SPi	Journal Code				Article ID				Dispatch: 18.04.13	CE:
	I	S	A	F	1	3	4	0	No. of Pages: 22	ME:

INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
Intell. Sys. Acc. Fin. Mgmt. (2013)
Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1340

UNDERSTAND THE GLOBAL ECONOMIC CRISIS: A TEXT SUMMARIZATION APPROACH

SHUHUA LIU[a]* AND BENOIT FAVRE[b]

[a] *Arcada University of Applied Sciences, Department of Business, IT and Media, Jan-Magnus Janssonin aukio 1, 00550 Helsinki, Finland*

[b] *LIF, Aix-Marseille Université, Parc Scientifique et Technologique de Luminy, 163, avenue de Luminy – Case 901, F-13288 Marseille Cedex 9, France*

SUMMARY

Economic crises are significant threats to macroeconomic stability. They can incur large costs and bring devastating effects on economies, with the effects often spilling over into other economies. Since 2007 we have witnessed the most severe and widely spread economic crisis since the Great Depression of the 1930s. In the meantime, a huge amount of ongoing media coverage, reporting, analysis and debate concerning the global economic crisis has been generated. In this study we explore the possibilities of applying text summarization tools to learn from text documents the various discussions surrounding the global economic crisis. Included in our analysis are blog posts and articles of highly influential economists, as well as official speeches and publications of government organizations. The ICSI-ILP extractive summarizer is applied in a large number of experiments, and the summary outputs are manually examined and evaluated. The results provide us with insights into the potential and limitations of state-of-the-art summarization systems when used to help us quickly learn and digest large amounts of textual information. The results also suggest different ways to break the limitations of text summarization technology. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: text summarization; content extraction; text analysis; economic crisis

1. INTRODUCTION

The world economy of today has evolved into an extensive interdependent network through globalized trade, investment, financial markets, supply chains, commodity flows and people flows. This greatly increases dynamics and introduces elements of instability into economic systems. Since 2007 we have witnessed the worst global economic crisis since the Great Depression of the 1930s. The subprime mortgage crisis that started in the USA in 2007 turned into a heated credit crunch in 2008 and became a global recession in 2009. It reduced global real activity and trade to a degree unprecedented since World War II, contributed to the failure of big businesses and significant decline in economic activity and necessitated substantial financial commitments from governments (Cecchetti *et al*., 2009). As of 2010, after more than 3 years of turbulent movements and digestions, the crisis seemed to be under control, and significant risks for the world economy seemed to be fading away. However, a new wave of turbulence in the form of debt crises has brought new threats to the stability of the world economy.[1]

* Correspondence to: Shuhua Liu, Arcada University of Applied Sciences, Department of Business, IT and Media, Jan-Magnus Janssonin aukio 1, 00550 Helsinki, Finland. E-mail: Shuhua.liu@arcada.fi

[1]As of October 2012, Nouriel Roubini is predicting that the global economy is again on course for a 'perfect storm' in 2013.

In the course of the world experiencing one of the most severe and widely spread economic crises in history, huge amounts of writings and discussions concerning various aspects of the crisis have been produced in the past few years in newswire, social media, research institutions and government organizations. We now have easy online access to large amounts of information-rich and opinion-rich accounts and debates of what has happened in the economic world. Is it possible and how can we make use of existing computational and technology tools to explore such abundant data so that one can quickly develop a sufficient overall understanding of the crisis and can learn about key economic issues related to the crisis? What would be the problems and limitations of the technologies? To answer these questions, in this study we draw upon developments in methods and techniques for natural language processing, especially in text summarization, to analyse a large collection of text documents concerning the global economic crisis. Included in our analysis are selected blog posts, commentary columns and articles by highly influential economists, together with speeches, reports and publications of governmental and international organizations.

