Applying natural language processing to text analysis in educational context

Tuomo Kakkonen, Niko Myller, Patrik Scheinin, Erkki Sutinen, Jari Timonen

Introduction

In this paper, we use the term academic text to refer to any free text composed in an academic setting, covering the whole spectrum from first year students' reviews of available scientific texts up to a scientist's struggling, but still fragile text concerning his or her new discoveries. These academic texts can be processed by various information technologies. We can now outline the landscape of applying technologies for assessing academic texts as a cube (see Figure 1) where a particular application can be identified as its position indicated by the three dimensions of the purpose of assessment, the kind of feedback, and the type of contribution.

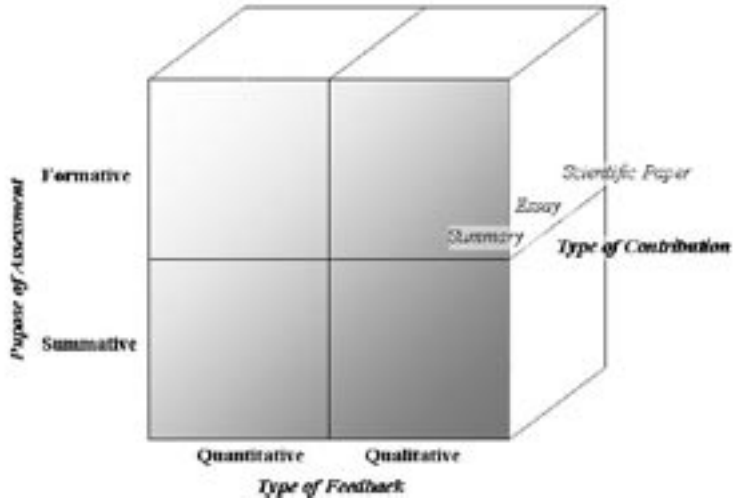

Figure 1 *Dimensions of the analysis of scientific texts.*

As illustrated in the Figure 1, a particular academic text can be assessed at different stages of its compilation process, the two extremes being the very beginning and after the process. Although the dimension is continuous, we can identify the main purpose of assessment at its two ends as formative and summative. Formative assessment aims at supporting the writing process, whereas summative assessment gives feedback after the process has ended. Feedback can be primarily quantitative (scoring, grading, marking, profiling) or qualitative (descriptive), but combinations are also possible. The type of an individual text's contribution varies from a summary up to an independent scientific text.

To give an example of the idea of using the categorization just introduced, let's consider a student's essay that needs to be assessed in an examination. The purpose of the assessment is summative, feedback quantitative, and writer's own contribution is considered to be quite limited. Second example is a scientist's original paper that needs to be analyzed as a part of her departmental evaluation. Purpose of the evaluation is summative, feedback between quantitative and qualitative, and contribution highly original.

The trend is to move from summative, quantitative assessments towards formative qualitative ones. While the categorization introduced above describes the opportunities of applying technologies to text assessment or the what-map, we still need another scheme for analyzing the available technologies, i.e., the how-map.

Methods for analyzing scientific texts can be divided into comparison-based methods and ones that rely on analyzing the texts not by comparing them to some source, but more to analyze their properties like structure, cohesion, and use of words. While methods based on text comparison are feasible for quantitative evaluation, like assigning scores to student's writings, they are less suitable for other types of analyses. Methods such as discourse structure analysis, on the other hand, are more suitable for giving detailed feedback to writers and for descriptive text analysis, but they can also be used to aid the summative and quantitative type of assessment.

On the other hand, methods can be divided into ones concentrating on surface features, to the content, or to the structure of the text. The earliest essay grading systems were solely based surface features such as the length of the essay in words, the number of commas and so on. More recent and advanced systems have moved into measuring the content and the structure of the texts. In this paper, we will consider the two latter types of methods.

Text visualization is one of the available methods for text assessment (Card et al. 1999). It is primarily useful for illustrating the structural features of a text, such as in indicating the different types of texts (descriptive, argumentative etc.) with colors. It can also represent the whole text as nodes, with edges in between for internal dependencies.

Comparing the what-map with the how-map reveals that there is no one-to-one relationship between the challenges and the methods, but a particular technique can be applied to instances at several different positions in the cube.

