

# Analysing Semantic Flow in Academic Writing

*Stephen T. O'Rourke, Rafael A. Calvo*

*School of Electrical and Information Engineering, The University of Sydney*

Many approaches have been proposed for providing feedback in academic writing, however, few of them are visually based. We describe a novel visualisation method for providing feedback to support formative essay assessment. The visualisation method makes use of text mining techniques to provide insight on the semantics of the topics in an essay. We propose that visualisation can be used to mitigate many of the problems associated with the subjectivity of formative essay assessment. The visualisation method involves a process of Non-negative Matrix Factorisation (NMF), to uncover topics in an essay, followed by Multidimensional Scaling, to map the essay topics to a 2-dimensional representation. We evaluate our approach with a subset of the British Academic Written English corpus of 2761 essays written by university students.

Keywords: writing support, visualisation, multidimensional scaling, non-negative matrix factorization

## 1. Introduction

An essay is a piece of writing that can be used to assess a learner's knowledge on a particular topic. Essays are typically graded according to a rubric, which codifies assessment standards and brings greater transparency to the assessment process. However, essay assessment is a subjective process that is costly, time consuming and prone to several types of errors [1], leading to many inconsistencies in the grades given by different assessors. While some aspects of essay writing can be assessed based on objective features, such as the word count and spelling errors, others require a more subjective and insightful interpretation, such as the flow of topics and ideas. These aspects of essay writing are somewhat inexact and cannot be easily abstracted without existing domain knowledge.

North [2] discussed how visualisation can be used to intuitively capture insight, listing some of its important characteristics as being complex, deep, qualitative, unexpected and relevant. Features possessing these characteristics can be abstracted from complex datasets, such as unstructured text or a high-dimensional vector model, by transforming the data to a low-dimensional visual representation. This approach provides a qualitative view of a dataset, which can be used to bring insight to its latent structures and relationships [3].

We propose visualisations to enhance formative assessment. The goal is not to assess an essay based on a visual formula, but rather use visualisation to bring greater insight to features that require a subjective interpretation.

The next section reviews some of the extensive literature on tools for supporting academic writing. Section 3 presents a short description of the text mining techniques used to analyse the essay features. In Section 4, the visualisation is introduced and explained using examples. In Section 5, we provide an evaluation of the techniques used in our approach and assess to what extent they allow one to capture topic flow. Finally, Section 6 concludes the paper.

## 2. Background

Researchers have developed tools to support academic writing, including simple feedback on so called ‘surface features’, such as the number citations or spelling errors, as well as feedback on more subjective aspects, such as the flow of ideas. The Writer’s Workbench tool provides automatic feedback on spelling, style and diction by analysing English prose and suggesting possible improvements [4]. The Sourcer’s Apprentice Intelligent Feedback tool [5] provides automatic feedback on sourcing by detecting citations and plagiarised sentences and suggesting ways to resolve them. Glosser [6] provides automatic feedback by highlighting important essay features and using thought provoking questions to promote reflection. Other tools have been used to review essays.

Several automated feedback methods have been proposed for analysing and interpreting the semantic features of an essay. Foltz [7] used Latent Semantic Analysis (LSA) to measure the coherence of a document by calculating the degree of semantic relatedness between consecutive text passages. Using LSA, Foltz was able to successfully predict the effect of text coherence on readers’ comprehension. LSA is a corpus based technique that relates documents through term co-occurrence [8]. LSA uses a matrix factorisation technique called Singular Value Decomposition (SVD) to find a low rank approximation of a term-by-document occurrence matrix. This low rank approximation identifies a set of basis vectors which capture most of the variance of the corpus in a linear space. These basis vectors can be linearly combined to represent any document in the space. Thus, allowing documents to be indirectly related through the semantics of the basis vectors they span in the linear space.

Although LSA performed fairly well in Foltz’s experiments, it has some theoretical limitations in the context of measuring topic flow. This is primarily due to the interpretability of its basis vectors, which contain both positive and negative values. As a document vector is represented in a linear space as a combination of the basis vectors it spans, negative values can at times be contradictory and can lead to ambiguity in relating the conceptual features of a document. Moreover, this makes it difficult to interpret the topics of a document as its representation in the linear space is not an additive combination of the topics it spans.