Natural language technology has experienced fast and extensive developments in the past two decades. Advancements in the field of artificial intelligence, especially in machine learning methods and ontology development efforts, fuelled by the easy access to large amounts of textual information on the web and active research in information retrieval and search engines have contributed to the development in computational linguistics and text analytics. A wide variety of useful methods and tools have flourished, tackling issues in topic detection and tracking, information extraction, text summarization, question answering and, most recently, sentiment analysis and opinion mining from text documents. These developments are shown in the explosion in the number of publications in computational linguistics, natural language processing, information retrieval, text mining and text analytics conferences.

Text summarization is a process of distilling the most important content from text documents. The first text summarization methods were invented by Luhn (1999) and Edmundson (1999). Very active and intensive research efforts from the computational linguistics community are seen since the 1990s. Much progress has been made in exploring a variety of text summarization methods and techniques (Paice, 1990; Salton *et al*., 1994; Mani and Maybury, 1999; Hovy and Lin, 1999; McKeown and Radev, 1999; Erkan and Radev, 2004; Filatova and Hatzivassiloglou, 2004; Radev *et al*., 2004). In most recent years, research on summarization continues in the direction of incorporating more and more progress made in computational linguistics/natural language processing, domain-specific ontology development efforts, advanced machine learning methods and summary evaluation methods (Hennig *et al*., 2008; Li *et al*., 2008; Otterbacher *et al*., 2008; Ouyang *et al*., 2010; Wei *et al*., 2010). In the meantime, in addition to news summarization, application research started to appear, for example, in the summarization of emails, product reviews, medical dialogues, multilingual or multimodal sources of varying types on the Web, such as blogs, and talk-show transcriptions (the DARPA-funded GALE project; Galley, 2006).

Different from most of the research work reported in the literature that deals with controlled environment and laboratory problems, in this study we apply text analysis and summarization methods to a real-life problem. This raises various questions that we will try to address in this paper:

1. *Data*. What sources should be covered and how can the data be collected?
2. *Summary output*. What would be a proper length constraint for the summaries and how can the results be presented to the user in an easier to understand and useful way?
3. *The system*. How does a state-of-the-art generic system perform on a real-life topic? How can we adapt it to process large amounts of data or to be deployed in interactive scenarios? What is the impact of retrieval noise/off-topic data on summarization?

DOI: 10.1002/isaf

4. *Evaluation*. How can the generated summaries be evaluated and have we really been able to get anything valuable out of the summarization process?

In the following, we first describe the data collection in Section 2 and then introduce the text summarization methods and tools used in our experiments in Section 3. In Section 4 we describe the text summarization tasks and present the results and discussions around the issues concerned. Section 5 concludes the paper.

2. THE GLOBAL ECONOMIC CRISIS: BACKGROUND AND TEXT COLLECTION

Economic crises refer to great turbulence and decline in economic development activities and economic performance. An economic crisis can be a recession or a depression, with a recession referring to a general downturn in the business cycle and a depression referring to a long, sustained downturn in one or more economies, and thus an extreme form of recession. Economic crises, especially depressions, are characterized by symptoms of large numbers of defaults, sharp decline in asset and equity prices, abnormal increase in unemployment, restriction of credit, shrinking output and investment, reduced amounts of trade and commerce, numerous bankruptcies and restructuring, volatile currency value fluctuations, increased government debts and so on (Reinhart and Rogoff, 2008a, 2008b; Gupta and Mulas-Granados, 2009).

Since its breakout in 2007, and especially after the heating up in 2008, there has been a great deal of ongoing media coverage, reporting, analysis and debate concerning the global economic crisis. In this study we choose to focus on learning about the crisis from well-known and reliable sources. Data are collected from sources that are easily accessible on the Web, including blog posts and newswire commentary columns of highly influential economists, the Beige Book and speeches of US Federal Reserve officials, together with a collection of articles written by leading economists at VoxEU since the financial crisis hit. All texts are retrieved from the Web directly. They are listed in Table 1. The **T1** selection of data is by no means systematic and comprehensive. All issues and events in the crisis may not have been covered and addressed in the texts. But at this stage our purpose is to see what we can learn about the crisis quickly from selected sources rather than trying to be exhaustive in coverage. The size of the text collections varies quite a bit, ranging from the smallest one of 820 sentences to largest with 20,830 sentences and from 17,426 words to 449,037 words, which may partly explain the performance differences in summarizing data from different sources.