In this paper, we discuss possibilities offered by current information retrieval (IR), natural language processing (NLP) and machine learning (ML) methods to analyses of academic texts. In chapter 2 we describe our essay grading system based on comparison between the course materials and essays and discuss some research results on other similar types of approaches. Chapter 3 discusses the possibilities of using statistical and discourse analysis methods for more fine-grained text analysis, supporting automated feedback and other semi-automatic assessment features. Finally, in chapter 4, we discuss the findings and outline some future opportunities.

Automated essay grading

Automated grading using Latent Semantic Analysis

AEA (*Automaattinen esseiden arvioija*) developed at the University of Joensuu is a computer based application which can assess essay answers written in the Finnish language (Kakkonen & Sutinen, 2004). AEA is based on an IR method called *Latent Semantic Analysis* (LSA). LSA is an implementation which models human knowledge representations into a computer. It searches for conceptual similarities of words and passages (Foltz, Laham & Landauer 1999). Discovery of the content similarity between two documents is one of the strengths of this method. One interesting feature of LSA is the ability to recognize similarities between two documents although they contain only few words which are exactly the same in both documents (Deerwester et al. 1990).

LSA has been developed since the beginning of the 1990's. Laham, Landauer and Kintsch are researchers that developed LSA and have later applied it to automatic essay grading for essays written in English. They have developed an Intelligent Essay Assessor (IEA) which is based on LSA (Foltz, Laham & Landauer 1999). IEA's performance has been tested with diverse materials, varied from psychology and history to biology based essay answers and book material (Foltz, Laham & Landauer 1999). One of the IEA's test materials consisted of 1205 essays from 12 different topics (Foltz, Laham & Landauer 1999). In that research, they compared the correlation between the grades given by two trained human graders to the grades given by their LSA based application. The main result of that research was that the correlation between two human graders was 0.707 and between human grader and LSA almost the same (0.701). Figure 2 also shows the main result of the research by Foltz, Laham & Landauer (1999). In our research at the University of Joensuu, the

correlation between a human grader and the grade given by LSA based AEA reached 0.8 (Kakkonen & Sutinen, 2004). This research result indicates that AEA produces similar grades compared to human graders.

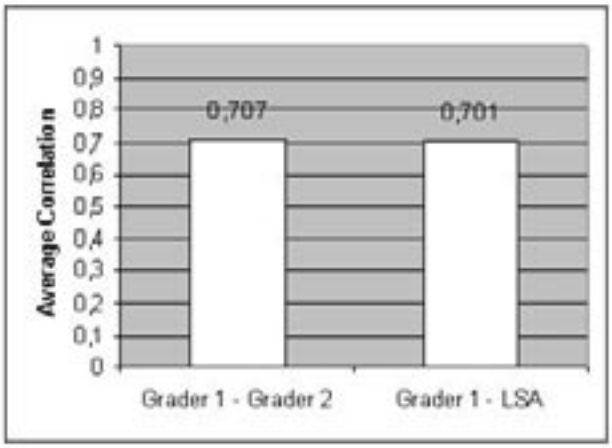

Figure 2 Correlation between two human grader versus the correlation between grade given by human and LSA (Foltz, Laham & Landauer 1999).

Automated selection of LSA dimensionality

LSA processes a large corpus of text (in this case essays and text book materials). After processing LSA creates a representation of sentences and paragraphs which are taken from the original corpus. This representation, called “semantic space”, is very high dimensional (e.g. 50-1500). Reducing the dimensionality is very important stage which transforms the surface information into deeper abstraction. Dimensionality can also be described as the number of parameters by which the original corpus is described. The idea of the dimension reduction step is to find small but still large-enough number of dimensions which produces good approximations compared to the human cognitive relations and models. (Landauer et al., 1998)

In addition to dimension reduction step, there are also other weaknesses in LSA. First of all, it can be only used for an analysis of the contents of the essays and it does not take the word order into consideration. Nevertheless, there are a number of research results indicating that passage and content meaning can be derived without using word order (Landauer et al., 1997). As described before, dimension reduction is one vital stage inside LSA. In automated assessment, it means that there is certain number of dimensions in LSA which produces the most accurate grades compared to the grades given by a human grader. One problem is how to find this dimension. As a solution for the problem, a training phase is implemented in AEA. Before the actual grading process, the AEA system must be trained with essays graded by human graders. We applied different machine learning methods to implement the training phase. These data validation methods are *k-fold cross-validation*, *holdout* and *bootstrap* (Witten & Frank, 2000). Descriptions of the k-fold cross-validation and holdout can be found from (Witten & Frank, 2000) and bootstrap is best described by (Efron & Tibshirani, 1993). Following three paragraphs describes shortly these three methods and their implementation in AEA. Figure 3 illustrates the general principle of dimension selection with the data validation method as a part of the process.

