O'Rourke, S. T., Calvo, R. A., & McNamara, D. S. (2011). Visualizing Topic Flow in Students' Essays. *Educational Technology & Society*, *14* (3), 4–15.

Visualizing Topic Flow in Students' Essays

Stephen T. O'Rourke, Rafael A. Calvo and Danielle S. McNamara[1]
University of Sydney, Australia // [1]University of Memphis, USA // stephen.orourke@sydney.edu.au // rafael.calvo@sydney.edu.au // dsmcnamara1@gmail.com

ABSTRACT

Visualizing how the parts of a document relate to each other and producing automatically generated quality measures that people can understand are means that writers can use to improve the quality of their compositions. This paper presents a novel document visualization technique and a measure of quality based on the average semantic distance between parts of a document. We show how the visualization helps tutors mark essays more efficiently and reliably, and how a distance index calculated for the visualizations correlates with grades. The technique is further evaluated using three dimensionality reduction techniques. The results provide evidence that the degree of topic flow between consecutive sentences and paragraphs is related to essay quality.

Keywords

Document visualizations, writing, argument flow

Introduction

Writing is an important learning activity common at all educational levels and disciplines. But incorporating writing activities into the curricula faces many challenges, including the cost of providing meaningful and timely feedback as the one provided in assessment. Technically researchers are tackling these challenges by producing automated feedback (including grades) directly targeted to the students or support for human assessors who then write the feedback. In this line of research studies have shown tools and techniques for automated feedback in academic writing (Beals, 1998; Graesser & McNamara, in press; Thiesmeyer & Thiesmeyer, 1990; Wade-Stein & Kintsch, 2004; Wiemer-Hastings & Graesser, 2000). Feedback can be genre specific, as for example, in argumentative writing. Many studies in this area have focused on argument visualization (Kirschner, Shum, & Carr, 2003) where the students are visually presented with the way in which claims are interrelated, showing evidential structures. Other forms of feedback focus on quality measures that apply to multiple genres and disciplines. For example, automatically generated visualizations can be used as support material that students can use to reflect on a set of trigger questions designed around issues with which students normally have difficulty (Calvo & Ellis, 2010). Thanks to new cloud computing technologies these computationally intensive forms of support can be provided in real-time, at any stage of the writing process and to large numbers of students (Calvo, O'Rourke, Jones, Yacef, & Reimann, 2011).

Many different features of a document can be quantified and therefore represented visually as a form of feedback or support to handling writing activities. The challenge is to find visual representations of features that are both meaningful to actual writing tasks (e.g. putting together evidence into an argument) rather than mathematical artifacts, and that do actually provide useful information that correlates with the quality of writing. Several linguistic features of quality writing have been identified with evidence that they can predict high and low proficiency essays. For example, McNamara (2010) provided evidence for syntactic complexity (measured by the number of words before the main verb), lexical diversity, and word frequency (as measured by Celex, logarithm of all words). While other measures, such as cohesion indices, have received much attention in the literature, they generally offer more of a guide to a text's connectedness, which does not necessarily correlate well to that of experts (Crossley & McNamara, 2010). The way an argument is structured and the flow in a composition are important quality features. We evaluate here different techniques for producing topic flow visualizations. These were proposed by O'Rourke and Calvo (2009b) as a way of helping students reflect about their writing, particularly issues related to the flow in the composition. The visualization techniques include several steps, all of which must be taken into account if the actual visual representation (the final outcome) is to be semantically valid and useful.

In most Natural Language Processing techniques, each text segment is converted into a high dimensional representation using the Vector Space Model. This high dimensional space is then reduced by using one of several mathematical techniques that preserve as much of the original information as possible. The first challenge is to find the optimum dimensionality reduction technique that produces meaningful visualization. The optimum value is the measure of how it relates to a writing quality attribute (e.g. flow). This reduced space can then be used to produce a

2-dimensional space that can be made into a visual representation. In our topic flow visualizations, textual units (e.g. paragraphs) are represented in a two dimensional space, with their distances being proportional to the distance between topics (in the reduced dimensionality space). This visualization can provide the means to see how the parts of a document interrelate and follow each other. Once the vector data is in lower dimensional representation and can be represented visually we need to understand how the visualization (produced using optimum dimensionality reduction techniques) will improve an aspect of a learning activity such as reading and assessment. This is the second important question addressed in this paper

The new techniques for document visualization and for quantifying topic flow are described in the next section. First the mathematical framework to represent text segments as vectors (i.e. the vector space model), and then the combination of the three dimensionality reduction techniques and Multidimensional Scaling, used to bring these representations to a 2-dimensional visualization. Since the algorithms need to be validated as producing data that represents the actual semantics of the documents we show how they can be used to quantify topic flow, in a collection of real documents assessed by experts. We then explore the value of the actual visual representations by measuring the impact they have on tutors assessing a collection of essays. This evaluation is made on a new corpora of student essays and assesses to what extent semantic features can be captured using these techniques.

Visualizing Topic Flow

The Mathematical Framework

Documents can be represented in high dimensional spaces (e.g. all its possible topics or terms) but their visualization can only be done in 2 or 3 dimensions. It is therefore essential to choose the dimensions that are meaningful representations of quality features found in the text. Textual data mining approaches have been applied to analyze the semantic structure of texts (Graesser, McNamara, Louwerse, & Cai, 2004) and to build automatic feedback tools (Villalon, Kearney, Calvo, & Reimann, 2008; Wiemer-Hastings & Graesser, 2000). Such approaches typically involve performing computations on a text's parts to uncover latent information that is not directly visible in the surface structure of text. These approaches have been applied to solve a number of problems in educational settings, including automatic summarization, automatic assessment, automatic tutoring and plagiarism detection (Dessus, 2009). These applications often use measures of semantic similarity to compare text units. The most common approach involves the creation of a term-by-document matrix; derived from frequency vectors of distinct terms in each document to create a topic model, which describes the mixture of different topics in the document corpus.

