What is my essay really saying? Using extractive summarization to motivate reflection and redrafting

Nicolas Van Labeke[1], Denise Whitelock[1], Debora Field[2], Stephen Pulman[2], John Richardson[1]

[1] Institute of Educational Technology
The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
Nicolas.Vanlabeke@open.ac.uk
Denise.Whitelock@open.ac.uk
John.T.E.Richardson@open.ac.uk

[2] Department of Computer Science
University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, UK
stephen.pulman@cs.ox.ac.uk
debora.field@cs.ox.ac.uk

ABSTRACT

This paper reports on progress on the design of OpenEssayist, a web application that aims at supporting students in writing essays. The system uses techniques from Natural Language Processing to automatically extract summaries from free-text essays, such as key words and key sentences, and carries out essay structure recognition. The current design approach described in this paper has led to a more "explore and discover" environment, where several external representations of these summarization elements would be presented to students, allowing them to freely explore the feedback, discover issues that might have been overlooked and reflect on their writing. Proposals for more interactive, reflective activities to structure such exploration are currently being tested.

Keywords

Essay writing; Extractive Summarization; Formative Feedback; External Representations; Reflective Activities.

1. INTRODUCTION

Written discourse is a major class of data that learners produce in online environments, arguably the primary class of data that can give us insights into deeper learning and higher order qualities such as critical thinking, argumentation and mastery of complex ideas. These skills are indeed difficult to master as illustrated in the revision of Bloom's Taxonomy of Educational Objectives (Pickard 2007) and are a distinct requirement for assessment in higher education. Assessment is an important component of learning and in fact (Rowntree 1987) argues that it is the main driver for learning and so the challenge is to provide an effective automated interactive feedback system that yields an acceptable level of support for university students writing essays.

Effective feedback requires that students are assisted to manage their current essay-writing tasks and to support the development of their essay-writing skills through effective self-regulation.

Our research involves using state-of-the-art techniques for analyzing essays and developing a set of feedback models which will initiate a set of reflective dialogic practices. The main pedagogical thrust of e-Assessment of free-text projects is how to provide meaningful "advice for action" (Whitelock 2010) in order to support students writing their summative assessments. It is the combination of incisive learning analytics and meaningful feedback to students which is central to the planning of our empirical studies. Specifically, we are investigating whether summarization techniques (Lloret & Palomar 2012) could be used to generate formative feedback on free-text essays submitted by students.

This paper is organized as follows. We briefly describe the context and research questions that are informing the design principles of our platform, OpenEssayist. We then describe the basic processes behind the summarization techniques implemented in the system and, finally, demonstrate the current stage of design of the prototype, in particular the use of external representations for the summarization elements. We conclude this paper by sketching our current and planned evaluations.

2. DEFINING A DESIGN SPACE FOR OPENESSAYIST

2.1 WRITING SUMMARIES VS. REFLECTING ON SUMMARIES FOR WRITING.

Writing summaries has been a long-standing educational activity and has received some serious attention in delivering computer-based support. For example, systems such as SummaryStreet (Wade-Stein & Kintsch 2004) or Pensum (Villiot-Leclercq *et al.* 2010) aim to help students *write* summaries as a learning, skills-based, task.

But using summaries as a source of reflection on your own writing seems to be a more open issue. Recent research on formative feedback suggests indeed that essay summarization, understood to comprise both a short summary of the essay and a simple list of its main topics, could be useful for students, e.g. *"to help determine whether the actual performance was the same as the intended performance"* (Nelson & Schunn 2009, p. 378).

With this in mind, one of our research questions is how to use advances in Natural Language Processing to design an automated summarization engine that would provide a good foundation for a dedicated model of formative feedback. Can we use summarization elements to help students identify or visualize patterns in their essays, as explored by (O'Rourke & Calvo 2009)? Or to trigger questions and reflective activities, as implemented in Glosser (Villalon *et al.* 2008)?

2.2 SUPPORTING ESSAY WRITING IN DISTANCE LEARNING

The context of application of our research agenda is supporting students at the Open University (OU) in writing assignment essays. Specifically, we have been working closely with a postgraduate module *Accessible online learning: Supporting disabled students* (referred to as H810). This postgraduate module runs twice a year for about 20 weeks and contributes to a Master of Arts (MA) in Online and Distance Education. All courses, materials and support are delivered online. Students on this module, as is the case for most of the students at the OU, are typically part-time, mature students, who have not been in formal education for a long period of time. It is therefore unsurprising that writing essays, a common assignment in most of the OU courses, proves to be a challenging task for students (and, anecdotal evidence suggests, a common reason for drop-out).

At the same time, OU students often have extensive work experience in a wide variety of areas, and that experience is explicitly capitalized on in the assignments. This means that essays can vary greatly in subject matter. To illustrate this point, two examples of assignment tasks are given in Table 1.

Table 1. Examples of assignment tasks.