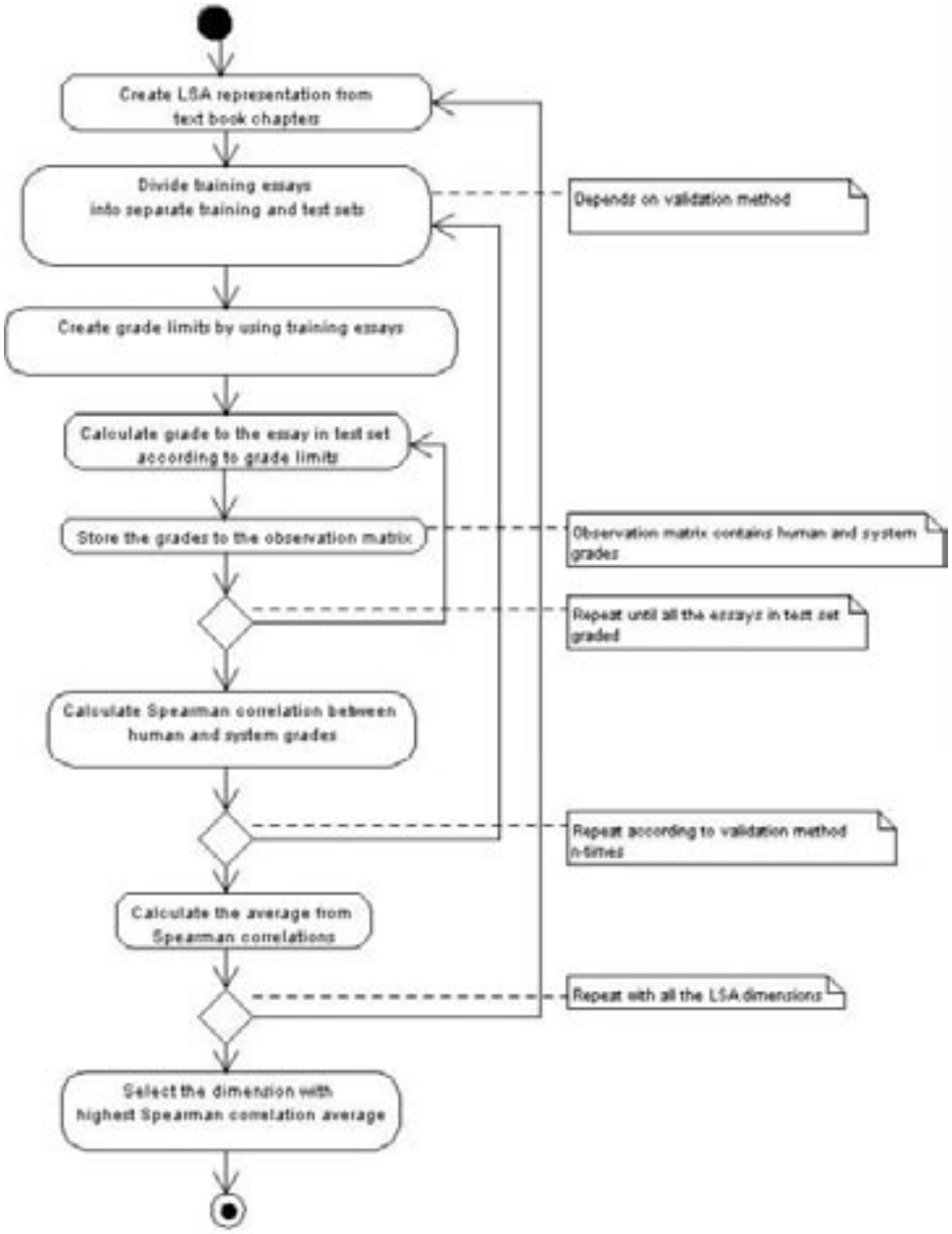

Figure 3 UML state diagram from dimension selection phase inside AEA.