The interpretability problems with SVD’s basis vectors have led researchers to propose alternative matrix factorisation methods which maintain the non-negativity of original term-by-document matrix. Non-negative Matrix Factorisation (NMF) is one such method that offers a more intuitive approach for creating a topic model of a document. By permitting only positive entries in its basis vectors, a document can be represented in a linear space as a non-subtractive combination of its parts to form a whole [9]. Using NMF, each basis vector gives rise to a distinct topic, allowing for a document to be modelled as an additive non-negative combination of topics. A non-

negative solution also allows for topic overlap and thus provides a more direct measure of a document's topic mixture.

### 3. Data representation and processing

Text mining techniques are used here to model the topic mixture of paragraphs and map them to a 2-dimensional space for visual consumption. The automated mapping approach involves performing the following steps. First, a term-by-document matrix is prepared, after stop-words and low frequency words are removed, and stemming is applied. Second, a topic model is created using NMF. Third, the topic model is projected to a 2-dimensional space using Multidimensional Scaling, and finally, a visualisation of the document is produced.

This is the same process used in LSA, except for two variations. First, the NMF model is used instead of SVD. Second, since the visualisations are performed for a single document instead of a collection, paragraphs are used instead of documents. The factors produced by NMF decomposition can still be interpreted as 'topics' as they are in LSA.

The elements of the initial term-by-paragraph matrix can be weighted using a number of schemes (i.e. Chi-squared, Log-entropy, TF-IDF) [10]. The results in the next section are produced using Log-entropy, although the same visualisation can be produced with the other approaches. Log-Entropy 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 weight scheme for our purposes because it assigns higher weights to terms that appear fewer times in a smaller number of paragraphs. Thus, emphasising the importance the infrequent terms in the paragraphs while also eliminating the 'noise' of frequent terms.

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

NMF 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 factorisation of  $X$  which can be approximated by minimising the squared error of the Frobenius norm [11] of  $X - WH$ . Finding this solution defines the NMF problem which can be mathematically expressed as,

$$F(W, H) = \|X - WH\|_F^2 \quad (2)$$

The details of an algorithm for solving the NMF problem are available in [12]. The NMF algorithm uses an iterative procedure to multiplicatively update the initial values of  $H$  (equation 3) and  $W$  (equation 4) so that the product approaches  $X$ . The initial values of  $H$  and  $W$  are randomly generated such that  $W_{ij} > 0$ . Once the NMF model is calculated, each topic is represented as a vector of its distribution of terms and each of a document's text passage is represented as a vector of its distribution of terms over these topics.

$$H_{cj} \leftarrow H_{cj} \frac{(W^T V)_{cj}}{(W^T WH)_{cj}} \quad (3)$$

$$W_{ic} \leftarrow W_{ic} \frac{(VH^T)_{ic}}{(WHH^T)_{ic}} \quad (4)$$

The distance between any two paragraphs can be calculated in the reduced topic representation using standard measures (cosine, Euclidean...). In this way, the distance between any two paragraphs (not only the consecutive ones) is calculated. Multidimensional Scaling uses this paragraph-paragraph triangular distance table to produce a 2-dimensional representation [13]. For example, given the distances between all the cities in a country, Multidimensional Scaling could be used to plot the relative location of each city on a 2-dimensional map. The Multidimensional Scaling transformation is performed using a procedure called iterative majorisation [14]. The iterative majorisation algorithm undertakes an iterative, least-squares approach to Multidimensional Scaling by attempting to minimise a loss function called Stress. Stress (equation 5) can be expressed as the normalised sum of the squared errors between the vector dissimilarities  $\hat{d}_{ij}$  and their approximated distances  $d_{ij}$  in the low dimensional space:

$$\sigma = \sum_{i < j} \frac{(d_{ij} - \hat{d}_{ij})^2}{\hat{d}_{ij}^2} \quad (5)$$

For each iteration, the iterative majorisation algorithm minimises stress by creating a new configuration of points that yields an optimal arrangement of the documents. The final result is a least-squares representation of the paragraphs described in the distance matrix, with the directions of the axes being arbitrary.

