (2004). From paragraph to graph: Latent semantic analysis for information visualization. Proceedings of the National Academy of Science, 101, 5214-5219.

[3] Foltz, P. W., Martin, M. J., Abdelali, A., Rosenstein, M. B. & Oberbreckling, R. J. (2006). Automated Team Discourse Modeling: Test of Performance and Generalization. In *Proceedings of the 28th Annual Cognitive Science Conference.*

[4] LaVoie, N, Streeter, L., Lochbaum, K. Boyce, L., Krupnick, C., and Psotka, J. Automating Expertise in Collaborative Learning Environments. *International Journal of Computer-supported Collaborative Learning.* Submitted.

[5] Landauer, T. K, McNamara, D.S., Dennis, S., Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Human Hierarchization of Semantic Information in Narratives and Latent Semantic Analysis

Guy Denhière, Vigile Hoareau, Sandra Jhean-Larose,
Wolfgang Lehnard, Herbert Baïer, & Cedrick Bellissens

***Abstract* — The goal of this paper was to investigate the children and adults cognitive processes implied in relative importance judgements of sentences and to simulate them with Latent Semantic Analysis (Landauer & Dumais, 1997). In a previous study, Lemaire, Mandin, Dessus & Denhière (2005) did not obtain significant correlations between human judgments of relative importance of sentences from narratives and models based on Latent Semantic Analysis. In this study, we used 16 calibrated narratives extracted from DIAGNOS™ material. We have compared human judgments of relative importance (JIR) of the 30 sentences for each narratives to the predictions derived from LSA models using adult and children semantic spaces (Denhière, Lemaire, Bellissens, Jhean-Larose, 2004). As predicted, significant correlations were obtained with children semantic spaces but not with adult semantic spaces.**

I. Introduction

THE goal of this study was to investigate cognitive processes implied in judgement of relative importance (JRI) of narratives' text sentences. We hypothesised that Latent Semantic Analysis model (LSA; Landauer & Dumais, 1997), were useful to simulate human judgments of relative importance (JIR) (Lemaire, Mandin, Dessus & Denhière, 2005), considering that important text sentences should have a high semantic similarity with the whole text.

II. Human Experiment: Material and Judgments of Relative Importance

16 narratives extracted from DIAGNOS™ tests (Baudet & Denhière, 1990) were used here. They did not differ for the number of words, sentences, semantic propositions, number of arguments, etc… Each narrative was divided in 30 sentences and 4 groups of 30 adults (students) were asked to judge the relative importance of each of the 30 statements (of 4 narratives) on a 5 points scale. So, for each of the 16 narratives, we obtained an importance value for each sentence, importance value that was significantly correlated with recall and summary by children and adolescents of these texts (Denhière, Thomas & Legros, 2002).

G. Denhière is with the Laboratoire CHArt, Cognitions Humaine et Artificielle, EA 4004, EPHE – CNRS, 46, Rue Gay Lussac, 75006 Paris, France (guy.denhiere@ephe.sorbonne.fr).

V. Hoareau is with the Université de Paris VIII, Laboratoire CHArt, Cognitions Humaine et Artificielle, EA 4004, EPHE, 46, Rue Gay Lussac, 75006 Paris, France (yann-vigile.hoareau@cognition-usages.org).

S. Jhean-Larose is with the I.U.F.M. de Paris et Université de Paris VIII, Laboratoire CHArt, Cognitions Humaine et Artificielle, EA 4004, EPHE, 46, Rue Gay Lussac, 75006 Paris, France (jhean@paris.iufm.fr).

W. Lehnard and H. Baïer are with the Lehrstuhl für Psychologie IV, Universität Würzburg, Röntgenring 10, Raum 109, 97070 Würzburg, Germany (wolfgang.lenhard@mail.uni-wuerzburg.de, baier@psychologie.uni-wuerzburg.de).

C. Bellissens is with the Institute for Intelligent Systems of the University of Memphis, Office 410, 365 Innovation Drive, Memphis TN 38152, USA (cbellissens@mail.psyc.memphis.edu).

III. Semantic spaces

We used different semantic spaces that are implemented at the Cognitive Science Institute, University of Colorado at Boulder (http://lsa.colorado.edu), in the University of Würzburg (http://www.summa.psychologie.uni-wuerzburg.de) and at the Ecole Pratique des Hautes Etudes (EPHE) in Paris and Grenoble. Part of these semantic spaces were supposed to represent adult semantic memory (around 20 million words, with several months of Le Monde daily journal, and literature books)? An other part of the semantic spaces were supposed to represent 7-11 years old children semantic memory. The children corpus, of about 3 million words, included children textbooks and tales (see Denhière & Lemaire, 2004). These corpora were processed by LSA in Boulder and Paris and in a similar system developed by Lehnard & Baier in Wurzbourg and all words were represented as vectors in a 300-dimensional space.

IV. Models derived from LSA

As in Lemaire & al. (2005) paper, we postulate that important sentences have a high semantic similarity with the whole text. Each of the 16 narratives and all of their 30 sentences were represented as vectors in the different semantic spaces. All sentences were assigned a measure of relative importance which was their cosine with the whole text. These measures were correlated with human adult judgments. The second model, instead of considering the text as a whole, breaks it into sentences, the important sentences being supposed to be highly connected to the others.

V. Results

The two models gave similar results. Table 1 and table 2 present correlations between LSA measures and human data

for children and adult semantic spaces (Model 1).

Correlations are higher with children semantic spaces than adult spaces and, for the two kinds of spaces, important differences between narratives are obtained. We tried to explain the differences observed in children spaces implemented in Würzburg by comparing the semantic characteristics of verbs, nouns and modifiers for each of the 16 narratives according to the variables identified by the Würzburg system.