Holdout method implemented into AEA divides the training essays to separate training and testing sets. Training set contains 2/3 of the total amount of training essays and the rest 1/3 is placed to test set. The sets can also be stratified, meaning that the same proportion of essays is placed on each grade category on the training and testing sets, as there was in the original set of training essays. After the division, the grade limits for each grade category are computed. Once the grade categories are defined, AEA gives the grades to the essays in test set as described in Kakkonen & Sutinen (2004). Because all the training essays contained human grades, the grades given by the system and the human grader can be compared. This is done by calculating the Spearman correlation. Because the original problem was, what dimension to use, this division and grading process must be repeat by using different LSA dimensions. As a result, we have Spearman correlation for each LSA-dimension. We select and use in the actual grading the LSA dimension, which gave the highest Spearman correlation.

K-fold cross-validation randomly splits the training essays into *k* essay sets of same size. Division can also be stratified which means that the resulting *k* essay sets contain as many essays from different grade categories as the whole training essay set. When we have *k* separate essay sets, AEA

goes through them using one of the sets a testing set and remaining *k-1* sets as training set. Figure 4 illustrates the essay division of 3-fold cross-validation. Otherwise the training phase goes like in Holdout method described in the above paragraph and in the Figure 3.

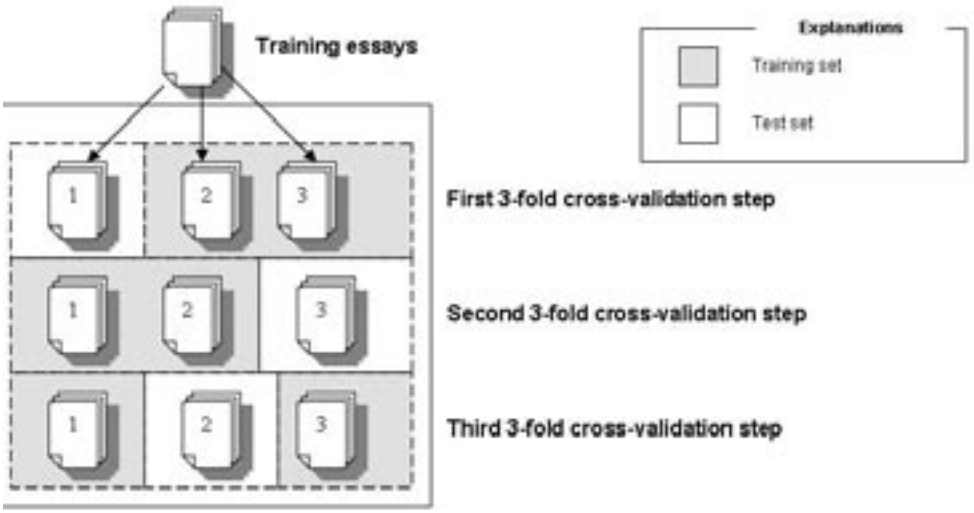

Figure 4 Different division of 3-fold cross-validation into training and testing sets.

Bootstrap is a method for statistical analysis which is based on experimental data (Efron & Tibshirani 1993). In AEA, the data consists of graded training essays. Main difference between the cross-validation based methods and Bootstrap is that the Bootstrap uses replacement. It means that one element (in this case essay) already selected into Bootstrap training set, can be selected again. In cross-validation based methods (such as k-fold cross-validation and holdout), same essay can only occur once in a training set. Bootstrap creates a training essay set by using replacement. If size of all the training essays is *n,* the bootstrap draws randomly out *n* essays. Those essays in the original essay set which were not selected are used for testing. After creating Bootstrap training and testing samples training phase in AEA goes like in figure 3.

As described in figure 3 and above, the machine learning methods extracts the best LSA dimension based on the training essays and course materials, e.g. passages from the course textbooks. It can be said that the machine learning methods train AEA to use the best dimension by using human graded essays. After the dimension search, the essays without human grade can be assessed. Currently, dimension research is in progress and the results will appear later. As a summary we can present the research goals for this ongoing research as follows.