3. TEXT SUMMARIZATION AND SUMMARY EVALUATION METHODS

3.1. Text Summarization Methods

Most of the text summarization systems today are extraction based, which regard text summarization as a process of extracting (as opposed to abstracting) important content without having to distinguish between factual content and perceptions or beliefs. A summary is created based on extraction of text segments (phrases, sentences or paragraphs) and then the sequential reorganization of the extracted content without any rewriting. Sometimes the extraction process may be smoothed to some degree (for example, by adding source sentences preceding selected sentences containing anaphoric references).

Different extractive methods make use of different text features to represent the text content and employ different scoring and ranking methods to determine the importance of the content. For

DOI: 10.1002/isaf

Table I. Data and sources

Source and author	Description	No. of words	No. of sentences
Paul Krugman (S1)	Blog posts related to the crisis April 2006 to July 2009	65,505	3,229
Paul Krugman (S2)	Op-ed column for *The New York Times* (NYT) related to the crisis April 2006 to July 2009	179,588	8,344
Dean Baker (S3)	Blog posts and commentary column for *The Financial Times* related to the crisis November 2005 to June 2009	449,037	20,830
Carmen Reinhart (S4)	Articles on the crisis March 2008 to January 2009	17,246	821
Barry Eichengreen (S5)	Articles on the crisis May 2007 to June 2009	21,822	1,102
John Taylor (S6)	Articles on the crisis September 2006 to November 2008	75,542	3,536
Richard Baldwin (S7)	Articles on the crisis July 2007 to April 2009	29,984	1,393
Federal Reserve Beige Book (S8)	Beige Book: briefing about economic activities in all regions January 2005 to June 2009	111,176	5,466
Federal Reserve officials (S9)	Official speeches January 2007 to June 2009	146,166	5,680
Total		1,096,066	50,401

example, *position-based methods* weight the sentences in the different parts of a document differently (Brandow *et al*., 1995; Ouyang *et al*., 2010). Often, sentences under headings and sentences near the beginning and end of a document or a paragraph are given extra weight compared with those in the middle. Sometimes they are simply selected automatically; for example, in the *lead* method, sentences are added to the summary based on their position in the source articles alone (Brandow *et al*., 1995). Sometimes the location of a sentence or word in a text is used to adjust the normal sentence score (Ouyang *et al*., 2008). This has proven to be an effective method in summarizing news articles as well as professional documents, with clear and consistent ways to emphasize content at unique locations (Liu, 2009).

In addition to simple surface text features, many other more complicated summarization methods and systems have emerged over the years. The *centroid* method (Radev *et al*., 2003, 2004) is a baseline method used in single and multi-document summarization systems. Given a document or a collection of documents to be summarized, all sentences in the cluster are represented as a tf–idf (term frequency–inverse document frequency) weighted vector space model. A pseudo sentence, which is the average of all the sentences in the cluster, is then calculated. This pseudo sentence is regarded as the centroid of the document (cluster), and the significance of each sentence is determined by calculating how similar each sentence is to this 'centroid'. A centroid consists of a set of the most important/frequent words of the whole cluster. It is regarded as the best representation of the entire document.

The query-based method is another frequently used content extraction method. It borrows techniques from information retrieval, applied at text segment level instead of document level. Given a query (e.g. a set of single words, phrases or short passages), the query-based method will calculate the similarity between the query and the sentences in the documents. Sentences with the highest similarity values are selected to compose a summary. The query-sentence similarity can be computed using a simple cosine similarity measure or tf-idf weighted cosine similarity measure. If the query contains words with high idf values, the sentences containing these words will get high scores and are more likely to be chosen for the summary (Radev *et al*., 2003).

DOI: 10.1002/isaf

When both the collection of original documents and the corresponding collection of model summaries (especially extractive summaries) are available, empirical rules for extracting text segments from the documents can be learned using machine learning methods. The problem of summarization then becomes a text classification problem (Kupiec *et al*., 1995; Wong *et al*., 2008). However, as we do not have model summaries available for our text collection, we need to focus on other approaches.