	Paris			Boulder		Würzbourg	
	Enfants	Enfant2	Enfant Total	Contes	Mixte	M1 TEnfant	M2 TEnfant
Géant	0,61	0,61	0,60	0,40	0,48	0,63	0,60
Dragon	0,58	0,58	0,57	0,48	0,36	0,56	0,52
Taureau	0,47	0,46	0,45	0,30	0,44	0,46	0,42
Ane	0,44	0,44	0,43	0,20	0,44	0,41	0,31
Chamois	0,42	0,42	0,41	0,41	0,24	0,47	0,32
Clowns	0,35	0,34	0,33	0,38	0,11	0,21	0,30
Poule	0,26	0,27	0,25	0,02	0,24	0,48	0,35
Rennes	0,22	0,21	0,22	0,44	0,16	0,23	0,21
Araignée	0,20	0,20	0,18	0,17	0,07	0,22	0,32
PièceOr	0,20	0,06	0,04	0,13	0,04	0,05	-0,01
Ourson	0,17	0,17	0,13	0,23	0,28	0,16	0,07
Camion	0,17	0,17	0,24	0,26	0,45	0,33	0,26
Lebrac	0,16	0,16	0,14	0,24	0,10	0,13	-0,06
Dauphin	0,14	0,14	0,19	-0,09	0,22	0,28	0,16
Lion	0,06	0,19	0,23	0,22	0,11	0,21	0,22
Singes	-0,05	-0,07	-0,14	-0,02	-0,20	-0,02	-0,06
Mean	*0,27*	*0,27*	*0,27*	*0,24*	*0,22*	*0,30*	*0,25*
Std. Dev.	*0,19*	*0,19*	*0,19*	*0,17*	*0,18*	*0,19*	*0,19*
Median	*0,21*	*0,20*	*0,23*	*0,24*	*0,23*	*0,26*	*0,28*

Table 1. Correlations between human judgments of relative importance of sentences and LSA measures between sentences and the whole text for children semantic spaces.

	Paris					Boulder		
	Monde93	Monde 95	Monde97	Monde 99	Littérature	Total	Monde	Livres
Poule	0,43	0,41	0,49	0,03	0,22	0,69	0,40	0,61
Camion	0,21	-0,05	0,05	0,32	0,09	0,34	0,40	0,48
Taureau	0,09	-0,06	-0,02	-0,18	0,42	0,36	0,23	0,48
Rennes	0,55	0,29	0,50	-0,15	0,11	0,44	0,29	0,46
Géant	0,27	0,34	0,40	0,07	0,45	0,29	0,38	0,46
Dragon	0,00	0,41	-0,14	-0,08	-0,10	0,09	0,37	0,39
Lebrac	0,29	0,37	0,32	0,01	0,29	0,33	0,23	0,38
Ourson	0,25	-0,01	0,14	0,00	0,18	0,30	0,30	0,35
Chamois	0,20	-0,02	0,12	0,06	0,01	0,35	0,37	0,35
Ane	0,34	0,18	0,15	0,32	0,32	0,26	0,35	0,31
Araignée	0,09	-0,05	0,27	0,21	-0,04	0,16	0,12	0,22
Singes	0,05	-0,05	-0,01	0,26	-0,09	0,10	0,02	0,21
Clowns	-0,09	-0,17	-0,15	-0,10	-0,01	0,02	0,16	0,13
Lion	0,06	-0,05	-0,02	-0,06	0,02	0,11	0,14	0,12
PièceOr	0,14	0,05	0,03	0,16	0,02	0,17	0,20	0,12
Dauphin	0,10	0,09	-0,10	0,25	-0,02	-0,05	0,19	-0,05
Mean	*0,19*	*0,10*	*0,13*	*0,07*	*0,12*	*0,25*	*0,26*	*0,31*
Std. Dev.	*0,17*	*0,20*	*0,21*	*0,17*	*0,18*	*0,18*	*0,12*	*0,17*
Mediane	*0,17*	*0,02*	*0,09*	*0,05*	*0,05*	*0,27*	*0,26*	*0,35*

Table 2. Correlations between humans judgments of relative importance of sentences and LSA measures between sentences and the whole text for adult semantic spaces.

REFERENCES

[1] Denhière, G., Lemaire, B., Bellissens, C. & Jhean-Larose, S. (2005). A semantic space for modeling a child semantic memory. In W. Kintsch & T. Landauer (Eds), Latent Semantic Analysis: A road to meaning (pp. 155-176) Hillsdale, N.J.: Lawrence Erlbaum Associates.

[2] Lemaire, B., Denhière, G., Bellissens, C. & Jhean-Larose, S. (a paraître). A model and a computer program for simulating text comprehension. Behavior Research Methods.

Eye movement analysis and Latent Semantic Analysis on a comprehension and recall activity

David Tisserand, Sandra Jhean-Larose, Guy Denhière

***Abstract* — This study was twofold: First, to use eye-tracking analysis to study the progressive construction of the meaning of a text. And second, to measure the influence of sentences' semantic prominence on the coding realised during the reading activity. This prominence was modulated using some calibrated texts. We measured this semantic prominence using LSA. We expected to find some coding differences between readers' initial treatments (fixation time during the first pass) and rereading treatments (fixation time during the second pass). The first pass analysis explains the construction of sentences' local meaning and the second pass analysis debriefs the integration phase. We also expected some coding variations according to the semantic prominence of the read sentences. The high prominence sentences were expected to be more treated than the less prominent sentences. Finally we analysed the semantic proximity between texts recalls and each of the texts' sentences according to the semantic prominence. We were expected some correlations between the semantic prominence of texts' sentences, the semantic prominence of recalls and the eye-movements during reading.**

I. Experiment

From the University of Nice Sophia Antipolis, 24 students took part in this experiment. We used a personal computer, a 1024*680 CRT screen and an EYEGAZE system to record individual ocular movements.

This experiment was in 2 phases: first subjects were asked to read the text (as many times as they wanted) in order to understand and recall it. Then they were asked to recall the maximum amount of information content in the text.