- Is it possible to search the model (dimension) needed by LSA automatically when applying it to automatic essay assessment?
- What is the accuracy between the optimal dimension and the dimension found in training phase when comparing the human and system grades together?
- What are the accuracies of different data validation methods?

Text analysis-based methods

In addition to currently ongoing research on the selection of dimensionality, there are also other challenging research issues related to the essay assessment system. Semi-automatic assessment features alongside AEA are the next research step to be taken. They mean different kind of feedback features added to AEA application. Together with grade the feedback enables not only the students but also teachers go towards constructive learning process (Kakkonen, Myller & Sutinen 2004). Automatic feedback and summary generation from the essay given to the students as well as analysis of the structure, coherence and cohesion of the essay text as a help for a teacher's grading process, are examples of the future planning. More detailed description of the semi-automatic features can be found from Kakkonen, Myller & Sutinen (2004). In the rest of the chapter, we introduce some current

methods and ongoing research that could be applied for analysis for texts of higher scientific contribution as well as for giving qualitative feedback, but that also can be applied to aid essay scoring and other quantitative type of evaluation.

Cohesion identification

Text is *cohesive* when it is continuous and the transitions between different sentences or paragraphs are smooth. This means that sentences share some amount of semantic similarity and the paragraphs are linked together. If text is cohesive it is not necessarily coherent. *Coherence* means that the adjacent sentences or paragraphs are logically related to each other. This is harder to be detected automatically. For example, next sentences, *"Jack is tall. He is happy."*, are cohesive but not coherent, because even though they share the same subject, the relationship between the sentences is not logical. However, if the first sentence is changed to, *"Jack got a job"*, the sentences are both cohesive and coherent.

The cohesion and coherence of the text are important measures of the writing style (Shermis et al. 2001). Foltz, Kitsch and Landauer (1998) used LSA to determine the cohesion between two adjacent sentences or paragraphs. This could be applied as a simple measure of cohesion because LSA measures how much the compared two sentences or paragraphs have in common at the semantic level. This means that if a text is cohesive two adjacent sentences or paragraphs share something in common in semantic level and thus LSA should be able to identify them as similar. In the case of abrupt shift in a topic the cohesion is not retained and the semantic structure changes and LSA will identify that with no correlation between paragraphs. This will also indicate some of the coherence problems but some of the cases cannot be detected automatically and human intervention is needed.

Word Co-occurrences

When comparing texts from different disciplines, it is also meaningful to look for words that have similar or different meanings between disciplines. Computers cannot automatically determine the meanings of the words in the context but there are ways that can give some information about the senses of the words. For example, word sense is manifested in different texts by the co-occurring words. If we have large corpus of text from the different disciplines, we can compute the word co-occurrences and have a distribution of distinct words co-occurring with it (Lee 2001). These distributions can be compared across the disciplines and this can give insights into how the meaning of the words is similar or different between the disciplines. Similar technique has been used in word sense disambiguation. Other possibility is to find words which have high distributional similarity inside the texts of the disciplines and then compare these groups of words to each other.

Analyzing Discourse Structure

A possible method to be used for examining scientific papers is that of discourse structure analysis (Allen, 1995). Discourse analysis here refers to study of the organization of written texts in larger units, above the sentence and clause levels. A discourse can be divided into segments, pieces of discourse in which the sentences are addressing the same topic, thus displaying coherence. Two main problems can be identified from the analysis. First, the sentences within the sequence must be analyzed with some method and second, the relationships between the segments must be identified. A certain type of phrases and words, called *cue phrases*, can be used for finding topic changes in a discourse. Cue phrases can be divided into two classes: the other is used for finding semantic relationships and the other for detecting discourse structure without identifying a semantic relationship. For example, words such as 'anyway', 'by the way', and 'first' can be used as structure cue phrases,

and sentence connectives such as 'because', 'but', and 'however' can be applied for detecting semantic relations.

Burstein & Marcu et al. (2000 & 2001) have applied analysis of discourse structure in educational context by a system that identifies a *thesis statement*, the sentence that previews the main idea of the essay, and the connections between the thesis statement and the main points of the essay. The system uses a set of essays in which the thesis statement has manually been marked and applies a Bayesian classifier to automatically detect the thesis statements based on the training. The identification is based on the position of the sentences in the essays, words commonly used in thesis statements, and most interestingly on our perspective, on Rhetorical Structure Theory (RST) parses.

