Visual Analytics of Academic Writing

Duygu Simsek[1], Simon Buckingham Shum[1], Anna De Liddo[1],
Rebecca Ferguson[2], Ágnes Sándor[3]

[1] Knowledge Media Institute
[2] Institute of Educational Technology
The Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
{firstname.lastname}@open.ac.uk

[3] Parsing & Semantics Group
Xerox Research Centre Europe
6 chemin Maupertuis, F-38240 Meylan
France
agnes.sandor@xrce.xerox.com

ABSTRACT

This paper describes a novel analytics dashboard which visualises the key features of scholarly documents. The Dashboard aggregates the salient sentences of scholarly papers, their rhetorical types and the key concepts mentioned within these sentences. These features are extracted from papers through a Natural Language Processing (NLP) technology, called Xerox Incremental Parser (XIP). The XIP Dashboard is a set of visual analytics modules based on the XIP output. In this paper, we briefly introduce the XIP technology and demonstrate an example visualisation of the XIP Dashboard.

1. INTRODUCTION

As literatures expand and fields become increasingly multidisciplinary, it is common for researchers to find themselves navigating papers produced in a variety of research fields; some of which are written according to norms and conventions that are different from those of their 'home' disciplines. When engaging with this literature, a core competency is a 'critical mind', which includes an ability to identify when significant claims and arguments are being made in articles. The ability to decode such moves in texts is essential, as is the ability to make such moves in one's own writing. For this reason, we are interested in analytics tools which help readers to make sense of the scholarly, and which provide feedback to writers, ranging from students to experienced researchers, on the quality of their own writing.

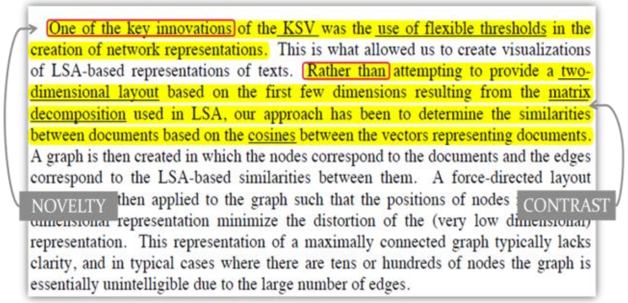
Research into academic writing draws attention to the question of whether or not there are universal conventions for such scholarly moves (see [1, 2] for example studies). Literature shows that scholarly articles have typical argument structures regardless of their discipline. Therefore, while engaging with the literature; researchers make use of specific linguistic cues in the text. These are referred to technically as 'metadiscourse' markers. Metadiscourse is an important element of a document. It allows readers to make sense of a text; understand viewpoints, arguments and claims; thus engaging with the author's intended meaning [3]. Authors signal argumentative moves by using well-established patterns known as 'metadiscourse markers'. Such elements inside the text are discipline-independent clear forms that are actually identifiable. However, identifying these can be challenging for literature reviewers.

Although identifying such argumentative elements within the documents can be challenging for the human readers, 'metadiscourse markers' can be automatically identified. An increasing number of electronic publications through electronic library databases has prompted an increase in research and development in the field of machine processing. Research into this field has been providing more effective ways of navigating the literature and helping readers to engage with ideas. And work in natural language processing (NLP) technology has made it possible to detect 'metadiscourse markers' automatically.

In this paper we will introduce an NLP tool, called Xerox Incremental Parser (XIP). XIP's discourse analysis module identifies rhetorically salient sentences on the basis of the 'metadiscourse markers'. We, then demonstrate the XIP dashboard, which builds on the XIP output. It is designed to help readers to make sense of the scholarly papers more rapidly and conveniently; and to assess the current state of the art in terms of trends, patterns, gaps and connections.

2. XEROX INCREMENTAL PARSER

Xerox Incremental Parser (XIP) is an NLP tool which carries out automated metadiscourse analysis of scholarly text documents. It aims at highlighting the main research issues that the article handles. The idea behind the tool is that "rhetorical moves can be detected from the author's language use" [4]. Therefore, XIP highlights metadiscourse that conveys the author's rhetorical strategy and labels the rhetorical functions such as: summary, background knowledge, contrasting ideas, novelty, surprising idea, and open question. For example, in the following paragraph of text (Figure 1), XIP extracted two salient sentences (highlighted in yellow) and identified rhetorical type of these sentences (as Novelty and Contrast) on the basis of the metadiscourse markers (shown in red). In addition, XIP identifies *concepts* (nouns and noun phrases) within rhetorically salient sentences, which indicate the topics dealt with by the sentence (underlined in the Figure 1).



