Open Research Online

The Open University's repository of research publications and other research outputs

Visual analytics of academic writing

Conference or Workshop Item

How to cite:

Simsek, Duygu; Buckingham Shum, Simon; De Liddo, Anna; Ferguson, Rebecca and Sándor, Ágnes (2014). Visual analytics of academic writing. In: Proceedings of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14, pp. 265–266.

For guidance on citations see [FAQs](#).

© 2014 ACM

Version: Accepted Manuscript

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.1145/2567574.2567577>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

Duygu Simsek¹, Simon Buckingham Shum¹, Anna De Liddo¹,
Rebecca Ferguson², Ágnes Sándor³

Walton Hall, Milton Keynes, MK7 6AA, UK
 {firstname.lastname}@open.ac.uk

agnes.sandor@xrce.xerox.com

One of the key innovations of the KSV was the use of flexible thresholds in the creation of network representations. This is what allowed us to create visualizations of LSA-based representations of texts. Rather than attempting to provide a two-dimensional layout based on the first few dimensions resulting from the matrix decomposition used in LSA, our approach has been to determine the similarities between documents based on the cosines between the vectors representing documents. A graph is then created in which the nodes correspond to the documents and the edges correspond to the LSA-based similarities between them. A force-directed layout NOVELTY then applied to the graph such that the positions of nodes CONTRAST dimensional representation minimize the distortion of the (very low dimensional) representation. This representation of a maximally connected graph typically lacks clarity, and in typical cases where there are tens or hundreds of nodes the graph is essentially unintelligible due to the large number of edges.

XIP's raw output is a semantically tagged file suitable for subsequent machine analysis. While such plain textual output is well suited for researchers to analyse manually, or with other tools, this is not a form which could be usefully or attractively

presented back to either learners, educators seeking to assess their progress, or to other kinds of information analyst for whom this work is relevant. Therefore the XIP Dashboard, a visual analytics on XIP output, has been implemented to solve this problem.

3. THE XIP DASHBOARD

The XIP dashboard has three parts, each of which visualises the XIP output in a different way. This paper illustrates one example visualisation (Figure 2), for other visualisations and detailed information please see the demo video and our previous paper [5].

As shown in Figure 2, the XIP dashboard consists of a bubble chart which displays the occurrence of scholarly papers on specific topics, filtered by user-selected concepts. As shown by the colour spectrum at the top, saturation represents the total number of papers mentioning the selected concepts as salient (the darker the bubble, the higher the number of papers); and while the size of the bubble represents the 'density' of the concept in the paper based on the number of XIP classified sentences in which it occurs (the bigger the bubble, the denser the use of the concept). When a user mouse overs a concept bubble, it displays a pie chart showing the relative distribution of rhetorical types. For example, it could display the total number of contrasting idea statements made in relation to the selected concept within papers. When a user selects a segment of the pie, these sentences are listed. This then enables users to see the full paper that mentions the selected sentence with all other highlighted salient sentences within the paper.

In the conference demo, we will be showing the XIP Dashboard visualisations with a specific test corpus. The dataset we will be using is the Learning Analytics and Knowledge (LAK) dataset. This dataset is published by the Society for Learning Analytics Research (SoLAR) [6]. Prior to implementation of the XIP Dashboard, the LAK Dataset papers were analysed through XIP. This produced a semantically tagged output file of each paper for subsequent machine analysis. Output files were then imported into a relational MySQL database and the user interface was implemented using PHP and JavaScript, making use of Google Chart Tools for the interactive visualisations.

Current version of the XIP dashboard is built from locally stored data. All the LAK dataset papers were rendered by the XIP tool's developers, who then shared the output files with us. By the time of the LAK '14 conference, we plan to be able to build an architecture where XIP can be called as a web service.

Therefore in the conference demo, the XIP dashboard will be accessing XIP services remotely through its API.

4. CONCLUSION

Authors use specific discipline-independent argumentative patterns, called 'metadiscourse markers'. These help readers to identify significant claims and arguments made by the authors. However the identification of such patterns can be difficult for the human readers. As computational techniques are maturing, NLP tools are becoming available to automatically identify these markers. However the results are not very user-friendly. The XIP Dashboard has been implemented to address these problems and provide user-friendly visual analytics of academic writing.

In the LAK '14 conference demo session, we will demonstrate the different types of visualisations offered by the XIP dashboard on a specific test corpus (LAK dataset). In the live demo, we will also be explaining how the XIP dashboard can be used to improve the quality of the academic writing process.

The current version of the XIP dashboard restricts itself with enabling its users to make sense of the published literature. Our longer term goal is using the XIP Dashboard as a formative assessment tool for one's own writing, in which users will get visual analytics results of their own contributions.

5. REFERENCES

- [1] T. Thonney, "Teaching the Conventions of Academic Discourse," *Teaching English in the Two Year College*, vol. 38, p. 347, 2011.
- [2] S. North, "Disciplinary variation in the use of theme in undergraduate essays," *Applied Linguistics*, vol. 26, pp. 431-452, 2005.
- [3] K. Hyland and P. Tse, "Metadiscourse in academic writing: A reappraisal," *Applied Linguistics*, vol. 25, pp. 156-177, 2004.
- [4] S. Ait-Mokhtar, J.-P. Chanod, and C. Roux, "Robustness beyond shallowness: incremental deep parsing," *Natural Language Engineering*, vol. 8, pp. 121-144, 2002.
- [5] D. Simsek, S. Buckingham Shum, A. Sandor, A. De Liddo, and R. Ferguson, "XIP Dashboard: visual analytics from automated rhetorical parsing of scientific metadiscourse," 2013. <http://oro.open.ac.uk/37391>
- [6] Society for Learning Analytics Research. (2013) <http://www.solaresearch.org/resources/lak-dataset/>

Figure 2 An example XIP Dashboard Visualisation