ISSN 1744-1986

Technical Report N° 2005/10

Can Online Learning Materials Improve Student Access to Digital Libraries?

Alistair Willis

17th June 2005

Department of Computing
Faculty of Mathematics and Computing
The Open University
Walton Hall,
Milton Keynes
MK7 6AA
United Kingdom

http://computing.open.ac.uk

Can Online Learning Materials Improve Student Access to Digital Libraries?

Alistair Willis
Department of Computing
Faculty of Maths and Computing
The Open University
Milton Keynes, UK

A.G.Willis@open.ac.uk

Abstract

We present preliminary investigations into how text alignment techniques can be used to align the content of undergraduate textbooks against the journal papers from which they were developed, or which may be recommended further reading. We propose that such methods could be used to improve student access to academic material, particularly in distance learning environments.

We show that our initial techniques determine which passages of textbooks align against appropriate academic documents, and consider what techniques might be needed for the required finer grained alignment.

1 Introduction

Over the course of an undergraduate degree, a student may be expected to move from learning materials (for example, textbooks) provided and written by the host institution, to the literature published within the domain of study. Particularly towards the end of the undergraduate studies, an appreciation and understanding of the state of the art in their chosen domain represents the transition from a student of the subject to a practitioner.

However, students may find the transition from learning materials to domain texts difficult. The intended audiences for the two genres are very different, reflected both in the content of the writing and in the style. In terms of content, the author of (say) a journal paper is generally able to assume that the intended readership is fully *au fait* with standard techniques and concepts and the subject matter generally goes beyond what is covered in typical learning materials. The written style of the genres is also different; Lee (2001) has noted that the instructional writing and persuasive writing styles which are generally adopted in learning materials and domain texts respectively form distinguishable genres.

We are investigating how techniques for aligning comparable texts (section 2) might assist students who wish to extend their knowledge from the textbook to the large amount of papers held online. Such techniques should enable students to navigate the world of the real academic literature more readily. The key benefit would be to demonstrate to students how the techniques being taught at the undergraduate level are applied in practice in the state of the art of their chosen field of study and, therefore, how they might use those same techniques themselves. (Kato et al., 1999) has described how computer science students often believed (incorrectly) that activities carried out in-class did not reflect the activities of professional software development. There is benefit in being able to demonstrate to students that the techniques that they are introduced to during their studies are also those that are used outside the classroom.

In fact, suggestions for suitable papers are often provided as further recommended reading. A successful system would align that content of the learning materials against the equivalent passages of the domain texts (figure 1).

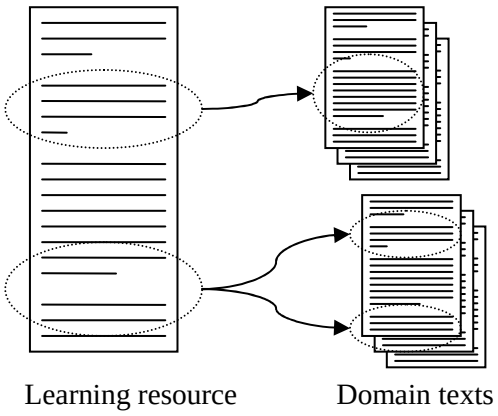

Figure 1: Align learning text content against the equivalent in the domain texts

The transition becomes necessary when the student is expected to carry out tasks at research level; the marking guidelines for Masters projects at the Open University recommend that students should quote academic texts, rather than the textbooks that they have been using prior to the project. We can therefore state the hypothesis motivating this work:

Hypothesis A student may improve his or her knowledge and understanding of a subject area by having key concepts aligned between learning materials and the associated domain texts.

2 Aligning Text Across Genres

The elearning aid that we are proposing requires the alignment of monolingual texts that are comparable but nonparallel, that is, which share common content but where there is not usually a sentence to sentence mapping.

Our experiments have been using a course textbook from the biological sciences for the Open University and comparisons run against two related texts recommended by the document's author. The following two sentences (with closely related content) indicate some of the differences in the styles of writing:

> When given to young female mice... the hormone... enables them to breed earlier than the controls. *(Textbook)*

> Treatment of mice with leptin accelerates the maturation of the female reproductive tract and leads to an earlier onset of the oestrous cycle and reproductive capacity. *(Nature article)*

The passage from the text book has been reduced, partly for reasons of space, but also to indicate that the salient terms do not necessarily form a single block of text. In addition, the journal article's author is able to expect his audience to be more confident with the technical language than the author of the textbook.