16 explanatory texts of 12 sentences (from the 25 sentences' texts created by Jhean-Larose, 2005) were constructed following the same structure: an introduction, a conclusion, 2 Core sentences directly related to the main purpose of the text , and 4 Expansion sentences related to each core sentence. 2 of the Expansion sentences explained the concept used in each Core sentence, one related to the agent, the other related to the patient. The 2 other Expansion sentences explained the modification of state mentioned in each Core sentence, one describing a causal relationship with the Core sentence, the other describing a consequence of the action mentioned in the Core sentence.

Due to the lack of space to display items on the screen, we only presented 8 sentences by text. In order to counterbalance the sentences' order of display, we created 4 variants of each text. Text versions were counterbalanced within subjects. Each individual was confronted with the 4 variants. Four texts were displayed in variant 1, 4 texts in variant 2, etc.

Regarding the construction of the texts, the Core sentences were presumed to be more important for the construction of a cognitive mental model of the text. We made the hypotheses that we could predict differences in semantic proximities and in fixation times according to this constructed semantic hierarchy:

- We used a 100983 terms' database (AdulteTotal, Marseille) and LSA to measure the semantic proximity between each of the 8 sentences and:
  - the whole text
  - subjects' recalls
- The fixation duration were measured online during subject reading according to zones defined by sentences (sentence 1 = zone 1, etc.)

Several analyses have been conducted:

II. Latent semantic analysis on items

Data from the latent semantic analysis show the calculated semantic proximities between each text's sentence and the whole text in its displayed version. It seems there is a difference between the semantic proximity cosines revealed by the analysis according to the constructed hierarchy but too few relevant differences were found between these cosines.

III. Eye movement anlysis during reading activity

We measured the fixation times during the first pass (DF1), the second pass (DF2, including all the rereading passes) and in total (DFTot = DF1+DF2). Measures were filtered (±2 SD) and balanced by the sentences' number of letters.

Two analyses of variance were conducted: one according to the constructed semantic hierarchy (introduction, Core, Expansion 1, etc.); the other according to the sentences' order of presentation on screen. We used the second analysis in order to verify the results obtain in the first analysis. The

D. T. was with the Laboratoire de Psychologie experimentale et Quantitative, Nice, France (0044-797-3524101; e-mail: david.tisserand@orange-ftgroup.com).

S. J-L, is with IUFM de Paris et Université de Pris VIII, Paris, 75006 France (e-mail: jhean@pqris.iufm.fr).

G. D. is with the Laboratoire Chart EA 4004, Paris, 75006 France (e-mail: guy.denhiere@ephe.sorbonne.fr).

fixation time averages are presented in Table 2

HIERARCHY	DFTOT	DF1	DF2	ORDER	DFTOT	DF1	DF2
INTRO	101,2	43,3	55,4	**1. INTRO**	101,2	43,3	55,4
C1	97,9	39,2	59,0	**2. C1**	97,9	39,2	59,0
E11	94,3	35,0	58,0	**3. Exp 1**	85,1	36,3	47,6
E12	89,9	36,9	52,4	**4. Exp 2**	90,1	36,8	51,7
C2	91,8	37,3	53,6	**5. C2**	91,8	37,3	53,6
E21	84,0	38,0	44,9	**6. Exp 3**	87,9	37,4	49,4
E22	82,4	37,3	43,4	**7. Exp 4**	87,4	36,9	50,0
CONCL	65,5	36,2	29,1	**8. CONCL**	65,5	36,2	29,1

Table 2. Average fixation time according to the constructed semantic hierarchy and the display order

We observe a relevant modulation of ocular information gathering according to the semantic hierarchy (DFTOT, $F(7,161)=25.27$ $p<.001$). Without introduction and conclusion sentences this effect disappear during the first pass (DF1, $F(5,115)=1.40$ p=n.s) but there is still an effect during the second pass (DF2, $F(5,115)=6.72$ $p<.001$) (See Figure 1). The fixation time is higher on Core sentences (56.3ms/letter) than on Expansion sentences (all together) (49.7ms/letter) during the second pass (DF2, $F(1,23)=19,66$ $p<.001$). After a first pass were subjects seem to read quickly without deep analysis, at the second pass subjects take more time to encode sentences with a semantic content directly related to the main purpose of the text and try to construct a macrostructure to remember the text.

But in the same time we observe a relevant modulation of ocular information gathering according to the sentence's order of presentation (DFTOT, $F(7,161)=25,71$ $p<.001$). This effect is more relevant at the first pass [DF1, $F(7,161)=23,68$ $p<.001$; DF2, $F(7,161)=7,85$ $p<.001$]. A deep analysis reveals that the differences of fixation time between Core sentences C1 and Expansions all together are still relevant (DFTOT, $F(1,23)=12,62$ $p<.01$) and differences of fixation time between Core sentences C2 and Expansions all together are also still relevant (DFTOT, $F(1,23)=6,19$ $p<.05$). Then as there is a difference between fixation time on the Core sentences C2 (91.8ms/letter) and fixation time on the Expansion sentences all together (87.65 ms/letter) we can confirm that there is relevant effect of the semantic hierarchy independently to the display order effect at the second pass.

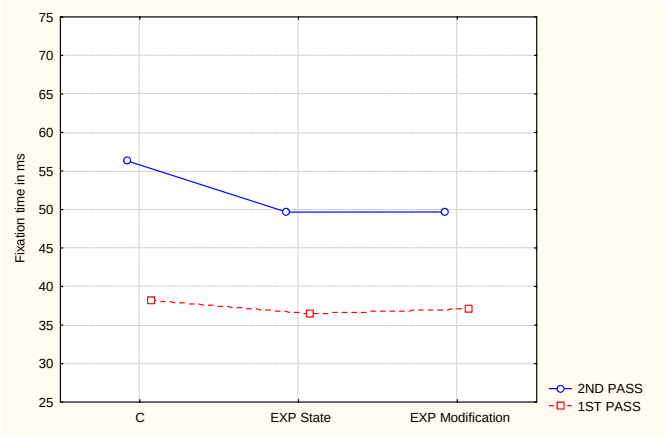

Figure 1. Fixation time at the first pass and at the second pass by letter according semantic hierarchy

The lack of significant results during the first pass indicates a coding mostly guided by the order of presentation. This step seems to be a phase during which individuals identify words, construct semantic proposition and locally organise them within the text (Kintsch and van Dijk, 1978). Then during the second pass the coding seems more structured. The most important sentences in the hierarchy catch more the individual attention. This effect is consistent with Van der Broek et al. theory (2001) in which a sentence with many causal relations is more recalled. It shows a coding at a macro structural level.