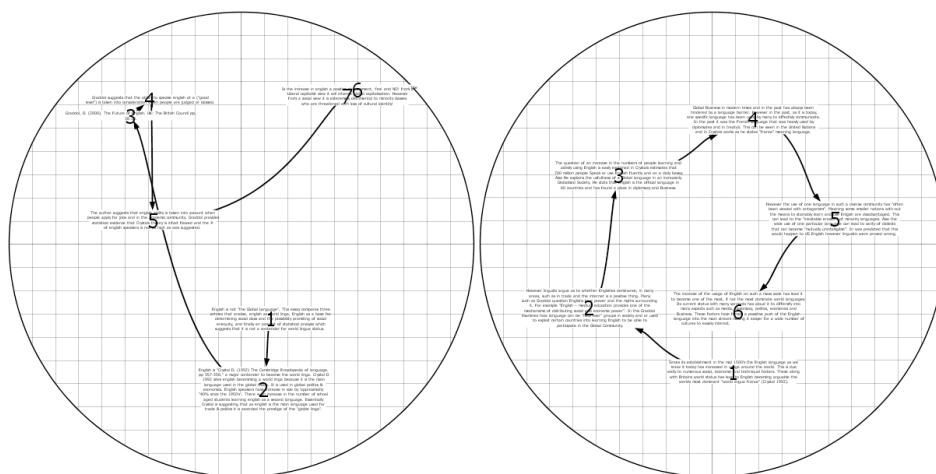
#### 4. Analysis of the 2-dimensional visualisations

A well structured essay should show a clear and logical flow of ideas through the flow in paragraphs. The 2-dimensional representation can be used to produce a final representation that students and tutors can use. To support feedback on the flow of an

essay, we use the paragraph visualisation aiming for students to gain insight on how they 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 [15], can be visualised in a paragraph map. In this paradigm, 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.

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 disorganised. The low grade essay shows signs of disconnected through rough topic shifts, as well as repetition through paragraphs of similar topic mixtures.



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

## 5. EVALUATION

The paragraph maps were evaluated using the British Academic Written English (BAWE) corpus. The BAWE corpus consists of 2761 documents written for assignments by university students over a four year period, from 2004 to 2007. The corpus contains documents written in a variety of genres on various topics. The documents have been graded as either Merit (60-70%) or Distinction (70-100%). A

subset of essays with between 5 and 50 paragraphs was selected for testing. The subset is comprised of 296 distinction graded essays and 577 merit graded essays (numeric grades were not available). As these essays are considered to be of a reasonable grade, we make the assumption that the essays have a measurable degree of topic flow, and that the topic flow of the distinction essays is greater than that of the merit essays. The essay corpus was divided into two graded subsets created from the merit and distinction graded essays. We performed an experiment on the graded essay subsets, which sought to quantitatively validate our approach for analysing topic flow.

The distance index, defined in equation 6, measures the sum of distances  $\hat{d}_{ij}$  between consecutive pairs of paragraphs, ‘centred’ and normalised by the average over all the  $n$  pairs of paragraphs in a document. These averages are equivalent to distances that would be expected from randomising the order of the paragraphs. A distance index value less than or equal to 0 indicates a random topic flow, while a value greater than 0 indicates the presence of topic flow.

$$DI(i) = 1 - \frac{\sum_{i \neq j}^n \hat{d}_{ij}}{\frac{1}{n} \sum_{j \neq k}^n \hat{d}_{jk}} \quad (6)$$

Since different dimensionality reduction techniques may affect this distance index, this evaluation compares the results of the NMF and SVD matrix factorisation algorithms. The number of dimensions used for the matrix factorisation algorithms was kept at 5 throughout the experiments.

For each essay, a term-by-paragraph weight matrix was calculated using the Log-entropy term weighting schemes (other schemes had similar results). The matrix was reduced to 5 latent factors using the SVD and NMF matrix factorisation methods to create topic models of the paragraphs. The distance index was used to calculate and compared the difference in topic flow between the graded essays subsets produced using NMF and SVD topics models. A summary of the experiment results is displayed in Table 1.