In earlier studies we have investigated the MEAD system and centroid method (Liu and Lindroos, 2006; Liu, 2009) in the automatic summarization of news and IMF country reports. In this study we shall explore another state-of-the-art system, the *ICSI-Integer Linear Programming Summarizer* (ICSI-ILP) from UC Berkeley (Gillick *et al*., 2008, 2009). The ICSI summarizer was designed with the Text Analysis Conference (TAC) summarization evaluation in mind and optimized to get good performance in those evaluations. In the 2008 and 2009 TAC campaigns, the system obtained the best results or results not significantly different from the best one on a range of manual and automatic metrics (Gillick *et al*., 2008, 2009).

3.2. Summary Evaluation Methods

Generally speaking, summaries are often evaluated along two dimensions: content coverage and readability. While one often has to make a trade-off between the two (adding more contexts to a sentence leaves less room for content), some correlation has been shown between readability and judgment of content: if the references in a sentence are not clear (i.e. low readability), readers consider the content to be harmed also. Content valuation methods can be divided into extrinsic and intrinsic: extrinsic evaluation being task-based evaluation through investigating how a summary affects the completion of some tasks, and intrinsic evaluation being content-based examination by comparing a summary with a target (often called a reference summary or a model summary).

At the core of intrinsic evaluation is the choosing of an appropriate content measurement unit, as well as a significance and similarity measurement for matching the content units. Different evaluation methods address these issues in different ways. Semantic similarity analysis aims to measure content similarity in terms of meaning, while lexical similarity measures only consider the words used without concern for the actual meaning. Lexical similarity is often measured by the word or small *n*-gram overlapping between two summary texts. Examples include simple cosine similarity with a binary count of word overlap, cosine similarity with tf–idf weighted word overlap, measurement of the longest common subsequence, bigram and *n*-gram overlap measurements such as BLEU (BiLingual Evaluation Understudy; Papineni *et al*., 2002) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation; Lin, 2004), which was found to produce evaluation rankings that correlate reasonably well ($\rho = 0.9$) with human rankings (Lin, 2004).

Measuring and comparing content at word- and sentence-level granularity was found not precise enough and unsatisfactory. The BE (basic elements) method (Hovy *et al*., 2006) and the pyramid method (Nenkova and Passonneau, 2004; Nenkova *et al*., 2007) acknowledge the fact that there is no single best model summary and they offer ways to address the content variation across multiple model summaries of the same source text. The BE method, as a further development of the ROUGE method, tries to evaluate a summary by the matching of the basic elements it contains to a set of basic elements extracted from a number of reference summaries. Basic elements are extracted from the syntactic parse trees of the text. Matched basic elements' scores are calculated based on the number of reference summaries they appear in. By adding the scores of each basic element in the summary, an overall score can be assigned to the summary.

DOI: 10.1002/isaf

The pyramid method offers another solution for semantic similarity analysis. It depends on manually identifying and annotating text units of different sizes that are considered to contain only 'important content' of the summary texts. Summaries are compared manually based on these 'semantic units', which are usually approximately clause-length chunks of continuous or discontinuous sequences of words or phrases shared by the reference summaries (Nenkova and Passonneau, 2004). Thus, the content is identified based on 'shared meaning' instead of shared words or word strings (*n*-grams). Pyramid evaluation was found capable of differentiating a system summary from human summaries. However, creating the pyramid and evaluating the peer summaries are very demanding tasks, as they require heavy manual work.

The difficulty of summary evaluation in this study lies in the lack of human-written gold standards to compare with. As manually made model summaries for such large collections of documents are impossible to obtain, we choose to manually examine the summaries to get first-hand knowledge of the extracted sentences. Instead of comparing the results with baselines for an intrinsic evaluation, we focus on assessing the *usefulness of the content* and the *readability* of all the summaries using qualitative judgment. We believe that content and readability evaluation cannot be clearly separated, as readability directly affects the user's feeling towards the content because it limits the eventual understanding of the content.