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 then applied to the graph such that the positions of nodes representation minimize the distortion of the (very low di 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.

Figure 1 An example XIP analysis

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.

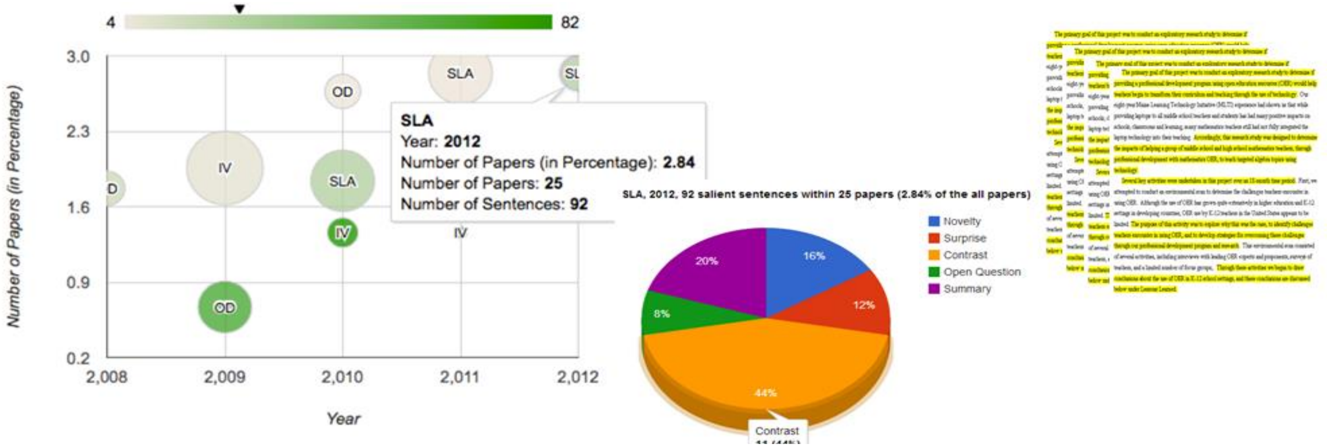
[2] S. North, "Disciplinary variation in the use of theme in undergraduate essays," *Applied Linguistics*, vol. 26, pp. 431-452, 2005.

[3] K. Hyland and P. Tse, "Metadiscourse in academic writing: A reappraisal," *Applied Linguistics*, vol. 25, pp. 156-177, 2004.

[4] S. Aït-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/



The primary goal of this project was to conduct an exploratory research study to determine if providing a professional development program using open education resources (OER) would help teachers begin to transform their curriculum and teaching through the use of technology. Our eight-year Maine Learning Technology Initiative (MLTI) experience had shown us that while providing laptops to all middle school teachers and students has had many positive impacts on schools, classrooms and learning, many mathematics teachers still had not fully integrated the laptop technology into their teaching. Accordingly, this research study was designed to determine the impacts of helping a group of middle school and high school mathematics teachers, through professional development with mathematics OER, to teach targeted algebra topics using technology.

Several key activities were undertaken in this project over an 18-month time period. First, we attempted to conduct an environmental scan to determine the challenges teachers encounter in using OER. Although the use of OER has grown quite extensively in higher education and K-12 settings in developing countries, OER use by K-12 teachers in the United States appears to be limited. The purpose of this activity was to explore why this was the case, to identify challenges teachers encounter in using OER, and to develop strategies for overcoming these challenges through our professional development program and research. This environmental scan consisted of several activities, including interviews with leading OER experts and proponents, surveys of teachers, and a limited number of focus groups. Through these activities we began to draw conclusions about the use of OER in K-12 school settings, and these conclusions are discussed below under Lessons Learned.

Figure 2 An example XIP Dashboard Visualisation