The most representative dimensionality reduction techniques (a.k.a. topic models) are Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), Non-negative Matrix Factorization (NMF) (Lee & Seung, 1999), and Probabilistic LSA (PLSA) (Hofmann, 2001). These dimensionality reduction techniques produce topic models, where individual terms in the term-by-document matrix are weighted according to their significance in the topic (McNamara, 2010). The aim of dimensionality reduction is to eliminate unimportant details in the data and to allow the latent underlying semantic structure of the topics to become more evident. These topic modeling methods do not take into consideration many important components of a text, such as word order, syntax, morphology, or other features typically associated with text semantics, such as linking words and anaphora. Nevertheless, the topic models built using these methods from many source documents have long been shown to be reliable in the domain of information retrieval and highly similar to topic models extracted by humans (c.f.p. Landauer, McNamara, Dennis, & Kintsch, 2007). Similarly, automatic essay scoring systems have also been shown to have agreement comparable to that of human assessors with the use of pre-scored essays (Wang & Brown, 2007) and course materials (Kakkonen & Sutinen, 2004) to build the mathematical models.

While the above-mentioned topic modeling techniques have been successfully used on a large scale, they can be somewhat problematic in the case of providing feedback on a single essay. The corpus used to create a topic model is essentially a set of baseline texts, which are used to define the topic semantics against which a document will be compared. Thus, the distances between the term vectors in a semantic space of a topic model is entirely dependent on the corpus upon which it is built. A document is compared to how it relates to the tendency words to co-occur in the source documents, which could bias its meaning.

Another approach is to create the topic model from only the information contained in the document itself. This would maximize the variance of the model, while focusing on the semantics of the actual document itself, rather than relating it to something external. The idea of constructing a topic model from a single document was first proposed by Gong and Liu (2001) for the purposes of automatic text summarization. The single document semantic space technique has also been used by other researchers for automatic summarization (Steinberger, Poesio, Kabadjov, & Jeek, 2007) and discovering labels for clustered results in information retrieval (Osinski, 2006). More recently, Villalon and Calvo (2009) showed that the rank order of document distances in a topic model created from a single document was similar to one created from many source documents.

The visualization technique introduced here involves performing the following steps. First, a term-by-sentence matrix is prepared, after stop-words and low frequency words are removed, and stemming is applied. Second, a topic model is created using the NMF dimensionality reduction technique. Third, the topic model is projected to a 2-dimensional space using Multidimensional Scaling, and finally, a visualization of the document's paragraphs is produced. For brevity, the details of the three techniques used are left to their original references.

The elements of the initial term-by-paragraph matrix can be weighted using a number of schemes (e.g. Chi-squared, Log-entropy, TF-IDF) (Baeza-Yates & Neto, 1999). Log-Entropy, used in this paper, weighs a term by the log of its frequency tf_{ij} in a paragraph offset by the inverse of the entropy of its frequency across all n paragraphs in a document. The formula for calculating the Log-entropy weight of a term entry is defined in equation (1). Log-Entropy provides a useful weighting scheme for our purposes because it assigns higher weights to terms that appear fewer times in a smaller number of paragraphs. Thus, emphasizing the importance of infrequent terms in the paragraphs while also eliminating the 'noise' of frequent terms.

$$x_{ij} = \frac{\log\left(1 + tf_{ij}\right)}{-\sum_{k=1}^{n}\left(\frac{tf_{ik}}{\sum_{l=1}^{n} tf_{il}}\right)\log\left(\frac{tf_{ik}}{\sum_{l=1}^{n} tf_{il}}\right)} \tag{1}$$

The NMF dimensionality reduction technique generates its topic model by decomposing X into the product of two k-rank non-negative matrices, W and H, so that X is approximately equal to $X \approx WH$. In our case, k is considered to be the number of latent topics in a document. This makes the choice of k entirely document dependent. Given that k represents the number of latent topics in a document, W becomes a term-by-topic matrix, indicating the weighting of each term in a topic, and H becomes a topic-by-document matrix, indicating the weight of each topic in a document. The product WH is called a nonnegative matrix factorization of X which can be approximated by minimizing the squared error of the Frobenius norm (Meyer, 2000) of $X - WH$. Finding this solution defines the NMF problem which can be mathematically expressed as:

$$F(W,H) = \left\| X - WH \right\|_F^2 \tag{2}$$

The distance between any two paragraphs (not only the consecutive ones) can be calculated in the reduced topic representation using standard measures (cosine, Euclidean, etc.). Multidimensional Scaling uses this paragraph-paragraph triangular distance table to produce a 2-dimensional representation (Borg & Groenen, 2005). The Multidimensional Scaling transformation is performed using a procedure called iterative majorization (de Leeuw, 1977). The iterative majorization algorithm undertakes a least-squares approach to Multidimensional Scaling by attempting to minimize a loss function called Stress. Stress (Equation 3) can be expressed as the normalized sum of the squared errors between the vector dissimilarities $\hat{d}_{ij}$ and their approximated distances d_{ij} in the low dimensional space. The final result is a least-squares representation of the paragraphs described in the distance matrix, with the directions of the axes being arbitrary.

$$\sigma = \sum_{i<j} \frac{\left(d_{ij} - \hat{d}_{ij}\right)^2}{\hat{d}_{ij}^2} \tag{3}$$

Visualizing Flow

A well structured essay should have a clear and logical flow of ideas represented through its flow in paragraphs. The 2-dimensional visual representation can be used by students to reflect on the correctness of each argument (i.e. how each argument point follows each other). The visualization may also help detect 'breaks' in the flow, which is when two consecutive paragraphs talk about very different topics. The visualization presents the document in a different way, as it might appear for an external reader. In a paragraph 'map' such as the one in Figure 1, the essay's paragraphs are plotted on a circular grid with the diameter of the grid equal to the maximum possible distance between any two paragraphs (i.e. no topic overlap). The paragraphs are represented using a node-link diagram with text labels and arrows used to indicate the paragraph sequence.