IV. Latent Semantic Analysis on Recall

Data from the latent semantic analysis show the calculated semantic proximity between each sentence of a text and the whole recall related to this text. It looks like there is a difference between the semantic proximity cosines revealed by the analysis according to the constructed hierarchy. An analysis of variance has been conducted according to the kind of sentences in order to measure the semantic hierarchy influence on the subjects' recalls.

We observe a relevant modulation of the semantic proximity between subjects' recalls and each sentence of the texts according to the constructed semantic hierarchy, $F(7,161)=44,66$ $p<.001$. Core sentences are in average (all texts together) semantically nearest from subjects recalls (0.3226) than Expansion sentences (0.2646), $F(1,23)=34,44$ $p<.001$.

V. Correlation Analysis

To measure the potential correlations between our three analyses (LSA on items, Eye movement analysis and LSA on recalls) the table 4 have been constructed.

	LSA ITEMS	LSA RECALLS	DF1	DF2	DF Tot
INTRO	0.552911	0.418853	43.3096	55.3958	101.167
C1	0.529598	0.342239	39.2047	59.0118	97.9132
E121	0.504520	0.282384	36.3454	47.6512	85.1017
E122	0.453972	0.261046	36.7818	51.6754	90.1417
C2	0.472325	0.302952	37.3524	53.5839	91.7801
E211	0.426685	0.240846	37.3640	49.3562	87.9474
E212	0.469885	0.274352	36.8572	50.0332	87.4457
CONCL	0.494159	0.314677	36.2045	29.0852	65.4821

Table 4. Average obtained on semantic proximities on items (LSA items) and on recall (LSA recalls) and eye movement analysis on the first pass (DF1), second pass (DF2) and in Total (DFTot) according to the constructed semantic hierarchy

VI. Conclusion

- The constructed semantic hierarchy modulate the ocular information gathering: individuals read more the Core sentences during the second pass.
- We found relevant differences between different kind of text's sentences and subjects' recalls according to the constructed semantic hierarchy: recalls are semantically nearest from the Core sentences.
- Semantic proximities between sentences and whole texts according to the constructed semantic hierarchy are actually computed with different LSA semantic spaces.

References

[1] T. Baccino (2002). Oculométrie cognitive.In G. Tiberghien (Ed.), *Dictionnaire des Sciences cognitives* (pp.100-101), Paris, Armand Colin.

[2] G. Denhière (2005). Rapport final Cognitique: *Le résumé de textes : de l'Analyse de la Sémantique Latente à l'élaboration d'un tuteur électronique*. Ministère délégué à la Recherche et aux nouvelles technologies, Ecole et Sciences cognitives, « La dynamique des apprentissages : des fonctions cognitives à l'élaboration des connaissances ».

[3] G. Denhière, B. Lemaire, C. Bellissens, & S. Jhean-Larose (2005). A semantic space for modeling a child semantic memory. In W. Kintsch & T. Landauer (Eds), Latent Semantic Analysis: A road to meaning (pp. 155-176) Hillsdale, N.J. : Lawrence Erlbaum Associates.

[4] A.C. Graesser, B. Olde & B. Klettke (2002). How does the mind construct and represent stories ?. In M. M.C. Green, J.J. Strange & T.C. Brock (Eds.), *Narrative impact : social and cognitive foundations* ((pp.231-263). Mahwah NJ : Lawrence Erlbaum Associates.

[5] S. Jhean-Larose, & G. Denhière (2006). Etude des processus cognitifs d'interprétation de combinaisons conceptuelles nouvelles, *L'Année Psychologique* , Numéro 2, 265-304.

Modeling Summarization Assessment Strategies with LSA

S. Mandin, B. Lemaire and Ph. Dessus

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

I. Introduction

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

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

II. Description of the Model

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

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

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

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

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

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

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

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

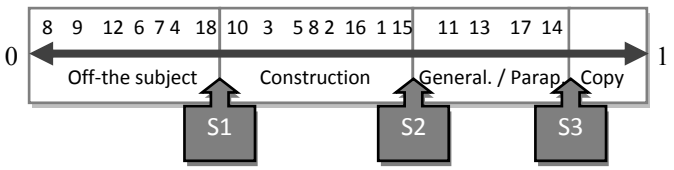

Fig. 1. Representation of the comparisons between a given summary sentence and each source text sentence (represented by numbers). In this example, the summary sentence is classified as a generalization since it is close to several source text sentences.

III. Validation of the Model

33 post-graduate students in educational science from our

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

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

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

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

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

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

IV. Towards a Learning Environment

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

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

References

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

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

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

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

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

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

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

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

Tuning an LSA-based Assessment System for Short Answers in the Domain of Computer Science: The Elusive Optimum Dimension

Debra T. Haley, Pete Thomas, Anne De Roeck, Marian Petre

I. Introduction

THIS paper discusses the results of using Latent Semantic Analysis (LSA) to mark short answers in the domain of computer science. Even though it was introduced almost twenty years ago, questions still remain on exactly how to implement a successful LSA marking system. Two factors contribute to the fact that many researchers have failed to match the results [1] of the early investigators. One factor is that developers must make many choices that affect the results of their assessment systems. The second factor is that there is no standard way of reporting the choices made or the results of those choices. Thus, critical features that improve LSA-based marking systems remain unpublished and unknown to the research community. Adding to the problem, researchers have difficulty comparing various systems and modifications to the basic LSA algorithm [2-5].

We conducted a 2-part experiment to answer one of the most important questions involved in implementing an LSA-based marking system: What is the optimum number of dimensions for the reduced matrix? Part 1 evaluated the results of varying the number of dimensions using the Euclidean distance measure. Part 2 used the best dimensions found in Part 1 to evaluate LSA as a marking tool by using several different evaluation metrics.