Barzilay and Elhdad (2003) have raised the key problems facing alignment between comparable rather than parallel corpora:

1. The relevant sentences are less likely to be in a continuous block, and
2. the content of interest may be more tangential than is the case with strictly parallel texts.

3 The Alignment Tasks

Our initial results have addressed only a subproblem of the full alignment task (the limitations of this are discussed in section 6):

Problem Given a pair of texts, $Text_1$ and $Text_2$, determine which passages of $Text_1$ and $Text_2$ contain similar content.

To run the specific experiments that we require here, we look at the question:

Task Given $Text_1$, a university course textbook, and $Text_2$, an academic text that covers the same content, determine the subsets, $t_1 \subseteq Text_1$ and $t_2 \subseteq Text_2$ that maximise the similarity of t_1 and t_2.

We have carried out two alignment tasks. The content of a second year undergraduate textbook in mammal physiology was been aligned against:

1. a journal paper (from *Nature*) whose content was very close to a section of the textbook (the textbook discussed that issue specifically),
2. a journal paper (*Journal of Anatomy*) that was broadly related to the content of the textbook, and was suggested by the course notes author as recommended further reading.

A textbook from a scientific discipline was chosen to reflect our interest in differing terminologies, reflected in the example sentences.

4 Matching Algorithm

The algorithm that we use is a greedy set coverage that selects sentences from each of the two texts until no further improvement is seen. The algorithm contains elements of the textTiling algorithm (Hearst, 1997) with the important difference that where textTiling uses similarity between paragraphs as a measure, our algorithm uses similarity across the two texts.

To measure the similarity between two blocks of text, S_1 and S_2, we use the cosine measure:

$$comp(S_1, S_2) = \frac{\sum_{i=1}^{n} w_{1i} w_{2i}}{\sqrt{\sum_{i=1}^{n} w_{1i}^2} \sqrt{\sum_{i=1}^{n} w_{2i}^2}}$$

where $\{t_1, t_2, \ldots, t_n\}$ denotes the set of terms which occur in S_1 and S_2, and w_{1i} and w_{2i} are the weights assigned to the term t_i in S_1 and S_2 respectively.

Input Two sets of sentences, TB and DT, where TB represents a textbook and DT an associated domain text.

1. Remove stop words from the sentences in TB and DT and perform stemming (Porter, 1980) on the remaining terms
2. Choose an initial pair of nonempty sets $tb \subset TB$ and $dt \subset DT$
3. **repeat**
4. add the sentence from TB to tb or from DT to dt that maximises $comp(tb, dt)$
5. **until** C converges on a maximum
6. **return** tb and dt

Figure 2: The matching algorithm

In this case, the weights w_{ij} assigned to each term t_j are simply the number of occurrences of the t_j in the sentences in S_i. As with Hearst, we found that term frequencies seem to give better results than including an inverse document frequency weighting.

To make the initial selection of subsets[1] (step 2) a pairwise comparison of the sentences in the two sets TB and DT is made and the pair of sentences that maximises the comparison function is chosen. Interestingly, in the initial selection only, a better initial pair is chosen by using an inverse document frequency in the weighting function (Manning and Schütze, 1999).

The body of the algorithm (step 4) proceeds by extending *either* tb by a member of TB or dt by a member of DT, whichever results in the largest increase in $comp(tb, dt)$. The process continues until $comp(tb, dt)$ reaches a maximum, although the curves are not completely smooth; we obtained good results using a window so that the maximum is considered passed if 6 of the 8 additions decreases $comp(tb, dt)$.

For the textbook with the closely related document, a clear maximum was reached after approximately 150 sentences were added altogether to tb and dt(figure 3). When the same algorithm was run on the text book and the loosely related text, the comparison function also peaked, but after the addition of fewer sentences, and with a much lower similarity measure (figure 4).