TMA1 (1500 words)
Write a report explaining the main accessibility challenges for disabled learners that you work with or support in your own work context(s). Use examples from your own experience, supported by the research and practice literature. If you're not a practitioner, write from the perspective of a person in a relevant context. Critically evaluate the influence of the context (e.g. country, institution, perceived role of online learning within education) on the: (1) identified challenges; (2) influence of legislation; (3) roles and responsibilities of key individuals; (4) role of assistive technologies in addressing these challenges.
TMA2 (3000 words)
Critically evaluate your own learning resource in the following ways: (1) Briefly describe the resource and its accessibility features; (2) Evaluate the accessibility of your resource, identifying its strengths and weaknesses; (3) Reflect on the processes of creating and evaluating accessible resources.

The questions we are considering, given this context, is how we can support these students as they write essays and what the implications are for the design of a computer- and summarization-based approach.

In the initial phase of the project, we ran a couple of focus groups with OU students that helped to identify many aspects of the students' personal approach to essay writing (Alden *et al.* 2013).

Writing an essay is a task that can involve several stages: preparation of material, drafting of essay, reflecting on feedback, summative evaluation by tutors. But not all of them are suitable, or even desirable, for support in an automated assessment system.

Moreover, writing a 1500+ word essay is not a casual operation, nor is it handled in the same way by different students. For example, we discovered that some students are not using computers to draft their essays, because of unease, lack of permanent access to a desktop computer or simply because they still prefer to write their text with paper-and-pencil before typing for the final submission.

Relying on embedded text editors or on cloud-based solutions such as Google Docs – as done by (Southavilay *et al.* 2013) for collaborative writing – is therefore not a viable solution in our context. The system will have to accept texts written with whatever platform students are using to organize, draft and revise their essay. Ultimately, the system will have to be seen and used as a resource, the way forums, online textbooks and other digital tools are used by OU students.

One of the consequences of such selective support is that the flow of activities during the overall writing process is likely to be highly scattered in time: the core of the activity (i.e. writing) will take place *outside* the system's ecology and its use will be mostly as an ancillary to that main task. Careful attention will have to be paid to trade-offs between support and distraction, especially when it comes to interaction, formal reflective activities, accessibility and usability[1].

Finally, the diversity of content in student essays is one of the motivations for investigating summarization techniques as a backbone for formative feedback. Unlike other NLP techniques such as Latent Semantic Analysis (LSA), used in many educational systems, we will not be relying on a corpus of essays to compare and grade new essays accordingly. Summarization using the text alone with no domain-specific knowledge will enable OpenEssayist to handle assignments which have open topics, as well as enabling it to be applied without extensive further development to new subject areas.

2.3 A WEB APPLICATION FOR SUMMARIZATION-BASED FORMATIVE FEEDBACK.

OpenEssayist is developed as a web application and is composed primarily of two components (Figure 1, see appendix). The first component, EssayAnalyser, is the summarization engine, implemented in Python with NLTK[2] (Bird *et al.* 2009) and other toolkits. It is being designed as a stand-alone RESTful web service, delivering the basic summarization techniques that will be consumed by the main system. The second component is OpenEssayist itself, implemented on a PHP framework. The core system consists of the operational back-end (user identification, database management, service brokers, feedback orchestrator) and the cross-platform, responsive HTML5 front-end.

The intended flow of activities within the system can be summarized as follows. Students are registered users and have assignments, defined by academic staff, allocated to them. Once they have prepared a draft offline and seek to obtain feedback, they log on to the OpenEssayist system and submit their essay for analysis, either by copy-and-paste or by uploading their document. OpenEssayist submits the raw text to the EssayAnalyser service and, upon completion, retrieves and stores the summarization data. From that point on, the students, at their own pace, can then explore the data using various external

[1] Worth noting is that students who mention that they don't use computers for drafting their essays also report that they are using smart phones. A focus on responsive user interface suitable for mobile (and tablet) and on asynchronous data access will be an issue for serious consideration in this project.

[2] Natural Language Processing Toolkit, see http://nltk.org/

representations made available to them, can follow the prompts and trigger questions that the Feedback Orchestrator might generate from the analysis and can then start planning their next draft accordingly.

Again, this rewriting phase will take place offline, the system merely offering repeated access to the summarization data and feedback, as a resource, until the students are prepared to submit and explore the summarization feedback on their second draft and on the changes across drafts. This cycle of submission, analysis and revision continues until the students consider their essay ready for summative assessment.

3. EXTRACTIVE SUMMARIZATION

We decided to start experimenting with two simpler summarization strategies that could be implemented fairly quickly: key phrase extraction and extractive summarization, following the TextRank approach proposed and evaluated in (Mihalcea & Tarau 2004). Key phrase extraction aims at identifying which individual words or short phrases are the most suggestive of the content of a discourse, while extractive summarization is essentially the identification of whole key sentences. Our hypothesis is that the quality and position of key phrases and key sentences within an essay (i.e., relative to the position of its structural components) might give an idea of how complete and well-structured the essay is, and therefore provide a basis for building suitable models of feedback.

The implementation of these summarization techniques is based on three main automatic processes: 1) recognition of essay structure; 2) unsupervised extraction of key words and phrases; 3) unsupervised extraction of key sentences.

