In-Browser Summarisation: Generating Elaborative Summaries Biased Towards the Reading Context

Stephen Wan and Cécile Paris
ICT Centre*
CSIRO
Locked Bag 17, North Ryde, Sydney
NSW 1670, Australia
Firstname.Lastname@csiro.au

Abstract

We investigate *elaborative summarisation*, where the aim is to identify supplementary information that expands upon a key fact. We envisage such summaries being useful when browsing certain kinds of (hyper-)linked document sets, such as Wikipedia articles or repositories of publications linked by citations. For these collections, an elaborative summary is intended to provide additional information on the linking anchor text. Our contribution in this paper focuses on identifying and exploring a real task in which summarisation is situated, realised as an *In-Browser* tool. We also introduce a *neighbourhood* scoring heuristic as a means of scoring matches to relevant passages of the document. In a preliminary evaluation using this method, our summarisation system scores above our baselines and achieves a recall of 57% annotated gold standard sentences.

1 Introduction

It has long been held that a summary is useful, particularly if it supports the underlying task of the user — for an overview of summarisation scenarios see Spark Jones (1998). For example, *generic* (that is, not query-specific) summaries, which are often indicative, providing just the gist of a document, are only useful if they happen to address the underlying need of the user.

In a push to make summaries more responsive to user needs, the field of summarisation has explored the overlap with complex question-answering research to produce query-focused summaries. Such work includes the recent DUC challenges on query-focused summarisation,[1] in which the user needs are represented by short paragraphs of text written by human judges. These are then used as input to the summarisation process. However, modelling user needs is a difficult task. DUC descriptions of information needs are only an artificial stipulation of a user's interest.

In this work, we propose a tool built into an internet browser that makes use of a very simple heuristic for determining user interest.[2] The basic premise of the heuristic is that the text currently being read provides an approximation of the current user interest. Specifically, as a user reads a sentence, it potentially represents a fine-grained information need. We identify the sentence of interest without complex methods, relying instead on the user to move the mouse over the anchor text link to request a summary of the *linked* document, thus identifying to the browser plug-in which sentence is now in focus.

To generate the summary, the whole document, specifically the *linking* sentence that contains the anchor text, serves as the *reading context*, a potential indicator of the user interest. An example of the current output on Wikipedia text is presented in Figure 1. It shows an *elaborative summary* of a document about the Space Shuttle Discovery expanding on the content of the linking sentence. In this case, it gives further information about a space walk in which the shuttle was repaired inflight.

Our summarisation tool, the In-Browser Elabora-

*Information and Communication Technologies Centre

[1] http://duc.nist.gov/guidelines/2006.html

[2] We currently work with the Firefox browser.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 129–132,
Columbus, Ohio, USA, June 2008.

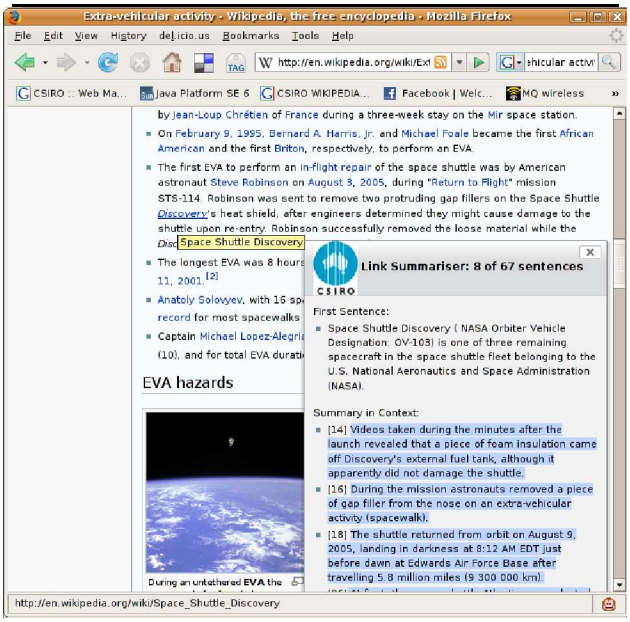

Figure 1: A summary generated when moving the mouse over the link "Discovery's" (mouse pointer omitted).

tive Summariser (IBES), complements generic summaries in providing additional information about a particular aspect of a page.[3] Generic summaries themselves are easy to generate due to rules enforced by the Wikipedia style-guide, which dictates that all titles be noun phrases describing an entity, thus serving as a short generic summary. Furthermore, the first sentence of the article should contain the title in subject position, which tends to create sentences that define the main entity of the article.