[1]in this case, each subset is only one sentence

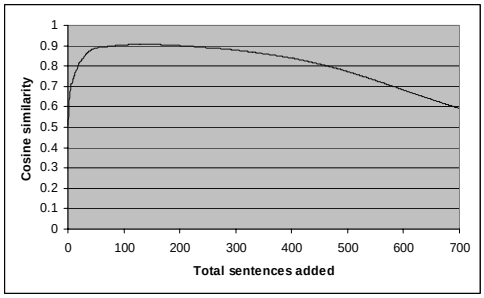

Figure 3: Closely related documents

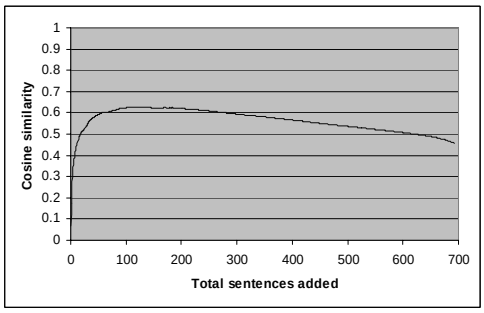

Figure 4: Loosely related documents

5 Results

The algorithm marks out those regions of the textbook whose content reflects that of the closely associated domain text. Figure 5 shows the sentences that are added by this method up to the point of convergence (approximately 155 sentences added). The graph shows which sentences were added from the text book. For example, a point on the graph at (30, 500) indicates that the sentence added on the 30th iteration of the algorithm was the 500th sentence from the textbook.

With a small number of outliers, the sentences are from the region sentence470 to sentence590, the approximate region where the textbook discusses the same content as the academic paper. Note that the region is clearly demarked, although only about half the sentences in that region are selected. The selection of sentences from the domain text forms a much less clearly defined block, reflecting the greater depth and smaller breadth of the domain text.

When the experiment was run to compare the textbook and the more loosely related paper, the sentences from the textbook were also relevant but formed a less clearly defined block (as we would probably expect). We are currently developing a method for more accurately mea-

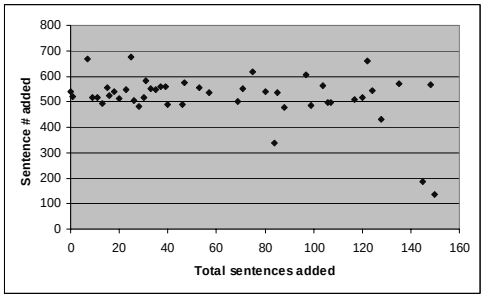

Figure 5: Sentences selected from textbook up to convergence

suring the technique's recall and accuracy.

While this algorithm successfully highlights section of the learning material whose content is closely related to that of the source texts, it is only a first step towards our aim of aligning blocks of the two texts by common content. We outline in the next section how we are addressing this task.

6 Further work

These results indicate that the blocks of text sharing common content can be found, but having found such blocks, we have not yet addressed the problem of aligning the content within those blocks. In fact, because there has been no alignment at this lower level, the matching algorithm can be thought of has having matched on *context*, rather than on *content*.

As the example sentences from section 2 illustrate, additional resources may be required for the lower level matching. Because the documents we are dealing with have widely divergent vocabulary, the words alone are unlikely to provide adequate information for the alignment. And as we have also seen, there is unlikely to be a direct sentence to sentence translation.

The longer term aim of our research is to demonstrate how additional domain resources, such as ontologies or taxonomies, might be used to go beyond contextual similarity and reason about the content of such documents. Publicly available ontologies are becoming available which we believe can be used to assist with aligning the terms between a textbook and associated domain texts.

7 Discussion and Conclusions

The primary objective of this research is to support students learning at a distance. The transition from learning materials to domain texts can be difficult, and can be more so in a learning environment with reduced teacher/student interaction. In addition, learning materials have generally not been written primarily to provide access to domain texts, and in fast moving areas of study such as computing the content of the learning materials may lag significantly behind the published state of the art. We believe that techniques to improve access to academic digital libraries using existing learning materials are a valuable addition to elearning environments.

A second issue relevant to the development of elearning environments is the maintenance of reusable learning components. An important aspect of these which has been identified is the "links that support the learning objective" (Leeder and Garrud, 2003, Universities Collaboration in Elearning). Techniques allowing *ad hoc* alignment between comparable texts should assist the process of creating and maintaining links between textbooks, designed solely to address learning objectives, and academic writing which supports those objectives.

References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in NLP*, July.

Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hiroshi Kato, Akiko Ide, and Hideyuki Suzuki. 1999. Providing access from classroom to community of practice through the Internet: The development of AlgoArenaWWW. In Geoff Cumming, Toshio Okamoto, and Louise Gomez, editors, *Advanced Research in Computers and Communications in Education*, pages 398–401. IOS Press.

David YW Lee. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

Dawn Leeder and Paul Garrud. 2003. Reusability: cultural, political and (occasionally) technical issues. Submitted for 'Shock of the Old 3', University of Oxford, April.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14:130–137.