Analyzing Semantic Flow in Academic Writing

Stephen T. O’ROURKE, Rafael A. CALVO
School of Electrical and Information Engineering, The University of Sydney

Abstract. 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 assignments written by university students.

Keywords. writing support, visualisation, multidimensional scaling, non-negative matrix factorisation

1. Introduction

An essay is a writing task that is often used to assess students’ attainment of learning outcomes. 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 variations in the feedback and grades given by different assessors. While some aspects of essay writing can be assessed based on objective features, such as the number of citations or spelling errors, others require a more subjective and insightful interpretation, such as the flow of topics. 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 the use of visualisation to enhance formative essay 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 automated approaches 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 an example. 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 numerous tools to support academic writing, including simple feedback on so-called ‘surface features’, such as the number of citations and spelling errors, as well as feedback on more subjective features, such as the flow of topics. 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.

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 overlap 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 base vectors which capture most of the variance of the documents in a linear space. These base vectors can be linearly combined to represent any document in the space. Thus, documents can be indirectly related through the semantic overlap of the base vectors they span.

Although LSA performed well in Foltz’s experiments, it has some theoretical limitations in the context of analysing topic flow. This is primarily due to the interpretability of its base vectors, which contain both positive and negative values. In LSA, these base vectors can be interpreted as ‘topics’, whose values describe the extent to which a topic does or does not contain a term. As a document vector is represented in a linear space as a combination of the base vectors it spans, negative values can at times be contradictory and lead to ambiguity in the meaning of a document’s topic mixture. Moreover, this makes it difficult to interpret the topics of a document as its representation in the linear space is not always a purely additive combination of the topics it spans.

The interpretability problems with SVD’s base vectors have led some 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 only permitting positive entries in its base 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 base vector gives rise to a distinct topic, allowing for a document to be modelled as an additive non-negative combination of topics.
negative solution also allows for greater 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 a document’s 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-paragraph 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 visualisation is generated for a single document instead of a collection, paragraphs are used instead of documents. The base vectors produced using NMF 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 weights a term $i$ by the log of its frequency $f_{ij}$ in a paragraph $j$ 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 Eq. (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, the scheme emphasises the importance of infrequent terms while also eliminating the ‘noise’ of frequent terms.

$$x_{ij} = \log(1 + f_{ij}) - \sum_{k=1}^{n} \frac{f_{ik}}{\sum_{a} f_{a}} \log \left( \frac{f_{ik}}{\sum_{a} f_{a}} \right)$$

NMF generates its topic model by decomposing the term-by-paragraph matrix $X$ 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-paragraph matrix, indicating the weighting of each topic in a paragraph. The product $WH$ is called a non-negative 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 is mathematically expressed in Eq. (2).
\[ F(W, H) = \|X - WH\|_F^2 \]  

(2)

A review of algorithms for solving the NMF problem is available in [12]. In our approach, we use The prototypical multiplicative algorithm developed by Lee and Seung [13]. This NMF algorithm uses an iterative procedure to multiplicatively update the initial values of \(H\) and \(W\) so that the product approaches \(X\). The update rules for \(W\) and \(H\) are defined in Eqs. (3) and (4).

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 paragraph is represented as a vector of its distribution of terms over these topics.

\[ H_{ij} \leftarrow H_{ij} \left( \frac{W^T V}{W^T WH} \right)_j \]  

(3)

\[ W_{ic} \leftarrow W_{ic} \left( \frac{V H^T}{W H H^T} \right)_c \]  

(4)

The distance between any two paragraph vectors can be calculated in the reduced topic representation using standard measures (cosine similarity, Euclidean...). Using these measures, the distance between any two paragraphs (not only the consecutive ones) is calculated. Multidimensional scaling uses this paragraph-by-paragraph triangular distance matrix to produce a 2-dimensional representation [14]. 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 [15]. The iterative majorisation algorithm undertakes an iterative, least-squares approach to multidimensional scaling by attempting to minimise a stress function. The stress function in Eq. (5) defines the squared relative errors between the paragraph vector distances \(d_{ij}\) and their approximated Euclidean distances \(\hat{d}_{ij}\) in the 2-dimensional space.