The results in Table 1 show that the distinction essays had a higher average distance index compared to that of the merit essays. This result is in agreement with our expectation of topic flow and essay quality. The p-value calculated from the NMF results are considered statistically significant, while on the other hand, the p-value from the SVD results do not indicate any statistical significance. Thus, the results in Table 1 can be seen as giving cause to the hypotheses that there is indeed measurable topic flow among the paragraphs of an essay. On average the NMF algorithm performed much better than SVD in measuring a documents topic mixture, leading us to conclude that in this case NMF is a much more useful technique for analysing topic flow compared to that of SVD.

**Table 1.** A comparison of the average distance indexes produced using the NMF and SVD matrix factorisation methods.

Matrix Factorisation	Distinction Essays	Merit Essays	p-value	effect size
----------------------	--------------------	--------------	---------	-------------

NMF				
Mean DI.	0.1908	0.1626	< 0.01	0.18
Std dev.	0.1579	0.1605		
SVD				
Mean DI.	0.0964	0.0866	0.19	0.06
Std dev.	0.1526	0.1570		

The effect size of the topic flow (with MSF) on the essay grade was small, but present. We expect to see a bigger effect on collections of essays where the quality differences are even more noticeable. Such research is in the domain of automatic essay scoring for which there exist many different measures. Importantly, it should be noted, that topic flow does not necessarily always strongly relate to a grade. Indeed, the results contained many examples of distinction essays with poor topic flow (i.e. essays that had a worse topic flow than would be expected from random chance).

## 6. CONCLUSIONS AND FUTURE WORK

We contribute a visualisation to support feedback in academic essay writing. The paragraph map is novel in visualising the semantics of a document to provide insight on the structure of its topic mixture. The mapping approach involves a process of NMF to uncover the topic mixture of the document's paragraphs, followed by Multidimensional Scaling to map the semantics of the paragraphs to a 2-dimensional representation.

The use of matrix factorisation for measuring topic flow was evaluated using a corpus of essays written by university students. We tested two matrix factorisation algorithms (NMF and SVD) with respect to the degree of measurable topic flow, according to our defined distance index. The experiments revealed that NMF was significantly better at capturing topic flow compared to that of SVD, but the effect size of the grades was too small. In future work, we hope to verify how well the topic flow measured by matrix factorisation compares to that of human assessors.

## 7. REFERENCES

- [1] L. M. Rudner, "Reducing Errors Due to the Use of Judges," *Practical Assessment, Research & Evaluation*, vol. 3, 1992.
- [2] C. North, "Toward measuring visualization insight," *Ieee Computer Graphics and Applications*, vol. 26, pp. 6-9, May-Jun 2006.
- [3] U. M. Fayyad, G. G. Grinstein, and A. Wierse, *Information visualization in data mining and knowledge discovery*. San Francisco, Calif.; London: Morgan Kaufmann, 2002.
- [4] N. H. Macdonald, L. T. Frase, P. S. Gingrich, and S. A. Keenan, "The Writers Workbench - Computer Aids for Text Analysis," *IEEE Transactions on Communications*, vol. 30, pp. 105-110, 1982.
- [5] A. Britt, P. Hastings, A. Larson, and C. Perfetti, "Using Intelligent Feedback to improve Sourcing and Integration in Students' Essays," *International Journal of Artificial Intelligence in Education*, vol. 14, pp. 359-374, 2004.

- [6] J. Villalon, P. Kearney, R. Calvo, and P. Reimman, "Glosser: Enhanced feedback for student writing," in *Proceedings of the International Conference on Advanced Learning Technologies*, 2008.
- [7] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis," *Discourse Processes*, vol. 25, pp. 285-307, 1998.
- [8] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990.
- [9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, Oct 1999.
- [10] Ricardo and Berthier, *Modern Information Retrieval*: ACM Press / Addison-Wesley, 1999.
- [11] C. D. Meyer, *Matrix analysis and applied linear algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 2000.
- [12] T. Lee, J. Hendlar, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, pp. 34-43, 2001.
- [13] I. Borg and P. J. F. Groenen, *Modern multidimensional scaling : theory and applications*, 2nd ed. New York ; London: Springer, 2005.
- [14] J. de Leeuw, "Applications of Convex Analysis to Multidimensional Scaling," in *Recent Developments in Statistics*, J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, Eds. Amsterdam: North Holland Publishing Company, 1977, pp. 133-146.
- [15] J. Davis and R. Liss, *Effective academic writing. 3, The essay*. New York: Oxford University Press, 2006.