For the elaborative summarisation scenario described, we are interested in exploring ways in which the reading context can be leveraged to produce the elaborative summary. One method explored in this paper attempts to map the content of the *linked* document into the semantic space of the reading context, as defined in vector-space. We use Singular Value Decomposition (SVD), the underlying method behind Latent Semantic Analysis (Deerwester et al., 1990), as a means of identifying latent topics in the reading context, against which we compare the linked document. We present our system and the results from our preliminary investigation in the remainder of this paper.

[3]http://www.ict.csiro.au/staff/stephen.wan/ibes/

2 Related Work

Using link text for summarisation has been explored previously by Amitay and Paris (2000). They identified situations when it was possible to generate summaries of web-pages by recycling human-authored descriptions of links from anchor text. In our work, we use the anchor text as the reading context to provide an elaborative summary for the linked document.

Our work is similar in domain to that of the 2007 CLEF WiQA shared task.[4] However, in contrast to our application scenario, the end goal of the shared task focuses on suggesting editing updates for a particular document and not on elaborating on the user's reading context.

A related task was explored at the Document Understanding Conference (DUC) in 2007.[5] Here the goal was to find new information with respect to a previously seen set of documents. This is similar to the elaborative goal of our summary in the sense that one could answer the question: "What else can I say about topic X (that hasn't already been mentioned in the reading context)". However, whereas DUC focused on unlinked news wire text, we explore a different genre of text.

3 Algorithm

Our approach is designed to *select* justification sentences and *expand* upon them by finding elaborative material. The first stage identifies those sentences in the linked document that support the semantic content of the anchor text. We call those sentences justification material. The second stage finds material that is supplementary yet relevant for the user. In this paper, we report on the first of these tasks, though ultimately both are required for elaborative summaries.

To locate justification material, we implemented two known summarisation techniques. The first compares word overlap between the anchor text and the linked document. The second approach attempts to discover a semantic space, as defined by the reading context. The linked document is then mapped into this semantic space. These are referred to as the Simple Link method and the SVD method, where

[4]http://ilps.science.uva.nl/WiQA/

[5]http://duc.nist.gov/guidelines/2007.html

the latter divides further into two variants: SVD-Link and SVD-topic.

3.1 Simple Link Method

The first strategy, Simple Link, makes use of standard vector space approaches from Information Retrieval. A vector of word frequencies, omitting stopwords, is used to represent each sentence in the reading context and in the linked document. The vector for the anchor sentence is compared with vectors for each linked document sentence, using the cosine similarity metric. The highest scoring sentences are then retrieved as the summary.

3.2 Two Singular Value Decomposition (SVD) Methods

In these approaches, the semantic space of the linked document is mapped into that of the reading context. Intuitively, only those sentences that map well into the reading context space and are similar to the linking sentence would be good justification material.

To begin with, the reading context document is represented as a term-by-sentence matrix, A, where stop words are omitted and frequencies are weighted using inverse document frequency. A Singular Value Decomposition (SVD) analysis is performed (using the JAMA package[6]) on this matrix which provides three resulting matrices: $A = USV^{tr}$.

The S-matrix defines the *themes* of the reading context. The U-matrix relates the reading context vocabulary to the discovered themes. Finally, the V-matrix relates the original sentences to each of the themes. The point of the SVD analysis is to discover these themes based on co-variance between the word frequencies. If words occur together, they are semantically related and the co-variance is marked as a theme, allowing one to capture fuzzy matches between related words. Crucially, each sentence can now be represented with a vector of membership scores to each theme.

The first of the semantic space mapping methods, SVD-link, finds the theme that the anchor text belongs to best. This is done by consulting the V-matrix of the SVD analysis to find the highest scoring theme for that sentence, which we call the linking theme. Each sentence in the linked document, after mapping it to the SVD-derived vector space, is then examined. The highest scoring sentences that belong to the linking theme are then extracted.

The second method, SVD-topic, makes a different assumption about the nature of the reading context. Instead of taking the anchor text as an indicator of the user's information need, it assumes that the top n themes of the reading context document represent the user's interest. Of the linked document sentences, for each of those top n reading context themes, the best scoring sentence is extracted.