Before extracting key terms and sentences from the text, the text is automatically pre-processed using some of the NLTK modules (tokenizer, lemmatizer, part-of-speech tagger, list of stop words).

3.1 STRUCTURE IDENTIFICATION

The automatic identification of essay structure is carried out using handcrafted rules developed through experimentation with a corpus of 135 essays that have been previously submitted for the same H810 module. The system tries to automatically recognize which structural role is played by each paragraph in the essay (summary, introduction, conclusion, discussion, references, etc.). This identification is achieved regardless of the presence of content-specific headings and without getting clues from formatting mark-up. With the essays in the corpus varying greatly in structure and formatting, it was decided that structure recognition would be best achieved without referring to a high-level formatting mark-up.

3.2 KEY WORD EXTRACTION

EssayAnalyser uses graph-based ranking methods to perform unsupervised extractive summarization of key words. The 'key-ness' value of a word can be understood as its 'significance within the context of the overall text'.

To compute this key-ness value, each unique word in the essay is represented by a node in a graph, and co-occurrence relations (specifically, within-sentence word adjacency) are represented by edges in the graph. A centrality algorithm – we have experimented with betweenness centrality (Freeman 1977) and PageRank (Brin & Page 1998) – is used to calculate the significance of each word. Roughly speaking, a word with a high centrality score is a word that sits adjacent to many other unique words which sit adjacent to many other unique words which…, and so on. The words with high centrality scores are the key words[3].

Since a centrality score is attributed to *every* unique word in the essay, a decision needs to be made as to what proportion of the essay's unique words qualify as key words. The distribution of key word scores follows the same shape for all essays, an acute "elbow" and then a very long tail, observed for word adjacency graphs by (Ferrer i Cancho & Solé 2001). We therefore currently take the key-ness threshold to be the place where the elbow bend appears to be sharpest.

Once key words have been identified, the system matches sequences of these against the surface text to identify within-sentence key phrases (bigrams, trigrams and quadgrams).

3.3 KEY SENTENCE EXTRACTION

A similar graph-based ranking approach is used to compute key-ness scores to rank the essay's sentences. Instead of word adjacency (as in the key word graph), co-occurrence of words across pairs of sentences is the relation used to construct the graph. More specifically, we currently use cosine similarity to derive a similarity score for each pair of sentences. Whole sentences become nodes in the graph, while the similarity scores become weights on the edges connecting pairs of sentences. The TextRank key sentence algorithm is then applied to the graph to compute the centrality scores.

3.4 ESSAY ANALYSIS OUTPUT

The text submitted for analysis is stripped of its surface formatting and returned as a *new* annotated structured text, reflecting the various elements identified by EssayAnalyser: sentences and paragraphs, labeled with their structural roles (body, introduction, headings, conclusions, captions, etc.) and confidence levels.

Key words and key phrases are returned as an ordered list of terms, associated with various metrics such as centrality, frequency of inflected forms, etc. Key sentences are identified within the annotated text by their ranked centrality scores.

In addition to the core summaries of the essay, various metrics and specialized data structures are made available, for use by the system for diagnosis purpose (or by researchers for analysis): word and sentence graphs, word count, paragraph and sentence density and length, number of words in common with the module textbook, average frequency of the top handful of most frequent words, etc.

Our task is now to look for ways of presenting and exploiting these results and, ultimately, to devise effective models of feedback using them.

4. OPENESSAYIST: EXTERNAL REPRESENTATIONS AND REFLECTIVE ACTIVITIES

The design of the first version of the system has focused on defining the essay summarization engine and integrating it into a working web application that supports draft submission, analysis and reporting, using multiple external representations.

[3] In the actual process, we are in fact ranking *lemmas* (the canonical form of a set of words) rather than their inflected forms in the surface text. For brevity's sake, we will keep the terms 'words' and 'key words' in this document.

At the front-end level, the instructional interactions have been deliberately limited to fairly unconstrained forms, leading the system towards a more "explore and discover" environment. Our aim was to establish a space where emerging properties of the interventions under investigation (i.e. using summarization techniques for generating formative feedback) could be discovered, explored and integrated into the design cycles in a systematic way, contributing to both the end-product of the design cycle (the system itself) and to its theoretical foundations.

Several external representations have been designed and deployed in the system, reporting the different elements described above in different ways, trying to highlight such properties in the current essay (or, in changes over successive drafts).

The main view of the system is a mash-up of the re-structured raw text, highlighting many of the features extracted by EssayAnalyser in context, using a combination of HTML markers and JavaScript-enabled interactive displays (Figure 2). Sentences, paragraphs and headings (as identified by EssayAnalyser) are displayed as blocks of text, with visual markers on the left-hand side indicating their diagnosed structural role (e.g. introduction, headings, conclusion, etc.). Key words and key phrases are also highlighted with specific visual markers, as with the top-ranked key sentences.

A control-box allows the student to change the visibility of selected elements of the essay: show/hide specific structural components (e.g. only show the introduction), key words (or user-defined categories, see below), top-ranked sentences, etc. (Figure 3).

The intended purpose of this dynamic essay representation is to attract the attention of the student away from the surface text to issues at a more structural level that might become apparent once an alternative viewpoint is considered.