For example, the clear sequence of topics in the five paragraph essay paradigm (Davis & Liss, 2006), can be visualized in our map. In this particular genre, the content of the 'introduction' and 'conclusion' paragraphs is expected to be similar, so these paragraphs should appear close in a map. The 'body' paragraphs address different subtopics and should ideally be linked through transitions so they should be sequentially positioned in the map. The map of a well structured ideal five paragraph essay would have a circular layout of sequential paragraphs, indicating a natural change in topic over the essay, with the introduction and conclusion starting and finishing on similar points. In contrast, we would expect a poorly structured essay to have many rough shifts in topic, with paragraphs positioned somewhat randomly around the map.

The visualization should also make evident the difference between a lower and a higher quality essay. Figure 1 illustrates the paragraph maps of two short essays. The essay on the left was given a low grade while the essay on the right was given a high grade. The topic flow of the high grade essays clearly resembles that of the prototypical five paragraph essay described above, while topic flow of the low grade essay appears disorganized. This qualitative evaluation implies that how much does statistically/geometrically the map/path deviate from a circle indicates the quality of its structure.

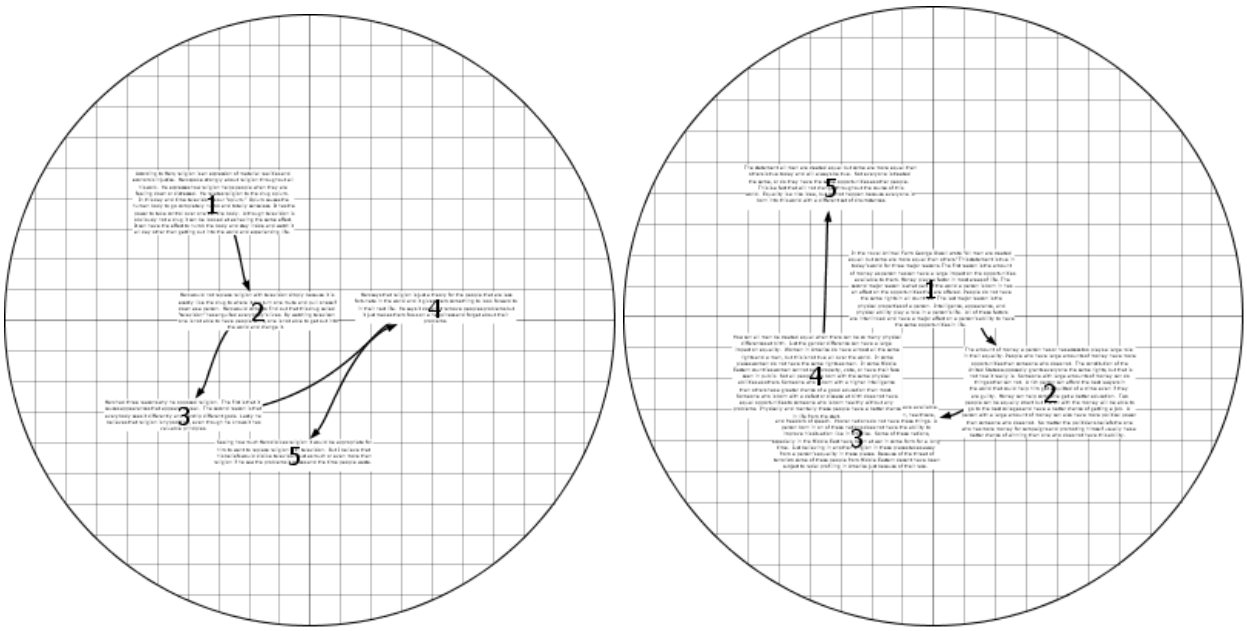

Figure 1. The paragraph maps of an essay with a low grade (left) and an essay with a high grade (right)

Quantifying Topic Flow

The topic flow analysis approach uses text mining techniques to model the topic mixture of a document's sentence and paragraphs, followed by document similarity comparisons to interpret the text's structure and flow. The automated approach involves performing the following steps:

- First, a term-by-sentence matrix is prepared, after stop-words and low frequency words are removed, and stemming is applied.
- Second, a topic model is created using dimensionality reduction techniques (NMF, SVD, PLSA).
- Finally, similarity comparisons are performed to identify features in the topic model of the document.

The approach is based on techniques that are well established and proven in the literature as well as the theoretical suitability of these techniques for the application. It is designed to be applied to most problems without parameter tuning, or substantial work on stopword lists. In the pre-processing step, after a given English document is decomposed into individual paragraphs, a list of 425 words from the Brown stopword list are removed, and stemming is performed using the Porter stemming algorithm. A term-frequency vector for each paragraph in the document is then constructed from the terms' stems.

The core of the approach is in the dimensionality reduction, which is used to model the topic flow of a document's text passages based on their associated topic mixtures. A separate topic model is built for each document, using a document's content to generate the term-by-sentence matrix. In the case of SVD and NMF, the matrix elements of the initial term-by-sentence matrix can be weighted using a number of schemes (Baeza-Yates & Neto, 1999).