Our LSA-based marking system: EMMA (ExaM Marking Assistant) is an LSA-based marking system we are developing to grade short answers to questions in the domain of Computer Science. To mark a student answer, EMMA chooses the five answers in the training data that are closest (using the cosine similarity measure) to the answer being marked. EMMA assigns the weighted average of these tutor-assigned marks to the answer being marked.

The work reported in this study was partially supported by the European Community under the Innovation Society Technologies (IST) programme of the 6th Framework Programme for RTD - project ELeGI, contract IST-002205. This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

The authors are with Centre For Research in Computing, The Open University, Walton Hall, Milton Keynes, UK MK7 6AA. They can be reached by email at: D.T.Haley, P.G.Thomas, A.De Roeck, M.Petre[at] open.ac.uk.

II. The experiment

A. The data

We have 1,000 tutor-marked answers for eighteen different questions. Approximately one-third of the questions were marked by EMMA and two-thirds were used to train EMMA.

B. The two parts of the experiment

We analysed the data in two separate ways. The first analysis found the optimum number of dimensions. Researchers suggest that somewhere between 100 and 500 is a good number [6-9]. We used the numbers 10 through 90 in increments of 10 and 100 through 800 in increments of 100 dimensions to try to bracket the commonly suggested 300 figure. Because of the large number of data points, we needed a quick way to evaluate the results. We chose the Euclidean distance measure, to determine the effectiveness of each of the number of dimensions[10].

The second part of the experiment used the best dimension found in the first part to determine how well EMMA could match human markers. We first used the simple metric of looking at the percentage of marks where EMMA and the human agreed exactly. We next used a more complicated metric that involved looking at the percentage of marks that differed by no more than 8%.

III. Evaluation metrics

A. Euclidean distance

The Euclidean distance measure, or L2, is a metric that tells how far apart two vectors are. It is a standard measure in the study of vectors [11]. If the vectors are identical, L2 is zero. The formula for calculating L2 is:

$$L2(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

where X = (x1, x2, ..., xn) and Y = (y1, y2, ..., yn) are two n-dimensional vectors.

For the purpose of grading short answers, we create two vectors, h and c, where h_i is the mark given by the human

tutor for question i and c_i is the mark given by the computer for question i. The Euclidean distance between these two vectors is the measure of how close the computer marks are to the human marks.

B. Intuitive metric

The purpose of the second part of the experiment was to interpret the best results of the previous part in a more intuitive way than using the Euclidean Distance. We used the simple metric of the percentage of computer-assigned marks that were identical to the human-assigned marks.

C. Modified intuitive metric

The intuitive metric in the previous section uses the percentage of marks where the human and the computer agree exactly. We argue that another reasonable evaluation metric would include marks with exact agreement plus those marks that differ by a "small" percentage. For example, suppose a question is worth twelve points. We suggest that a marking system where the computer and the human disagree by one or fewer points would be acceptable.

The alternative evaluation metric yields different results. To include marks that differ by a small amount means that a higher percentage of answers is assumed to be marked correctly. The question in our data set that has the highest number of points is worth twelve points. Using the idea that an acceptable answer can be off by one point for a question with twelve available points, we have somewhat arbitrarily decided that "small" is 1/12, or 8%. To modify the success metric for this question, we add the percentage of marks that agreed completely with the percentage of marks that differed by one point. Questions with a total point value less than 12 need to be considered a bit differently. For example, suppose a question is worth two points. A mark that is off by one point is quite a large percentage of the total. We can interpolate by taking 8% of the percentage of marks that disagree by 1 point, rather than the full 100% for questions worth 12 points, before adding the amount to the percentage of marks that were identical. For the general case, the formula to modify the success metric is

$$MIM = \%offByZero + (TotalMarksForQuestion / 12) * \%OffByOne$$

IV. Results

The range of optimum dimensions for the 18 questions is 10 to 800 with a mode of 70 and a median of 80. We used the evaluation metrics to understand these results further.

A. Part 1 of the experiment

When using the Euclidean distance measure, we found that changing the number of dimensions had only a small effect on the results. This result might lead us to suggest using a smaller number of dimensions to reduce processing time and memory requirements.

However, using the percentage of answers where the human and the computer agreed 100% of the time showed that, for one question, the percentage of identical answers went from 71% to 83% when using the optimum number of dimensions. The study showed that tuning the number of dimensions, one of the more critical factors of the basic LSA algorithm, can increase the success rate as much as 12%. Thus, we conclude that it is advantageous to determine the optimum number of dimensions individually for each question being marked.

B. Part 2 of the experiment

Part 2 analysed the results from Part 1 in two alternate ways. When using the simple, or intuitive, metric of the percentage of answers where the human and EMMA had identical answers, the results ranged from 36.6% to 97.3%, that is, EMMA agreed exactly with the human marker on 36.6% of the marks for the former question and 97.3% of the marks assigned to the latter question.

Another success metric, the modified intuitive metric, combines the percentage of identical results with a portion of the percentage of results off by ± 1 mark. This metric shows results ranging from 55.6% to 98.9%.

Which metric one prefers depends on whether one wishes to acknowledge that human markers do not always grade perfectly.

References

[1] [1] D. T. Haley, P. Thomas, A. De Roeck, and M. Petre, "A Research Taxonomy for Latent Semantic Analysis-Based Educational Applications," presented at International Conference on Recent Advances in Natural Language Processing'05, Borovets, Bulgaria, 2005.

[2] [2] P. Wiemer-Hastings, "Adding syntactic information to LSA," presented at 22nd Annual Conference of the Cognitive Science Society, 2000.

[3] [3] D. Kanejiya, A. Kumar, and S. Prasad, "Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA," presented at HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing,, 2003.

[4] [4] P. W. Foltz, D. Laham, and T. K. Landauer, "Automated Essay Scoring: Applications to Educational Technology," presented at ED-MEDIA '99, Seattle, 1999.