4 Evaluation

In lieu of a user-centered experiment, our preliminary experiments evaluated the effectiveness of the tool in terms of finding justification material for an elaborative summary. We evaluated the three systems described in Section 3. Each system selected 5 sentences. We tested against two baselines. The first simply returns the first 5 sentences. The second produces a generic summary based on Gong and Liu (2001), *independently* of the reading context.

4.1 Data

The data used is a collection of Wikipedia articles obtained automatically from the web. The snapshot of the corpus was collected in 2007. Of these, links from about 600 randomly chosen documents were filtered with a heuristic that enforced a sentence length of at least 10 words such that the link in the anchor text occurred after this minimum length. This heuristic was used as an approximate means of filtering out sentences where the linking sentence was simply a definition of the entity linked. In these cases, the justification material is usually trivially identified as the first sentence of the linked document. This leaves us with links that potentially require more complicated summarisation methods.

Of these cases, 125 cases were randomly selected and the linked documents annotated for varying degrees of relevancy. This resulted in 50 relevant document links, which we further annotated, selecting sentences supporting the anchor sentence, with a Cohen's Kappa of 0.55. The intersection of the selected sentences was then used as a gold standard for each test case.

[6] http://math.nist.gov/javanumerics/jama/

System	Recall	Precision
generic	0.13	0.05
SVD-topic	0.14	0.06
SVD-link	0.22	0.09
simple-link	0.28	0.11

Table 1: Recall and Precision figures for all summarisers without the first 5 sentences.

System	Recall	Precision
generic	0.27	0.04
SVD-topic	0.27	0.04
SVD-link	0.30	0.05
simple-link	0.38	0.06

Table 2: Recall and Precision figures using the *neighbourhood* heuristic (without the first 5 sentences).

4.2 Results

It is difficult to beat the first-5 baseline, which attains the best recall of 0.52 and a precision of 0.2, with all other strategies falling behind. However, we believe that this may be due to the presence of some types of Wikipedia articles that are narrow in scope and centered on specific events. For such articles, we would naturally advocate using the first N sentences as a summary.

To examine the performance of the summarisation strategies on sentences beyond the top-N, we filtered the gold standard sets to remove sentences occurring in positions 1-5 in the linked document, and tested recall and precision on the remaining sentences. This reduces our test set by 10 cases. Since documents may be lengthy (more than 100 sentences), selecting justification material is a difficult task. The results are shown in Table 1 and indicate that systems using reading context do better than a generic summariser.

Thinking ahead to the second expansion step in which we find elaborative material, good candidates for such sentences may be found in the immediate vicinity of justification sentences. If so, near matches for justification sentences may still be useful in indicating that, at least, the right portion of the document was identified. Thus, to test for near matches, we scored a match if the gold sentence occurred on either side of the system-selected sentence. We refer to this as the *neighbourhood heuristic*.

Table 2 shows the effect on recall and precision if we treat each selected sentence as defining a neighbourhood of relevance in the linked document. Again, performance on the first 5 sentences were ignored. Recall improved by up to 10% with only a small drop in precision (6%). When the neighbourhood heuristic is run on the original gold sentence set (with the first 5 sentences), recall reaches 0.57, which lies above an amended 0.55 baseline.

5 Future Work and Conclusions

We introduced the concept of a user-biased elaborative summarisation, using the reading context as an indicator of the information need. Our paper presents a scenario in which elaborative summarisation may be useful and explored simple summarisation strategies to perform this role. Results are encouraging and our preliminary evaluation shows that reading context is helpful, achieving a recall of 57% when identifying sentences that justify content in the linking sentence of the reading context. In future work, we intend to explore other latent topic methods to improve recall and precision performance. Further development of elaborative summarisation strategies and a user-centered evaluation are also planned.

References

Einat Amitay and Cécile Paris. 2000. Automatically summarising web sites: is there a way around it? In *Proceedings of the 9th international conference on Information and knowledge management*, NY, USA.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th ACM SIGIR conference*. New Orleans, USA.

Karen Spark Jones. 1998. Automatic summarizing: factors and directions. In I. Mani and M. Maybury (ed.), *Advances in Automatic Text Summarisation*. MIT Press, Cambridge MA.