For example, if confidence levels were low in the structural recognition of an introduction, the visual indicator would reflects that degree of (un)certainty about their exact role of this paragraph, requiring the student to reflect on his intention (or on the fact that an introduction might be missing in the essay or seems to be too long or too short).

Similarly, the highlighting of key words and key phrases, in context within the essay, is intended to trigger reflection on their occurrence within the text. Its purpose is different from a dedicated external representation of the key words as such (Figure 4), where the focus is more on individual terms, and on their relative importance in the essay (as indicated by their centrality score or frequency in the surface text). In the mash-up view, the key word centrality score is played down (we do not represent any attribute other than its identification as a key word) while we try to focus on whether key word *dispersion* across the essay might help identify the flow of ideas and arguments.

To complement the main mash-up view and to alleviate potential overload, we are also designing and deploying ad-hoc external representations on specific aspects of the summarization.

For example, we are exploring whether more compact representations of the dispersion of key words across the essay (Figure 5) might provide a more suitable ground for insight into its meaning. In this graph, each key word (or category of key words, if they have been defined) is plotted on a scale showing the flow of the essay (the figure uses words on the x-axis but sentences and paragraphs can also be used as units). By adding on the scale markers for the introduction, the conclusion (or any other structural elements), the student has immediate access to the overall flow of key words across the text and within specific parts of it: patterns of occurrence or omission might provide opportunity to detect an overlooked mistake (e.g. what can be said about the fact that "learning resource", ranked as a top key word by the system, only occurs in the first few paragraphs of the essay?).

On a more experimental approach, we are also exploring the possibility of visually exploiting the networks that constitute the core internal representation of the key word and key sentence extraction, using various visualization tools (e.g. force-directed graph, adjacency matrix). A case for their informational and – more importantly – formative values remains to be made.

However, we are also arguing that, to help students explore the significance of summarization elements in their essay, visualization on its own will not be enough. Support for reflective *action* is needed to resolve a key question students are likely to ask: "what are the key words (and key sentences) and how do they help me?"

Let's consider the key words. In the current version of the system, key words are presented in a very simple fashion (Figure 4): ranked by their centrality score and by their dimension (i.e. bigrams, trigrams and so on). This is a reflection of the domain-independent, data-driven design approach followed so far; key words are derived on the basis of co-occurrence, i.e. identity relation, not on the basis of semantic relations such as synonymy or hyponymy.

We can therefore have situations, as in Figure 4, where key words such as "learning experience" and "study experience" both occur as distinct bigrams, whereas, for the student who used them, they might mean very similar things. More fine-grained approaches could be implemented in EssayAnalyser to address such situation at detection level, but, ultimately, the *intention* of the student is the only safe ground for deciding on the usage of both terms. Hence the need to support some user interaction with the system, especially if it can act as a reflective scaffold.

A first example of support for reflective action is made available to the students immediately after a draft has been analyzed by the system: to let them organize key words according to their own schema, using as many categories as they wish or need (see Figure 6). This serves two purposes: it helps the students to reflect on the content of the essay and helps the system to adapt the content of every external representation accordingly, by clustering key words together (as seen in Figure 5).

Another key-word-related activity relies on the fact that a decision is made by the system on what constitutes a key word, a decision that might be at odds with the intention of the student. So we are offering the possibility for students to define – or select – their own key words. With the extraction process deriving a centrality score and frequency count for every unique word in the text, the student's decision to flag a word as a key word can be matched with that information, encouraging her to reflect on why it might be that the words she thinks should be key words are not being recognized by the system as such.

5. CONCLUSION

The first phase of the design of OpenEssayist, as reported in this paper, has focused on devising a range of external representations on the various elements that the summarization engine is extracting, notably key words, key sentences and the structural role of paragraphs in the essay.

We have implemented a working prototype that delivers a fairly unconstrained, unstructured exploration of these elements, The

drive of our design approach has been to consider how these elements, either separately or combined, would create a space where students (and researchers) could discover emerging properties of the essay, triggering deeper reflection on their own writing.

Our objective is now to consider how we structure these reflective episodes for support within the system, and how we design dedicated reflective activities that will prove to deliver formative feedback for students.

Our work is continuously focusing on three parallel but inter-connected lines of experimentation and evaluation:

1) improve the different aspects of the summarization engine;
2) experiment with it on various corpora of essays to identify trends and markers that could be used as progress and/or performance indicators (Field *et al.* 2013);
3) refine the educational aspect of the system, identify possible usage scenarios (Alden *et al.* 2013), test pedagogical hypotheses and models of feedback.

At the time of writing, several usability/desirability inspection sessions are underway, using both semi-structured walkthrough protocols in a usability lab and self-guided remote sessions with students from the last presentation of the H810 module. Part of the aim of these empirical studies is to identify tutorial strategies to be used to scaffold the student's exploitation of the system.

Finally, we are planning two empirical educational evaluations of OpenEssayist in an authentic e-learning context, to take place in September 2013 and February 2014. All students enrolled on two different Master's degree modules will be offered access to the system for two of the module's assignments and encouraged to submit multiple drafts of their essays. In-system data collection, post-module surveys, and interviews with selected participants and their tutors will give us valuable information on their learning experience with the system.