RST was originally developed by Mann & Thompson (1988) in the 1980's. Mann & Thompson did not introduce an algorithm for automatic construction of RS-trees. Such a method was later developed by Marcu (1997). The theory models aspects of the organization of natural text, and aims to identifying the relations between the parts of the text. Figure 5 shows an example of Rhetorical Structure tree (Burstein et al. 2001). As shown in the figure, the text is first divided into text spans. The relationships between the spans are represented as a tree structure, with the text spans as the leaf nodes. The spans are classified as Nucleus, the most important part of the sentences or as Satellites with less importance. The identification of text spans is based on cue phrases.

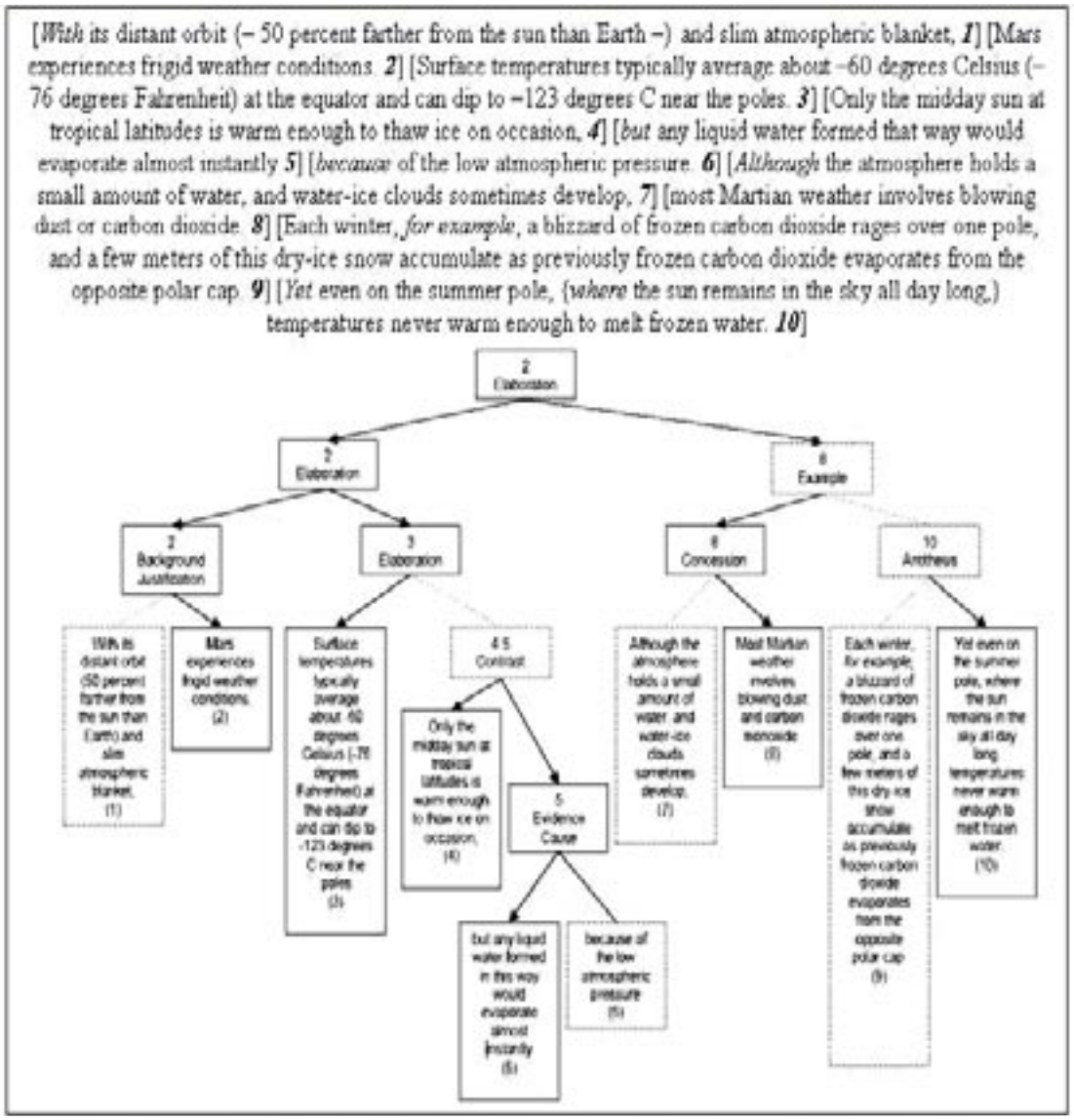

Figure 5 Example of RST tree (Marcu, 1997).