[5] [5] B. Lemaire and P. Dessus, "A system to assess the semantic content of student essays," *Journal of Educational Computing Research*, vol. 24, pp. 305-320, 2001.

[6] [6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990.

[7] [7] S. T. Dumais, "Data-driven approaches to information access," *Cognitive Science*, vol. 27, pp. 491-524, 2003.

[8] [8] P. Dessus, "Simulating Student Comprehension with Latent Semantic Analysis to Deliver Course Readings from the Web," *Cognitive Systems*, vol. 6, pp. 227-237, 2004.

[9] [9] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge," *Psychological Review*, vol. 104, pp. 211-240, 1997.

[10] [10] D. Haley, P. Thomas, A. De Roeck, and M. Petre, "Measuring Improvement in Latent Semantic Analysis-Based Marking Systems: Using a Computer to Mark Questions about HTML," presented at Proceedings of the Ninth Australasian Computing Education Conference (ACE2007), Ballarat, Victoria, Australia, 2007.

[11] [11] Gerald and Wheatley, *Applied Numerical Analysis*: Addison-Wesley, 1970.

Prior Learning Assessment with Latent Semantic Analysis

Marco Kalz, Jan van Bruggen, Bas Giesbers, and Rob Koper, *Educational Technology Expertise Centre, Open University of the Netherlands*

***Abstract* — Until now most approaches in technology enhanced learning that take into account prior learning stem from learner modeling. In the context of the TENCompetence project we are exploring alternatives to this top down approach for Prior Learning Assessment. We explore Latent Semantic Analysis as a technique to assess prior learning by correlating documents in a learner portfolio with documents in target learning activities.**

I. Introduction

Prior learning experiences are important for learning. In some European Countries like the Netherlands or the UK the process of Accreditation of Prior Learning (APL) is a standard procedure to assess a student and allow exemptions for a study program [1]. The result of this process is an individualized curriculum. In a traditional APL procedure students apply for exemptions with a portfolio that is subsequently assessed by experts from the domain who then decide about exemptions. The drawback of this procedure is that is very time and cost consuming.

In the TENCompetence project we are aiming at the development of an infrastructure for lifelong competence development [2]. In this context we explore approaches to assess prior learning experiences and to offer individualized learning paths through a collection of learning activities in a learning network. Traditionally this problem has been addresses by adaptive hypermedia research on learner modeling [3]. But the solutions from learner modeling have several limitations: On the one hand they only work in one adaptive system that "learns" over time about the learner's preferences and behavior. On the other hand a static pre-designed model of a learner does not fit to the dynamics in lifelong learning.

To overcome the limitations of existing approaches we are developing content-based approach to prior learning assessment in learning networks.

II. Prior Learning Assessment in Learning Networks

Our project is based on the assumption that content can be taken as a proxy to estimate prior knowledge of a learner. The rational behind the project is discussed in [4]. The estimation of prior knowledge is calculated through the similarity of content in the learner's portfolio and the content that is connected to his/her target learning activities. To calculate this similarity we use Latent-Semantic-Analysis [5].

The results of such a similarity analysis is an ordered list of correlations between documents in the learner portfolio/profile and the target learning activities. High correlations with target activities may, depending on the policies of the learning environment entitle the learner to exemptions. Since the number of documents for solving this educational problem will be quite small compared to an information retrieval scenario, Van Bruggen et al. conducted an exploratory study into the usability of LSA in small scale corpora and reported promising results [6].

III. The case study

To test our model we collected data in an introductory psychology course at the Open University of the Netherlands. The online course consisted of 18 learning activities based on a textbook. Every chapter covers a subtopic of the psychology domain. We asked participants of this course in advance to comment on prior learning experiences that they considered relevant to the course. We invited them to substantiate this by, wherever possible, uploading files they had produced or read during their prior education. Since we could not expect students to know exactly what topics were presented in the chapters we also questioned them after completion of each chapter on the novelty of the presented material. We also constructed some additional cases to reach a sufficient variety of profiles. Latent Semantic Analysis was used to analyze this material and to calculate correlations between the learner documents and the target learning activities.

To evaluate these results we will use an expert validation. Domain experts will analyze the material and decide about exemptions under a strict exemptions policy and under a more lenient policy. Another measure we are interested in is the time that experts spend to come to a decision because one of the main reasons for our project is to make the APL procedure more efficient. The decisions and the time needed for analysis of the portfolios will be compared to LSA results.

IV. The corpus and the Software

The final corpus contains 800 documents selected from the

course book, other psychology books and Wikipedia articles from the Dutch Wikipedia. Textstat [7] reports 35742 words. The corpus was filtered using a modified Dutch newspaper stop list [8].

For the analysis we followed the optimization procedure described in [6] and decided to use 20 singular values for the analysis, corresponding with 90% of the variance being explained. Visual inspection ("Scree test") of the singular values revealed a steep drop in the size of the singular values as well. We compared the results of analyses using 10, 20 and 40 singular values and found that the analysis with 20 singular values resulted in a.) a sufficient discrimination between the chapters; b.) a high correlation between the chapter and the learner portfolio when there is sufficient thematic overlap and c.) a low correlation when there is no or only a little overlap. For the analysis we used the GTP application by Giles, Wo & Berry [9].

V. Preliminary results and outlook

The provisional results are encouraging: portfolios with 'popular psychology' content produced no match. A portfolio of a student who had already finished several psychology courses produced several matches for the subchapters of the book. On the other hand student portfolios with only prior knowledge for one of the chapters (e.g. the chapter about perception) showed only a high correlation to this specific chapter but low correlations for the other chapters.

The current results are limited and provisional in many ways. First, the results need to be validated against expert assessments, where the main question is whether LSA-based decisions are comparable to expert placement decisions. Here, as well as in essay rating, the reliability of expert judgments has to be taken into account. More interesting, however, is whether experts operate by matching documents. For example in one case, a technical description of an experiment, LSA returned no matches. A human expert is capable of inferring prior knowledge. Second, the current analyses are based on the assumption that a one-to-one match exists between a student document and a target document (here a chapter). A more realistic scenario would be that there are several partial matches between student documents and target documents. For example, a student paper that addresses one particular topic would partially match a target document that deals with other topics as well. The type of automatic topic recognition in combination with segmentation of the documents is beyond the scope of our current research.