ACKNOWLEDGEMENTS

This work is supported by the Engineering and Physical Sciences Research Council (EPSRC, grant numbers EP/J005959/1 & EP/J005231/1).

REFERENCES

Alden, B., Van Labeke, N., Field, D., Pulman, S., Richardson, J. T. E., and Whitelock, D. (2013). Using student experience to inform the design of an automated feedback system for essay answers. In *Proceedings of the 2013 International Computer Assisted Assessment Conference* (CAA'13, Southampton, UK). pp. to appear.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Cambridge, MA: O'Reilly Media, Inc.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1), pp. 107–117.

Ferrer i Cancho, R., and Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 268(1482), pp. 2261–2265.

Field, D., Richardson, J. T. E., Pulman, S., Van Labeke, N., and Whitelock, D. (2013). Reflections on characteristics of university students' essays through experimentation with domain-independent natural language processing techniques. In *Proceedings of the 2013 International Computer Assisted Assessment Conference* (CAA'13, Southampton, UK). pp. to appear.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* 40(1), pp. 35–41.

Lloret, E., and Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review* 37(1), pp. 1–41.

Mihalcea, R., and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of Empirical Methods in Natural Language Processing* (EMNLP'04, Barcelona, Spain). , pp. 404–411.

Nelson, M. M., and Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science* 37(4), pp. 375–401.

O'Rourke, S. T., and Calvo, R. A. (2009). Analysing Semantic Flow in Academic Writing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (AIED'09, Brighton, UK). IOS Press, pp. 173–180.

Pickard, M. J. (2007). The new Bloom's taxonomy: An overview for family and consumer sciences. *Journal of Family and Consumer Sciences Education* 25(1), pp. 45–55.

Rowntree, D. (1987). *Assessing Students: How Shall We Know Them?* London: Kogan Page.

Southavilay, V., Yacef, K., Reimann, P., and Calvo, R. A. (2013). Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (LAK'13, Leuven, Belgium). ACM, pp. 38–47.

Villalon, J., Kearney, P., Calvo, R. A., and Reimann, P. (2008). Glosser: Enhanced Feedback for Student Writing Tasks. In *Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies* (ICALT'08, Santander, Spain). IEEE Press, pp. 454–458.

Villiot-Leclercq, E., Mandin, S., Dessus, P., and Zampa, V. (2010). Helping Students Understand Courses through Written Syntheses: An LSA-Based Online Advisor. In *Proceedings of the 10th International Conference on Advanced Learning Technologies (ICALT)* (ICALT'10, Sousse, Tunisia). IEEE Press, pp. 341–343.

Wade-Stein, D., and Kintsch, E. (2004). Summary Street: Interactive Computer Support for Writing. *Cognition and Instruction* 22(3), pp. 333–362.

Whitelock, D. (2010). Activating Assessment for Learning: Are We on the Way with Web 2.0? In *Web 2.0-Based E-Learning: Applying Social Informatics for Tertiary Teaching*, eds. Mark J.W. Lee and Catherine McLoughlin. Hershey, PA: IGI Global pp. 319–342.