In the first part of our evaluation study, we shall give our assessment of the summary outputs source by source. Readability is rated according to how easy or hard it is to understand the extracted sentences. Content is rated in terms of how well a summary brings relevant and important information concerning the crisis. For example, can the summary help us learn and understand the major events in the crisis, the causes and consequences of the crisis and their effects on economic activities, the responses and actions, etc.? In the second part of our evaluation study we try to find out what others think about the usefulness and quality of the summaries through a questionnaire survey.

4. THE ICSI-ILP SUMMARIZATION SYSTEM

The ICSI-ILP Summarization System is a multi-document summarization system that treats summarization as an optimization problem with an integer linear programming (ILP) solution. Although the basic assumption about summarization is similar to the maximum marginal relevance method (CMU) and SumBasic (Microsoft/Columbia) – that is, that a good summary will be able to gather as many important facts with as little redundancy as possible – the idea is formalized rather differently as a global optimization problem. The basic idea of the approach, first proposed by Filatova and Hatzivassiloglou (2004), is to find a set of sentences that covers the most relevant information units from the input documents. Such information units are named as *concepts* generally, but they may refer to events, facts or entities and their relations. In this framework, the input is a large set of sentences that refer to concepts, a concept being referred to by multiple sentences. Each sentence has a length, and concepts are associated with relevance weights. The objective is to find the subset of sentences that fits a length constraint (derived from the sum of the selected sentence lengths) and that maximizes the total weight of the concepts referred to by those sentences (the sum of the relevance weights). Unlike most summarization approaches that measure relevance at the sentence level, the concept-based model measures relevance of pieces of information inside sentences (i.e. at the concept level). Provided that concepts are not redundant with each other, redundancy at the sentence level, and more importantly at the summary level, can be completely estimated by the number of times a concept is repeated. Instead of handling redundancy directly, the model counts each concept once, giving no credit for extra

DOI: 10.1002/isaf

occurrence, and can thus naturally avoid redundancy in groups of more than two sentences (something that sentence-level approaches cannot do).

A global optimum to the concept-based sentence selection problem can be found using ILP (Gillick *et al*., 2008). Takamura and Okumura (2009) also proposed a similar approach under the name maximum coverage with knapsack constraint (MCKP). The problem is stated as follows:Maximize:

$$\sum_i w_i c_i \tag{1}$$

Subject to:

$$\sum_j s_j l_j \leq L \tag{2}$$

$$s_j o_{ij} \leq c_i \quad \forall i, j \tag{3}$$

$$\sum_j s_j o_{ij} \geq c_i \quad \forall i \tag{4}$$

$$s_j \in \{0, 1\} \quad \forall j \tag{5}$$

$$c_i \in \{0, 1\} \quad \forall i \tag{6}$$

Here, c_i and s_j are binary indicators for the selection of concepts and sentences respectively ($c_i = 1$ if concept i is in the summary, $s_j = 1$ if sentence j is in the summary). l_j is the length of sentence j; L is the maximum length; w_i is the weight of concept i; o_{ij} is an indicator of the occurrence of concept i in sentence j. Equation 1 is the score of a summary, the sum of selected concepts weights. Equation 2 is the length constraint, meaning that a summary shall not exceed a given length. Equations 3 and 4 ensure the consistency of a selection: if a concept is not selected, no sentence where it occurs shall be selected (equation 3); if a concept is selected, at least one sentence where it occurs must be selected (equation 4). Equations 5 and 6 ensure that s_j and c_i are binary indicators. In theory, equation 3 is not needed if concept weights are all positive; but in practice, if a concept selection leads to a summary shorter than the length constraint, it would be padded with random sentences that do not refer to any new concept.