In this part of the project we only focus on content analysis while we will widen the scope in the future also on the use of metadata and ontologies for prior knowledge assessment. The whole project plan is described in [10].

VI. Acknowledgement

The authors' efforts were partly funded by the European Commission in TENCompetence (IST-2004-02787).

References

[1] Koper, R. & Specht, M. (2007). TenCompetence: Lifelong Competence Development and Learning. In: Sicilia, M.-A. (Ed.). Competencies in Organizational E-Learning: Concepts and Tools. Idea Group.

[2] Merrifield, J., McIntyre, D., & Osaigbovo, R. (2000) Mapping APEL: Accreditation of Prior Experiential Learning in English Higher Education. London. Retrieved July 1, 2006 from http://www.dfes.gov.uk/dfee/heqe/let_final.pdf

[3] Aroyo, L. (Ed.) (2006). Learner Models for Web-based Personalised Adaptive Learning: Current Solutions and Open Issues. Deliverable 1.3. Prolearn Network of Excellence Professional Learning. Internal Document.

[4] Van Bruggen, J., Sloep, P., van Rosmalen, P., Brouns, F., Vogten, H., Koper, R., & Tattersall, C. (2004). Latent Semantic Analysis as a Tool for Learner Positioning in Learning Networks for Lifelong Learning, *British Journal of Educational Technology*, 35(6), 729-738.

[5] Deerwester, S., Dumais, S.T., Furnas, G. W., Landauer, T. & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.

[6] Van Bruggen, J., Rusman, E., Giesbers, B. & Koper, R. (2006). Content Based Positioning in Learning Networks. Proceedings of the Sixth International Conference on Advanced Learning Technologies. IEEE Computer Society. 366 – 368.

[7] Hüning, M. (2007.) TextSTAT – Simples Text Analyse Tool. Concordance software for Windows, GNU/Linux and MacOS X. FU Berlin.

[8] Bouma, G., & Klein, E. H. (2001). Volkskrant database nad practicum [Data file]. Retrieved May 12, 2005 from http://hagen.let.rug.nl/~hklein/Volkskrant/Volkskrant

[9] Giles, J., Wo, L., & Berry, M. (2003). GTP (general text parser) software for text mining. In H. Bozdogan (Ed.). Software for text mining, in statistical data mining and knowledge discovery (pp. 455-471). Boca Raton, FL: CRC Press.

[10] Kalz, M., van Bruggen, J., Rusman, E., Giesbers, B. & Koper, R. (2007). Positioning of Learners in Learning Networks with Content, Metadata and Ontologies. *Interactive Learning Environments.* Special Issue on Lifelong Competence Development.

Training of Summarisation Skills via the Use of Content-Based Feedback

W. Lenhard, H. Baier, J. Hoffmann, W. Schneider, A. Lenhard

***Abstract* — conText is an intelligent tutoring system in German language that has been developed to improve text comprehension and writing abilities. The system is based on the model of Summary Street [4] and features comparable feedback mechanisms. Two laboratory experiments demonstrate that conText improves the quality of text summaries written by undergraduate students. We will assess the applicability of conText in schools in the following three years.**

I. Introduction

Training of the ability to summarize texts is one of the most effective starting points for interventions aimed at fostering text comprehension [1]. Summarization not only results in a more active reading and a deeper level of processing. It also leads to a more integrated knowledge representation of a given text [2]. However, mere practise without supportive feedback is unlikely to produce a noticeable learning effect. Unfortunately, frequent and individualized feedback, albeit necessary, is hardly realisable under real life school conditions. In order not to pose additional work load on the teacher and to free resources for at-risk students, LSA-based tutoring systems may be used to assist the learning process.

In prior experiments in German language, LSA showed an equal performance compared to university students in classifying animal species and achieved medium to high correlations with human raters in essay scoring. Using added scores, the correlations reached values comparable to the reliability of standardized tests [3]. The correlation of the LSA-based scores with expert ratings varied between .6 and .8. Inter-rater correlation among expert was not significantly higher compared to the correlation between LSA and the experts. LSA therefore seems to be a valuable tool in the construction of intelligent tutoring systems not only in English, but also in German.

Following the work of Steinhart [2], [4] and the commercial application Summary Street (Pearson Knowledge Technologies), an intelligent tutoring system named "conText" is currently under development for the German language [5]. It is aimed for the application in school grade 5 to 8 and features feedback on orthography, plagiarised passages, redundant and irrelevant sentences, as well as detailed and overall content coverage. Moreover, and contrary to Summary Street, it is planned to include feedback on style and composition as well and incorporates an educational text selection module that chooses the adequate degree of text difficulty on the basis of the student's performance in preceding passes.

First experiments have been conducted to examine the effects of LSA-based feedback on the performance of students in writing summaries.

This work was supported by the German Research Foundation (DFG) under Grant HO 1301/11-2 and SCHN 315/29-1.

W. L, H. B., J. H., W. S. and A. L. are with the Department of Psychology, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany (W. L. phone: +49 931 312626; fax: +49 931 312763; e-mail: wolfgang.lenhard@mail.uni-wuerzburg.de; baier@psychologie.uni-wuerzburg.de; hoffmann@psychologie.uni-wuerzburg.de; schneider@psychologie.uni-wuerzburg.de; lenhard@psychologie.uni-wuerzburg.de).

II. Methods

We performed two experiments in which undergraduate students summarized an expository text. In experiments 1, there were two experimental groups: one that received LSA-generated content feedback, and one group that did not receive LSA-generated content feedback. In experiment 2, a third group that also received sentence analysis (feedback on potentially redundant and irrelevant sentences) was additionally examined.