Once the topic model is built, each topic is represented as a vector of its distribution of terms and each sentence is represented as a vector of its distribution of terms over these topics: thus producing the topic model from which an analysis of a document's semantic structure and flow can be performed. However, actually quantifying the topic flow in an essay is a difficult task and a methodology for doing so is missing from the literature. While a break in topic flow can sometimes be a good thing, on average it is reasonable to expect that a well written essay's topic flow should be better than that which would be expected from random chance.

In the context of the research presented in this paper, topic flow is quantified as the average amount of semantic overlap between successive sentences or paragraphs in an essay (O'Rourke & Calvo, 2009a). The Distance Index (DI), defined in Equation (4), measures the sum of semantic distances $\hat{d}_{ij}$ between consecutive pairs of sentences or paragraphs, 'centered' and normalized by the average over all the pairs of sentences or paragraphs in a document. These averages are equivalent to distances that would be expected from randomizing the order of the paragraphs. A DI value less than or equal to 0 indicates a random topic flow, while a DI value greater than 0 indicates the presence of topic flow.

$$DI = 1 - \frac{\sum_{i}^{n-1} \hat{d}_{ii+1}}{\frac{2}{n} \sum_{i<j}^{n} \hat{d}_{ij}} \tag{4}$$

Evaluation 1: Flow and Grades

Although the performance of various dimensionality reduction techniques has been examined by many researchers in the literature, the differing results have exposed how the suitability of each technique seems to be domain-dependent. Our approach of measuring the semantic flow of a single document with the use of a semantic space created solely from the actual document itself is unique, and makes the algorithm independent of any particular domain knowledge. This section evaluates the NMF, PLSA and SVD dimensionality reduction techniques in measuring the semantic flow of a text.

Experiment Dataset

The evaluation was performed using an corpus consisting of 120 essays written for assignments by undergraduate students at Mississippi State University (McNamara, et al., 2010). The corpus includes its associated essay grades from which a quality judgment about the essay semantics is inferred. The essays have been graded from 1 to 6. The essays contain an average of 726.20 (SD=114.37) words, 40.03 (SD=8.29) sentences, and 5.55 (SD=1.32) paragraphs.

Measuring Topic Flow

The aim of this experiment was to validate quantitatively whether dimensionality reduction techniques can be used to analyze the topic flow of an essay and measure whether the choice of dimensionality reduction techniques used in the approach is in line with the theoretical justifications discussed in the background sections. As topic flow is generally considered a positive feature of an essay, the assumption was made that these essays do have a measurable degree of topic flow, and that high graded essays have a higher amount of topic flow than that of low grade essays. In order to quantify topic flow the Distance Index (DI) measure in Equation (4) was used.

For each essay, a term-by-sentence weight matrix was calculated using the log-entropy term weighting scheme (other schemes had similar results). Since different dimensionality reduction techniques may affect the DI measure, this evaluation compares the results of the Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF), and Probabilistic Latent Semantic Analysis (PLSA) dimensionality reduction methods. The number of dimensions (topics) used for the dimensionality reduction algorithms was kept at $k = 5$ throughout the experiment. This parameter was chosen based on experimental experience with this dataset; however, similar results were achieved with other k values. The ideal of number of k dimensions is entirely document dependant and automatically determining this value is out of the scope of this paper.

The distance between the pairs of sentences and pairs of paragraphs in the reduced semantic space was calculated using the measure of cosine similarity. The DI was used to calculate and compare the difference in topic flow between the sentences and paragraphs for the graded essays subsets produced using the different dimensionality reduction methods.

Two evaluations were performed. In the first, the essay corpus was divided into a low grade and a high grade subset in order to define a quality benchmark on which to critically evaluate the experimental results. The low grade subset consisted of 67 essays graded 1 to 3 and the high grade subset consisted of 53 essays graded 3.2 to 6. In the second evaluation a correlation coefficient between DI and grades was calculated.

Results

The results for the sentence topic flow and the paragraph topic flow in the MSU corpus are summarized in Table 1. A graph illustrating the difference in the average sentence DI and the average paragraph DI produced using the SVD, NMF, and PLSA dimensionality reduction techniques is displayed in Figure 2.

The average sentence DI was found to be relatively larger in the high grade subset than the low grade subset for all the evaluated dimensionality techniques. The difference in the sentence DI between the graded essay subsets was found to be statically significant using NMF and SVD ($p<0.05$), but not PLSA ($p> 0.05$).

The average paragraph DI, on the other hand, was less present using either of the dimensionality reduction techniques, with a value of close to 0 and little difference between the graded essay subsets. Although the high grade essay subset had a slightly higher average paragraph DI than the low grade subset, the results were not found to be statistically significant ($p > 0.05$) using any of the dimensionality reduction techniques.