Conclusion

Automated essay scoring systems have produced good results, not only when applied to essays written in English but also the Finnish ones. LSA is one feasible method for such analysis. AEA system developed at the University of Joensuu is a combination of automated essay scoring and research work over a decade behind the LSA. Some issues, including the automated selection of dimensionality discussed in this paper, have to be solved. Our approach towards a solution is to apply some existing machine learning methods. Research results in near future will show how well the dimensionality can be found automatically.

Coherence and cohesion measures can be used to analyze diverse types of texts. These measures tell about the authors' ability to retain smooth transitions in the text and stay in the subject. Moreover, it has been shown that the automatic assessment of the essays became more accurate when cohesion measures were added. Word co-occurrences can be used to get some intuition how words are similar or different inside one discipline. This technique has been used to find semantically similar words in large corpora. Problem with this technique is that one word can appear in different senses inside the corpus and thus the material should be first annotated with the senses.

Research on discourse analysis has shown that indeed differences in different types on discourse can be identified. An interesting extension of this work would be to apply such an analysis to identify differences in the structure of scientific papers from different disciplines and to distinguish differences in structures between texts classified by humans as of high or low quality. Concerning applying the discourse analysis to texts written in Finnish, a list of cue words should be defined first.

REFERENCES

Allen, J. 1995. Natural Language Understanding (2nd Edition). California:The Benjamin/Cummings Publishing Company.

Burstein, J., Marcu, D. 2000. Towards Using Text Summarization for Essay-based Feedback. In Proceedings of the Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000. Lausanne (Switzerland).

Burstein, J., Marcu, D., Andreyev, S., Chodorow, M. 2001. Towards Automatic Classification of Discourse Elements in Essays. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse (France).

Card, S. K., Mackinlay, J. D., Shneiderman, B. 1999. Readings in Information Visualization: Using Vision to Think. Academic Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. 1990. Indexing By Latent Semantic Analysis. Journal of the American Society For Information Science, 41(6), 391-407.

Efron, B., Tibshirani R., J. 1993. An Introduction to the Bootstrap. Chapman and Hall.

Foltz, P. W., Kintsch, W., Landauer, T. K. 1998. Analysis of Text Coherence Using Latent Semantic Analysis, Discourse Processes, 25(2-3), 285-307.

Foltz, P. W., Laham, D., Landauer, T. K. 1999. The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2).

Kakkonen, T., Sutinen, E. 2004. Automatic Assessment of the Content of Essays Based on Course Materials. Proceedings of the 2nd International Conference on Information Technology: Research and Education (ITRE), London:London Metropolitan University, 126–130.

Kakkonen, T., Myller, N., Sutinen, E. 2004. Semi-Automatic Evaluation Features In Computer-Assisted Essay Assessment. Proceedings of Computers and Advanced Technology In Education (CATE), Anaheim:ACTA Press, 456-461.

Landauer, T. K, Foltz, P. W., Laham, D. 1998. An introduction to Latent Semantic Analysis. Discourse Processes, 25(2-3), 259-284.

Landauer, T. K., Laham, D., Rehder, B., Schreiner, M. E. 1997. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans. Proceedings of the 19th annual meeting of the Cognitive Science Society. New Jersey, Mahwah: Erlbaum, 412-417.

Lee, L. 2001. On the Effectiveness of the Skew Divergence for Statistical Language Analysis, Artificial Intelligence and Statistics 2001, 65-72.

Mann, W. C., Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8(3), 243-281.

Marcu, D. 1997. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts. PhD Thesis, Toronto:University of Toronto, Department of Computer Science.

Shermis, M. D., Mzumara, H. R., Olson, J., Harrington, S. 2001. On-Line Grading of Student Essays: PEG goes on the World Wide Web, Assessment & Evaluation in Higher Education, 26(3), 247-259.

Witten, I. H., Frank, E. 2000. Data mining: Practical machine learning tools and techniques with Java implementations. San Diego:Academic Press.