A. Participants

In the first experiment, 20 students (10 male and 10 female) participated in the experiment. In experiment two, 52 students (12 male, 40 female) took part. The students were randomly assigned to the different experimental conditions and received course credit for their participation.

B. Apparatus and procedure

After receiving an instruction, the participants worked with a laboratory version of conText. The program first shows the source text, and a text processor component is displayed next. The students receive online-feedback on the length of the summary, the spelling and plagiarised passages. After a length-threshold has been met, an analysis on sentences is carried out where irrelevant and redundant sentences are flagged. Finally, the content coverage of the different passages of the source text, as well as a global rating are computed via LSA and displayed as vertical bar charts. The student may now stop or engage in another trial to further improve the draft. In experiment one, students worked on a text about earth quakes. In the second experiment, a text about moa (extinct flightless birds from New Zealand) was presented. Students had one hour time for the summarisation, but could

stop working whenever they wanted to do so.

C. Measures

After completion of the experiments, participants rated the perceived accuracy of the content feedback of conText. The summaries were then scored by independent raters with a predefined rating scheme. Furthermore, time on task and the number of trials were recorded.

D. LSA-Platform

The LSA-platform consists of a server written in Java which runs on a desktop computer (Pentium IV, 3.2 GHz, 3 GB RAM, SuSE Linux 9.3). The server is administrated via a web interface [3] and deals with all aspects concerning the corpora administration, generation and weighting of frequency matrices, singular values decomposition (SVD), generation of semantic spaces and calculation of text similarities. The abridged Lanczos algorithm [6] is used for decomposing the singular values.

There are several client applications communicating with the server via internet, for example a system for automatic essay scoring of student writings in university lectures and laboratory prototypes of conText.

We use specialised semantic spaces for the computation of text similarities. The space underlying the described experiments consists of texts from the domains biology, geography and geology from school books, encyclopaedias and internet pages. The texts were extracted and split into paragraphs automatically. We converted all words in the texts to lower case and filtered stop words, words occurring less than three times as well as texts consisting of less than ten different words. The frequency matrix included 37 773 paragraphs with 83 369 different words (total size of corpus 2 178 432 words). Prior to the SVD, a log-entropy weighting was applied to the frequency matrix. We extracted 400 dimensions (duration of computation: 35min 17 sec) [3]. Other semantic spaces for different knowledge domains are at hand as well. When doing similarity judgments, missing words are automatically lemmatized in case, the semantic space contains the lemma.

III. Results

Whereas all data assessed in experiment one could be used for analysis, a total of five participants were excluded from experiment two, either due to the lack of German language skills (N=3), or because technical errors occurred during the experiment (N=2).

The LSA-scores of the summaries significantly correlated with the human ratings, with r=.552, p< .05 in experiment 1, and r=.738, p <.001 in experiment 2.

In experiment 1, students who received LSA-feedback tended to receive better scores by human raters compared to students who did not receive feedback, $F(1, 18)=2.531$, p=.064. The effect size was d(Cohen)=.71. Comparing the quality of their first and their final draft, they showed a higher increase of content coverage in the course of writing, $F(1, 18)=4.514$, $p< .05$. This effect amounted to d(Cohen)=.95.

In experiment two, due to a lower quality of the first draft, we failed to obtain a better final content quality in the groups who worked under LSA-based feedback as compared to the control group. However, the experimental groups showed a higher gain of content quality during writing than the control group, $F(2, 42)=4.34$, $p <.01$, with an effect size of d(Cohen)=.98. The group receiving both content feedback as well as feedback on redundant and irrelevant sentences showed the highest increase in content coverage. The difference to the content feedback only group was not significant, however. When feedback was given, students worked longer, $F(2, 42)=3.71$, $p <.05$, and did more revisions, $F(2, 42)=64.02$, $p < .001$. Moreover, the participants' ratings of the feedback quality given by conText increased with detailedness, $\chi^2(2)=4.60$, p=.05. This finding underlines the usefulness of the sentence analysis in terms of user acceptance of the learning environment.

IV. Conclusion

LSA-generated feedback scores showed medium to high correlations with human judgments. Students, who work under LSA-feedback tend to write better summaries and show higher increases in the quality of their summary during writing. The obtained effects are encouraging, especially when keeping in mind that feedback in well-designed studies on average yields an effect size of .46 [7].

The next step will be to assess the effects in schools with first experiments in the summer of 2007, and to create a robust learning environment. In order to measure the long-term effects on reading comprehension, longitudinal studies will take place between autumn 2007 and 2010.

References

[1] E. Souvignier and F. Antoniou, "Förderung des Leseverständnisses bei Schülerinnen und Schülern mit Lernschwierigkeiten - eine Metaanalyse," *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, vol. 1, pp. 46-62, January 2007.

[2] D. Wade-Stein and E. Kintsch, "Summary Street: Interactive computer support for writing," *Cognition and Instruction*, 22, 333-362, 2004.

[3] W. Lenhard, H. Baier, W. Schneider and J. Hoffmann, "Automatische Bewertung offener Antworten," Diagnostica, in press.

[4] D. Steinhart, "Summary Street: an Intelligent Tutoring System for Improving Student Writing through the Use of Latent Semantic Analysis," Ph.D. dissertation, Institut for Cognitive Science, University of Colorado, Boulder, CO, 2001.

[5] W. Lenhard, H. Baier, W. Schneider and J. Hoffmann. (2006). conText: Fostering text comprehension by working with texts [Online]. Available: http://www.summa.psychologie.uni-wuerzburg.de

[6] C. Lanczos, "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators," [Online]. Available: http://www.cs.duke.edu/courses/fall06/cps258/references/Krylov-space/Lanczos-original.pdf, date of retrieval: 10.01.2007), *Journal of Research of the National Bureau of Standards*, 48, 255-282, 1950.

[7] R. L. Bangert-Drowns, C. L. Kulik, J. A. Kulik and M Morgan, „The instructional Effect of Feedback in Test-Like Events," *Review of Educational Research, vol. 61, 213-238, 1991.*