Table 1. The mean sentence and paragraph distance indexes (DI) produced using the NMF, PLSA and SVD dimensionality reduction techniques

	Low grade	**High grade**	**p-value**	**Effect size**

Sentence				
NMF Avg DI (SD)	0.21 (0.13)	0.26 (0.15)	0.02	0.39
PLSA Avg DI (SD)	0.19 (0.13)	0.22 (0.14)	0.12	0.23
SVD Avg DI (SD)	0.13 (0.11)	0.19 (0.15)	0.02	0.39
Paragraph				
NMF Avg DI (SD)	-0.01 (0.17)	0.02 (0.14)	0.19	0.17
PLSA Avg DI (SD)	-0.03 (0.18)	0.00 (0.21)	0.17	0.16
SVD Avg DI (SD)	0.01 (0.18)	0.03 (0.18)	0.27	0.12

However, the measure of statistical significance does not take into consideration the actual size of the effect that topic flow has on an essay grade. This can be quantified by calculating the 'effect size', which, as the name suggests, measures the actual size of the difference between two datasets. At the sentence level, the Cohen's effect size of the DI between the graded essay subsets was calculated to be slightly larger using NMF (0.39) and SVD (0.39) compared to that of PLSA (0.23). This means that NMF and SVD were able to detect a greater effect of topic flow on a grade compared to that of PLSA. At the paragraph level, NMF had the highest effect size (albeit non-significant) followed by PLSA and SVD, with values of 0.17, 0.16 and 0.12, respectively.

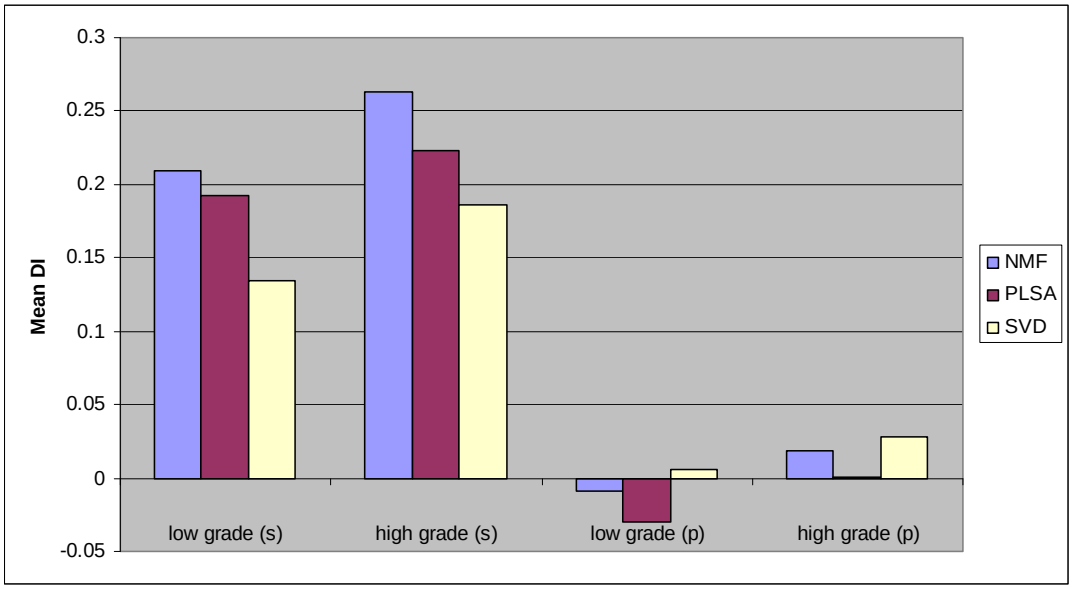

Figure 2. A comparison the average sentence and paragraph distance indexes (DI) produced using the NMF and PLSA and SVD dimensionality reduction techniques for the MSU corpus

With such a strong indication of sentence topic flow in the MSU corpus and the availability of numerical scores, the corpus was further analyzed in more fine-grained detail. Results in Table 2 show that on average the sentence DI of the essays increased with the score, with the results for the NMF algorithm being the most consistent. This result is in support the assumption in this paper of topic flow and essay quality. On average, there was a higher amount of topic flow between sentences than paragraphs.

Table 2. Correlation of the sentence and paragraph distance indexes (DI) with the numerical essay scores

	NMF	**PLSA**	**SVD**
Sentence			
Mean (SD)	0.23 (0.14)	0.21 (0.13)	0.16 (0.13)
Pearson's r	0.27	0.17	0.20
Kendall's τ_B	0.18	0.13	0.10
Spearman's ρ	0.25	0.17	0.15

Paragraph			
Mean (SD)	0.00 (0.16)	-0.02 (0.19)	0.02 (0.18)
Pearson's r	0.16	0.05	0.04
Kendall's τ_B	0.12	0.05	0.03
Spearman's ρ	0.17	0.06	0.04

Evaluation 2: Supporting Assessment

Achieving good consistency in the grading of written assignments is difficult and adds to effort to the already time consuming task. Such consistency is even more difficult when the tutors assessing the assignment are not trained in teaching writing or are non-native speakers (a common scenario). The aim of this evaluation was to study the use of the topic flow maps as a visual aid in the characterization of the texts' semantic structure and flow, supporting the instructors who assess the quality of the composition.

Methodology

The assessment of essays is a subjective task and there is evidence that the visualization of qualitative data can be used to enhance the speed and accuracy of a subjective task (North, 2006). The study evaluated the time tutors take to complete the marking of a collection of essays and the inter-rater agreement that the tutors had with two expert raters. The two tutors independently marked assignments with and without the visual aid of the topic flow maps. Following North's work (North, 2006), it is hypothesized that the structure and development of an essay can be subjectively assessed faster, more accurately, and more consistently with the visual aid of a topic flow map.