APPENDIX

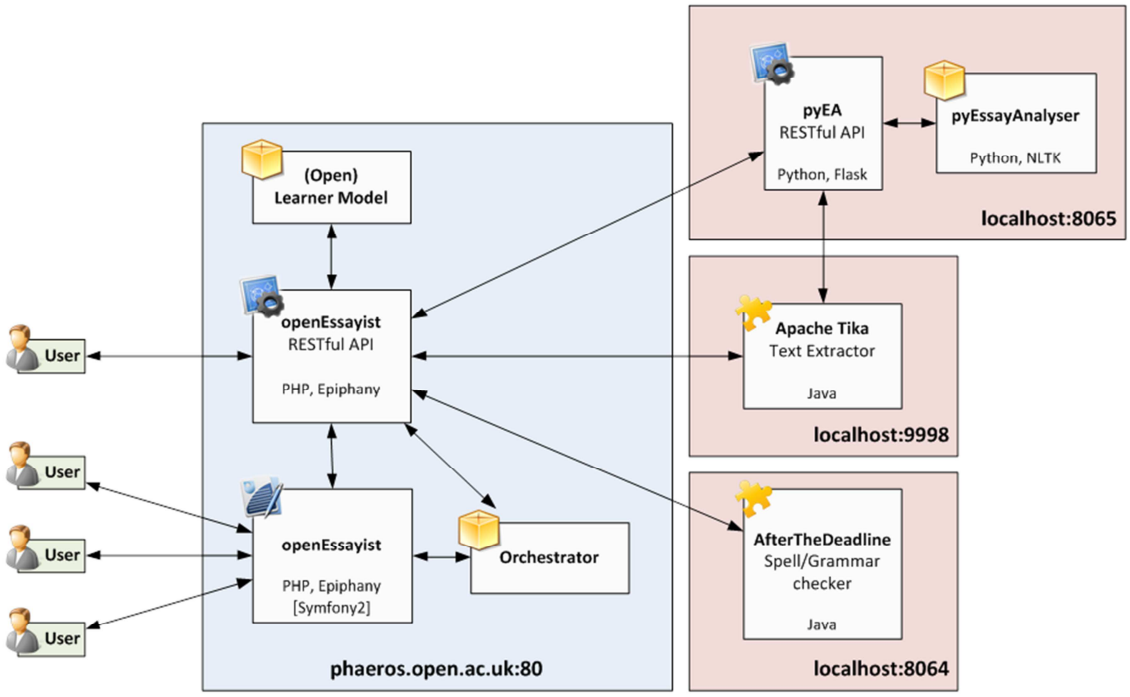

Figure 1. Architecture of OpenEssayist

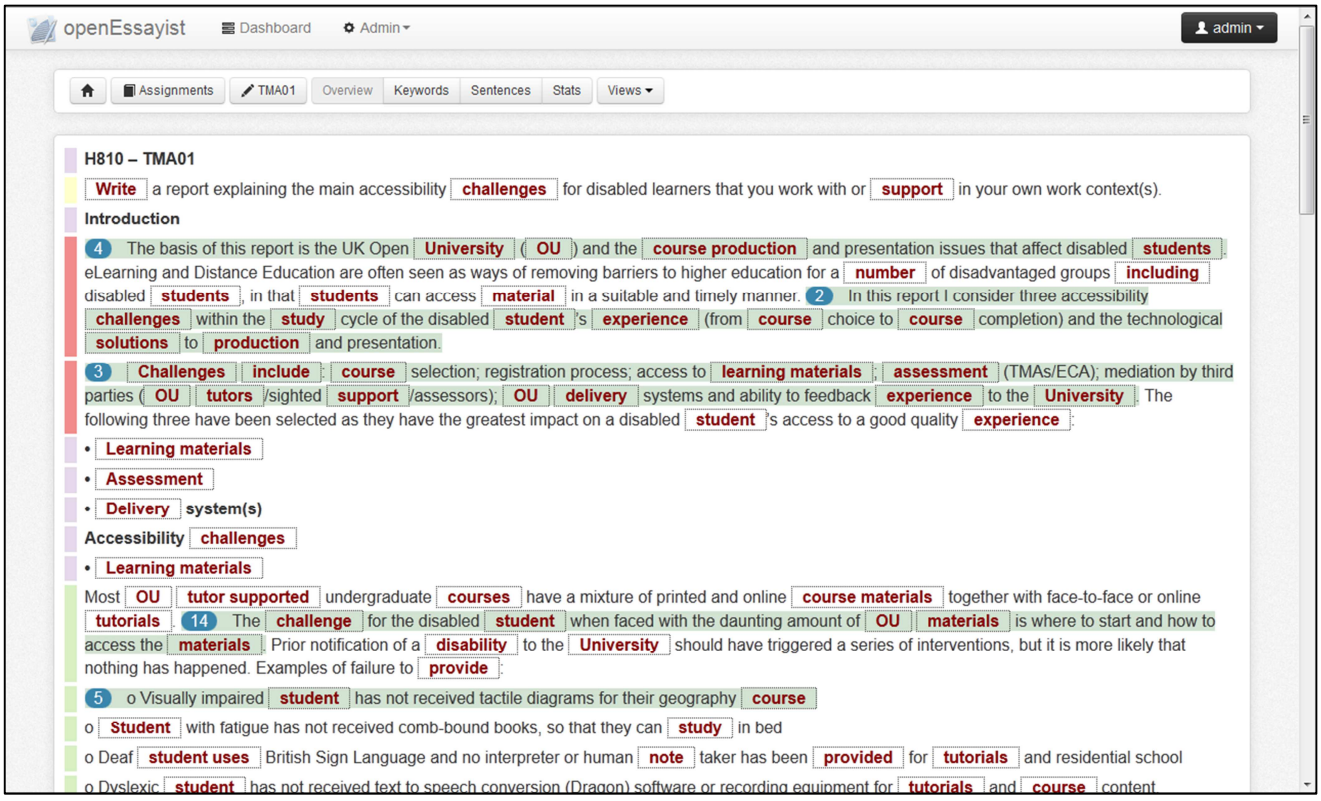

Figure 2. Key words, phrases and sentences visualized in the essay context. Sentences in light-grey (green) background are key sentences as extracted by the EssayAnalyser (the number indicates its key-ness ranking). Key words and key phrases are indicated in bold (red) and boxed.