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

(5)

For each iteration, the iterative majorisation algorithm progressively minimises the stress function by creating a new configuration of points until it converges to an optimal arrangement of the paragraphs at its minima. The final result is a 2-dimensional representation of the paragraphs described in the distance matrix, with the directions of the actual axes being arbitrary.
4. Analysis of the 2-Dimensional Visualisation

A well structured and developed essay answer should have a clear and logical flow of topics throughout its paragraphs. To support feedback on this aspect of essay writing, we use the 2-dimensional representation to provide insight into an essay's topic flow. In a paragraph ‘map’, such as those in Figure 1, an 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 sequence of paragraph.

Consider, for example, how the clear sequence of topics in the five paragraph essay paradigm [16] would appear 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. Thus, the map of an 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 paragraphs 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 almost 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 possible signs of disconnectedness through rough topic shifts as well as repetition through paragraphs of near identical 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, form 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%), with the exact numeric grades unavailable. We divided the BAWE corpus into two graded subsets of essay documents for the experiment. The subsets contain 575 merit essays and 295 distinction essays respectively, each of which consists of essays with between 5 and 50 paragraphs.

The aim of the experiment was to quantitatively validate whether matrix factorisation can be used to analyse topic flow of an essay. As topic flow is generally considered to be a positive feature of an essay, we make the assumption that these essays do have a measurable degree of topic flow, and that the distinction essays have a higher topic flow than that of the merit essays.

In order to quantify topic flow, we define it as the average amount of semantic overlap between successive paragraphs in an essay. The distance index, defined in Eq. (6), measures the sum of distances \( d_{ij} \) between consecutive pairs of paragraphs, ‘centred’ and normalised by the average over all the pairs of \( n \) 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 distance index with a value greater than 0 indicates the presence of topic flow. The distance between paragraphs was calculated using the measure of cosine similarity.

\[
DI = 1 - \frac{\sum_{i=2}^{n} d_{ij}}{\frac{1}{2} \sum_{i=1}^{n} d_{ii}}
\]  

(6)

For each essay, a term-by-paragraph weight matrix was calculated using the log-entropy term weighting scheme (other schemes had similar results). Since different dimensionality reduction techniques may affect this distance index, this evaluation compares the results of both the NMF and SVD matrix factorisation methods. The number of dimensions (i.e. topics) used for the matrix factorisation algorithms was kept at 5 throughout the experiments. The distance index was used to calculate and compare the difference in topic flow between the graded essays subsets produced by the different matrix factorisation methods. 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 assumption of topic flow and essay quality. The p-value calculated from the NMF results was statistically significant, while on the other hand, the p-value from the SVD results did not indicate any statistical significance. According these measures, the NMF algorithm performed much better than SVD in measuring a document’s topic flow.
This result leads us to conclude that in this case NMF is a much more appropriate technique for analysing topic flow.

<table>
<thead>
<tr>
<th>Matrix Factorisation</th>
<th>Distinction Essays</th>
<th>Merit Essays</th>
<th>p-value</th>
<th>effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>0.1908</td>
<td>0.1626</td>
<td>&lt;0.01</td>
<td>0.18</td>
</tr>
<tr>
<td>Std dev.</td>
<td>0.1579</td>
<td>0.1605</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVD</td>
<td>0.0964</td>
<td>0.0866</td>
<td>0.19</td>
<td>0.06</td>
</tr>
<tr>
<td>Std dev.</td>
<td>0.1526</td>
<td>0.1570</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The effect size of the topic flow on the essay grade was small, but present. We would expect to see a bigger effect on collections of essays where the quality differences are more significant. Importantly though, it should be noted, that topic flow did not necessarily always strongly relate to an essay 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). This is partly due to the fact that there are many factors which influence an essay grade, and thus no one measure can exactly account for a grade alone. Such research is in the domain of automated essay scoring, for which numerous different measures exist.

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 its topic flow. 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 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 methods (NMF and SVD) with respect to the degree of measurable topic flow according to our defined distance index. The experiment results revealed that NMF was significantly better at capturing topic flow compared to that of SVD, but the effect size in relation to grades was 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