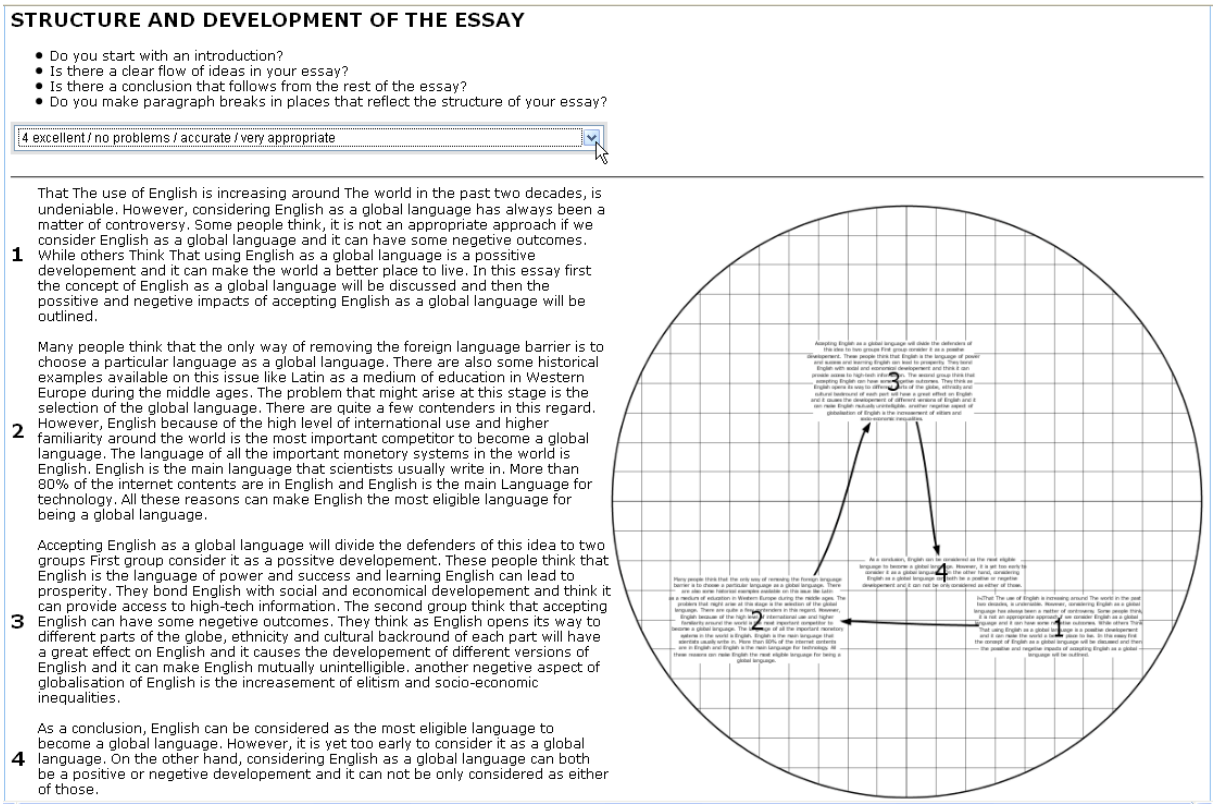

Figure 3. The main interface of the essay assessment tool. An essay (Left) is displayed side-by-side with it corresponding map (Right)

The two tutors marked the assignments using the interface shown in Figure 3 (or a version without the visualization). Each of the essay paragraphs was numbered, so they could easily be identified in the map. The rubrics for marking

are shown at the top of the interface. The rubric focused on questions like: 'Do you start with an introduction? Is there a clear flow of ideas in your essay? Is there a conclusion that follows from the rest of the essay? Do you make paragraph breaks in places that reflect the structure of your essay?' Below each question is a list box for the participant to record their score. Each time a participant changes the score in the list box, their assessment time is updated and logged by the tool.

Essay Subset Preparation

The evaluation was performed using a corpus of N=43 short essays handwritten in a timed assessment by students at the University of Sydney and then assessed by to expert applied linguistics. The essays were then typed in for this study. Due to the timed nature of the assessment and the fact that essays were originally handwritten, they are often quite erroneous. The corpus includes its associated numerical essay marks, as assessed according to the MASUS procedure (Bonanno, Jones, & University of Sydney. Learning Centre., 2007). Compared to the MSU corpus, the essays in the MASUS corpus are slightly shorter in length, with documents having an average of N=445.85 (SD=120.91) words, N=23.30 (SD=5.04) sentences, N=5.40 (SD=1.74) paragraphs.

Essays with only one or two paragraphs were excluded from the experiment, due to their lack of enough useful semantic information. The 40 essays remaining were divided into two subsets of 20 essays each, with roughly equal distributions of words, paragraphs and scores.

Table 3. A comparison of the statistical differences between the MASUS essay subset with maps and the MASUS essay subset with no maps

	Map	**No map**
Essays Total	20	20
Words Mean (SD)	455.80 (116.17)	435.85 (127.71)
Paragraphs Mean (SD)	5.60 (1.50)	5.75 (2.05)
Rating Mean (SD)	3.15 (0.65)	3.20 (0.59)

Results

The essays were assessed by two tutors with an engineering background, one a native English speaker (Rater 1) and the other a non-native English speaker (Rater 2). The average time taken for the raters to assess the essays as a function of condition is provided in Table 4. In order to eliminate the effect of essay length, the time taken to assess an essay was normalized to a measure of words per second. To eliminate interactions, the tutors did not discuss their work during the task. On average, both raters assessed the essays faster with the visual aid of the maps than without. This is an important result, because the raters are processing more rather than less information when they are provided with the maps, but they processed more words per second in the essays when they were provided with the maps.

Table 4. A summary of time taken for the two raters to score the MASUS essays, both with and without the visual aid of the topic flow maps

	Map - Time (words/sec)	**No map - Time (words/sec)**	**p-value**
Rater 1 Mean (SD)	4.60 (0.85)	3.24 (0.55)	< 0.01
Rater 2 Mean (SD)	4.39 (2.28)	3.64 (0.66)	0.08

Separate statistical comparisons for each rater revealed that the difference in the time taken to assess the essays with and without the maps for Rater 2 was not statically significant ($p = 0.08$), whereas it was for Rater 1 ($p < 0.01$). The

large variability (SD=2.28) in the time taken for Rater 2 to score the essays with the aid of the maps can be explained by an outlier entry in the data. However, excluding this entry from the analysis did not improve the statistical significance of the result.