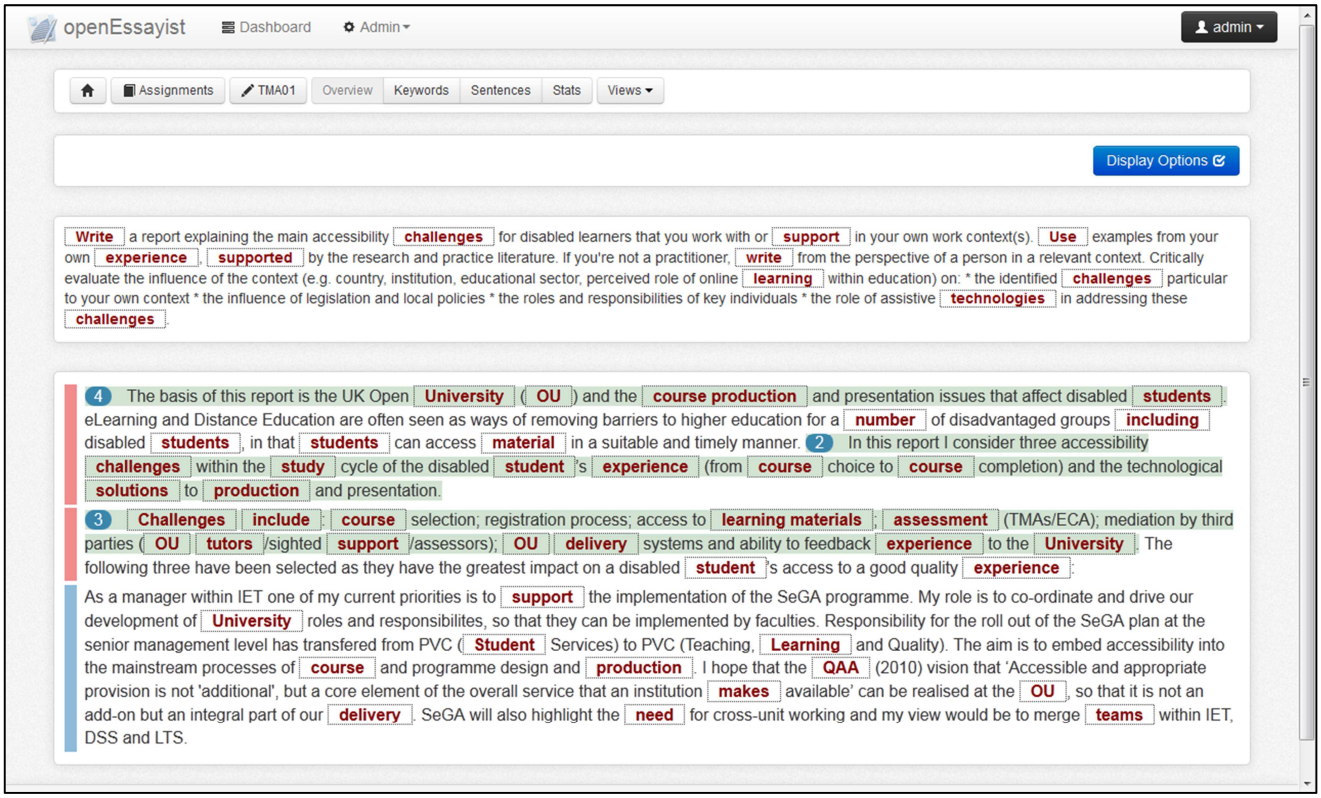

Figure 3. The structural elements of the essay can be used jointly with the key word extraction to highlight relevant information within specific parts of the essay, here in both introduction and conclusion (and the assignment question).