The most important dimension for evaluating whether the raters gained an insight into the semantics of the essay using the maps is by measuring how the maps affected the accuracy and consistency of the scores. This can be measured by calculating the correlation of the rater scores to that of the expert. Correlation determines the inter-rater agreement between two raters. In order to obtain a comprehensive view of this dimension of the results, both nominal and ordinal correlation measures are used.

Using Pearson's r, the respective inter-rater agreement with the expert MASUS raters for Rater 1 and Rater 2 was r = 0.69 and r = 0.63 using the maps compared to r = -0.14 and r = -0.02 without the maps. In both cases, there was a large different in Pearson's correlation for the two essay conditions. Therefore, both raters used the rating scale more consistently in the essay condition with the visual aid of the maps.

Table 5. A summary of the scores given by the two raters and their correlation with the expert MASUS raters

	Score - Map	**Score - No map**
Rater 1		
Mean (SD)	3.25 (0.79)	3.00 (0.92)
Kappa κ	0.47	-0.07
Pearson's r	0.69	-0.14
Rater 2		
Mean (SD)	2.35 (1.09)	2.15 (0.83)
Kappa κ	0.07	-0.07
Pearson's r	0.63	-0.02

Discussion

This paper aimed to show that the techniques used to compute these semantic spaces, and the flow in a document, can be used to 1) quantitatively predict the quality of a composition 2) produce visualizations that help readers (i.e. tutors) assess qualities of the composition more reliably and faster.

The study evaluated cognitive visualizations of student essays, particularly of the composition's flow as measured by distances between text units (sentences or paragraphs). We evaluated the correlations between the distances and the essay scores (i.e. quality) as well as the impact of providing the maps to raters on the time and reliability of assessing the quality of essays. This topic flow visualization method involves a process of dimensionality reduction to uncover topics in an essay, followed by multidimensional scaling to map the topic closeness of the essay's paragraphs to a 2-dimensional representation.

We compared different dimensionality reduction techniques and document similarity comparison for quantifying the semantic features of an essay, and evaluated how these techniques could be used to capture the topic flow of essays. The approach of quantifying topic flow was evaluated using a corpus of 120 essays written by university students. Three dimensionality reduction techniques, NMF, PLSA and SVD were evaluated with respect to the amount of measurable topic flow according to a defined distance index.

The results of this study indicate that NMF, PLSA and SVD capture some effects of semantic flow, which is likely due to greater argument overlap by literal repetition of topics from sentence to sentence. All of the dimensionality reduction techniques indicated that on average they were able to capture the degree to which consecutive sentences discuss similar semantic content. The average distance index between consecutive sentences was found to be greater in the high grade subset than the low grade subset for all the evaluated dimensionality techniques. While the effect size of topic flow between consecutive sentences on an essay grade was calculated to be roughly the same for NMF and SVD, but less for PLSA.

At the paragraph level, however, there was found to be little topic flow between consecutive paragraphs. The effect size of the topic flow on the essays grade was also small. This can be at least partly attributed to the size of the

documents used to build the semantic space. In the experiments of O'Rourke and Calvo (2009a), where much larger documents were used (averaging of 29.42 paragraphs, compared to 5.55 in the present experiment), the paragraph distance index was more closely tied to essay grades. In shorter essays it is likely that each topic will be more confined to individual paragraphs. Similarly it is likely that the topic content in the introduction and conclusions paragraphs will actually be more similar compared to their connecting paragraphs. These features could be observed in many of the topic flow visualizations produced from higher quality essays; where the introduction and conclusion paragraphs were often positioned in close proximity and more central to the other body paragraphs, which were generally positioned more equidistant to each other.

It is clear that the semantic space and thus the inter-document distances differ from one essay to the next, but the challenge remains in providing an automatic means of scoring coherence. The results of this study indicated that there is a measurable amount of topic flow in an essay that relates to the quality of an essay. The results showed that higher quality essays in the corpus had a higher amount of semantic overlap from sentence to sentence, but little from paragraph to paragraph. McNamara et al. (2010) and Crossley and McNamara (2010) have found that cohesion markers are not related to humans' estimates of essay coherence. Here we have examined the extent to which semantic flow is related to essay quality. There is some evidence that there is a relation, but in this case, the coherence resided primarily at the inter-sentence level, rather than between paragraphs (O'Rourke & Calvo, 2009a).

In the second study, topic flow visualization was evaluated using a corpus of 40 essays written by university students, which had been previously assessed by two expert linguists according to the MASUS procedure. The evaluation demonstrated the use of the visualization for assessing the structure and flow of an essay. On average the experiment participants were shown to assess the essays faster and more accurately and consistently with the aid of topic flow visualization.

Conclusions

Understanding the semantic space of an individual document and determining how its features relate to a grade is in no way simple. Intuitively, the semantic space of a document tells us something about the structure and the flow of topics within the document. How to best quantify this intuition using only the information contained in the document itself remains an open research question. A break in topic flow is not necessarily good or bad, but the semantic space should show that 1) the topic will shift such that it is likely that consecutive parts of a text should be more similar compared to those which are further away 2) there should still be a measurable semantic structure in an essay. Our Distance Index results show that this can be achieved, and they also highlight the impact of different feature selection techniques.

The evaluation also shows that the visualization can be a useful tool for assessment. We found that two raters' assessment of the essays was more accurate and consistent with the visual aid of the maps. The results are important because the raters assessed the essays more quickly but also more consistently with the expert raters. This has important educational implications because students may be able to better judge their own essays using such a tool and teachers may be able to score essays more quickly and consistently. Clearly more evidence is needed prior to making firm conclusions and further experiments are needed with more raters on various corpora. Nonetheless, these promising results highlight the need for further research on the use of visualization for interpreting essay semantics. They particularly point to the need of studying the impact of such visualizations on student learning, addressing the challenge of integrating them in realistic learning activities.

References

Baeza-Yates, R., & Neto, B. (1999). *Modern Information Retrieval*: ACM Press / Addison-Wesley.

Beals, T. J. (1998). Between teachers and computers: Does text-checking software really improve student writing? *English Journal, National Council of Teachers of English, 87*(1), 67-72.

Bonanno, H., Jones, J., & University of Sydney. Learning Centre. (2007). *The MASUS procedure : measuring the academic skills of university students : a diagnostic assessment* (3rd Ed.). Sydney: University of Sydney, Learning Centre.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling : theory and applications* (2nd Ed.). New York: Springer.

Calvo, R. A., & Ellis, R. A. (2010). Students' conceptions of tutor and automated feedback in professional writing. *Journal of Engineering Education, 99*(4), 427-438.

Calvo, R. A., O'Rourke, S. T., Jones, J., Yacef, K., & Reimann, P. (2011). Collaborative Writing Support Tools on the Cloud. *IEEE Transactions on Learning Technologies, 4* (1), 88-97.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Paper presented at the 32nd Annual Conference of the Cognitive Science Society*, August 11-14, Portland, OR, USA.

Davis, J., & Liss, R. (2006). *Effective academic writing. 3, The essay*. New York: Oxford University Press.

de Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier & B. Van Cutsem (Eds.), *Recent Developments in Statistics* (pp. 133-146). Amsterdam: North Holland Publishing Company.

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.

Dessus, P. (2009). An overview of LSA-Based systems for supporting learning and teaching *Artificial Intelligence in Education. Building learning systems that care: From knowledge representation to affective modelling (AIED2009)* (pp. 157-164): IOS Press.

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Paper presented at the 24th Annual International ACM SIGIR Conference, September 9-12, New Orleans, LA, USA.*

Graesser, A. C., & McNamara, D. S. (in press). Use of computers to analyze and score essays and open-ended verbal responses. In H. Cooper, P. Camic, R. Gonzalez, D. Long & A. Panter (Eds.), *APA handbook of research methods in psychology*. Washington, DC: American Psychological Association.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments and Computers, 36*(2), 193--202.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning, 42*(1-2), 177-196.

Kakkonen, T., & Sutinen, E. (2004). Automatic assessment of the content of essays based on course materials. *Paper presented at the 2nd International Conference on Information Technology: Research and Education*, June 28 - July 1, 2004, London, UK.

Kirschner, P. A., Shum, S. J. B., & Carr, C. S. (Eds.). (2003). *Visualizing argumentation: software tools for collaborative and educational sense-making*. London: Springer.

Landauer, T. K., McNamara, D., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*: Lawrence Erlbaum.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791.

McNamara, D. S. (2010). Computational Methods to Extract Meaning From Text and Advance Theories of Human Cognition. *Topics in Cognitive Science, 3* (1), 3-17.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication, 27*(1), 57-86.

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*. Philadelphia: Society for Industrial and Applied Mathematics.

North, C. (2006). Toward measuring visualization insight. *IEEE Computer Graphics and Applications, 26*(3), 6-9.

O'Rourke, S. T., & Calvo, R. A. (2009a). Analysing Semantic Flow in Academic Writing. *Paper presented at the 14th International Conference on Artificial Intelligence in Education (AIED 2009)*, July 6-10, 2009, Brighton, UK.

O'Rourke, S. T., & Calvo, R. A. (2009b). Visualizing paragraph closeness for academic writing wupport. *Paper presented at the Ninth IEEE International Conference on Advanced Learning Technologies (ICALT 2009)*, July 15-17, Riga, Latvia.

Osinski, S. (2006). Improving Quality of Search Results Clustering with Approximate Matrix Factorisations. *Paper presented at the 28th European Conference on Information Retrieval Research*, April 10-12, London, UK.

Steinberger, J., Poesio, M., Kabadjov, M. A., & Jeek, K. (2007). Two uses of anaphora resolution in summarization. *Inf. Process. Manage., 43*(6), 1663-1680.

Thiesmeyer, E. C., & Thiesmeyer, J. E. (1990). *Editor*. New York: Modern Language Association.

Villalon, J., & Calvo, R. A. (2009). Single document semantic spaces using Latent Semantic Analysis. Paper presented at the 22nd Australasian Joint Conference on Artificial Intelligence, December 1-4, Melbourne, Australia.

Villalon, J., Kearney, P., Calvo, R. A., & Reimann, P. (2008). Glosser: Enhanced Feedback for Student Writing Tasks. *Paper presented at the IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*, July 1-5, Santander, Spain.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333-362.

Wang, J., & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *Journal of Technology, Learning, and Assessment, 6*(2), retrieved June 20, 2011, from, http://escholarship.bc.edu/ojs/index.php/jtla/article/view/1632/1476.

Wiemer-Hastings, P., & Graesser, A. C. (2000). Select-a-Kibitzer: A Computer Tool that Gives Meaningful Feedback on Student Compositions. *Interactive Learning Environments, 8*(2), 149-169.