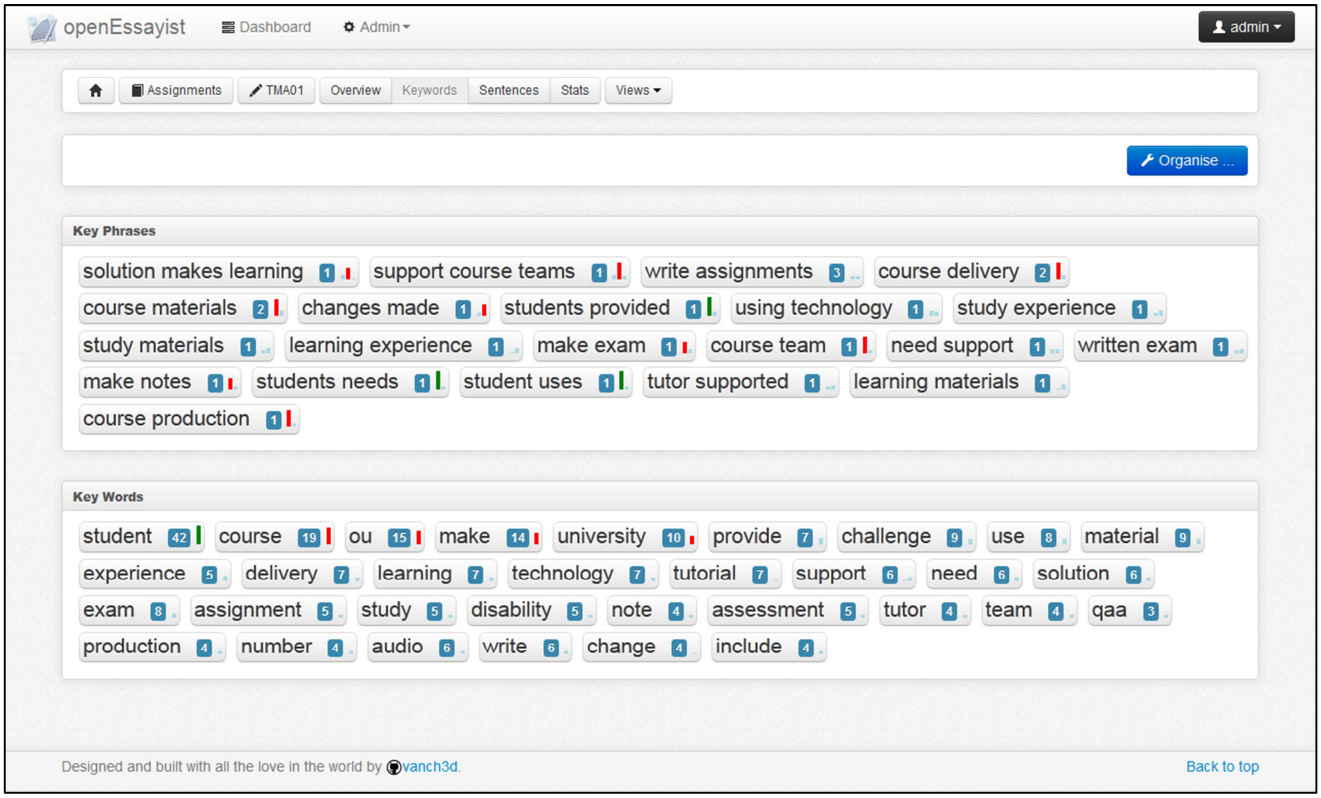

Figure 4. Key words and phrases as separate lists.

Figure 5. Dispersion of key words across the essay.

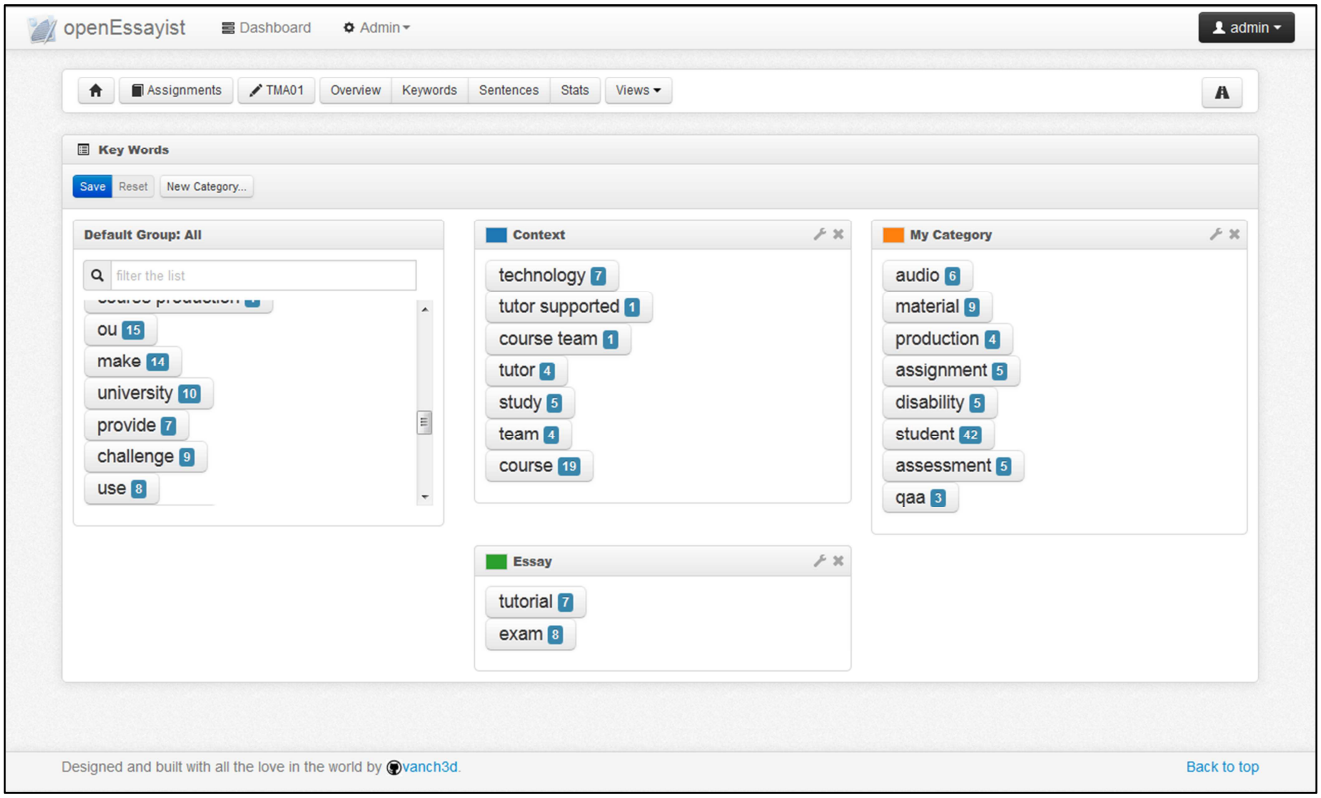

Figure 6. Key words extracted by the systems are re-organized by the students, using their own categories