In the ICSI-ILP summarizer, concepts are represented by word n-grams valued by their frequency in the input document, but more sophisticated concepts could be beneficial. Basic elements (pairs of words and the dependency relation that binds them; Hovy *et al*., 2005), summarization content units (Nenkova *et al*., 2004) or information nuggets (Lin *et al*., 2005) are good concept contenders, but one has to consider two factors when choosing the model of concepts: accurate concept extraction and effective concept weighing. While basic elements are appealing as concepts, their extraction relies

DOI: 10.1002/isaf

on dependency parsing, which can be erroneous even though state-of-the-art systems reach *F*-scores in the 80–90% range. Another problem lies in the difficulty in discriminating between relevant and nonrelevant basic elements. With word *n*-grams, better concept extraction accuracy can be obtained and concept weighting as well as relevance assessment has many mature methods to rely on. In the ICSI summarizer stop-word-only *n*-grams and *n*-grams that appear in a small number of documents are dropped. We used the 571 stop-word list available in NLTK (http://www.nltk.org).

ILP is expensive and does not scale well with the size of the input. It is desirable in the context of large input and/or user interaction to obtain quick solutions at the cost of accuracy. Therefore, we propose an alternative algorithm that provides a greedy solution to the sentence selection problem. Algorithm 1 presents the greedy selection variant, with an additional length normalizer to fit shorter sentences in the summary that generally yields higher objective function value.

$$
\begin{aligned}
&c_i = 0 \quad \forall i, \quad s_j = 0 \quad \forall j \\
&\textbf{while} \quad \sum_j s_j l_j \leq L \quad \textbf{do} \\
&\quad \hat{j} = \text{argmax}_j \frac{1}{l_j} \sum (1 - c_i) o_{ij} w_i \quad : \quad s_j = 0, \quad l_j + \sum_k s_k l_k \leq L \\
&\quad s_{\hat{j}} \leftarrow 1 \\
&\quad c_i \leftarrow 1 \quad \forall i \quad : \quad o_{i\hat{j}} = 1 \\
&\textbf{end}
\end{aligned}
\tag{7}
$$

Algorithm: Greedy algorithm for sentence selection under the concept model. Symbols are the same as in equations 1–6. While summary length is less than the length constraint, it selects, from the set of unselected sentences ($s_j = 0$) that would fit within the length constraint (equation 7), the sentence that adds the highest length-normalized weighted concepts to the selection. Length normalization is added to favour shorter sentences for the same concepts.

The greedy algorithm performs m times an argmax over n sentences, where m is the number of sentences that eventually enter the summary. Therefore, its complexity is in $O(n^2)$. On the other hand, the maximum coverage with a knapsack constraint problem is NP-complete (Gillick *et al.*, 2008), that is, an exponential number of solutions have to be evaluated to find the best solution. Solving an ILP formulation is also NP-complete, but solvers make use of branch-and-bound techniques for quickly discarding poor solution subspaces. When the number of input sentences is small enough (<2000), solutions can be found within minutes by off-the-shelf ILP solvers, while the greedy algorithm can achieve sub-second solutions. In the remainder of this work, we use ILP inference for offline-generated summaries and greedy inference when interaction is needed (e.g. querying a search engine). We use the CPLEX solver from ILOG (http://www.ilog.com/products/cplex) for solving ILP problems.

The ILP- and greedy-inference-based systems were evaluated following the standard evaluation setup of TAC. For example, the TAC 2009 task consisted of generating 100-word summaries of about 10 documents on various news topics, such as disasters, trials or elections (50 topics/document sets). Systems were ranked according to a range of metrics, including ROUGE, human rating of summaries for responsiveness and linguistic quality, hand-aligned events (pyramid), and the recall of word *n*-grams with four reference summaries written by experts. When comparing the ILP and greedy inference systems using ROUGE-2 (involving word bigrams), and only considering the TAC 2009 Set A task (which matches our scenario), the score for greedy inference (0.1174) is lower than the score for ILP inference (0.1274), but still better than the second best system (0.1084, sysid 35) in that evaluation

DOI: 10.1002/isaf

(Dang and Owczarzak, 2009), showing that according to ROUGE the degradation due to using a greedy algorithm should not affect the results by a large amount.

5. LEARNING ABOUT THE GLOBAL ECONOMIC CRISIS THROUGH TEXT SUMMARIZATION

5.1. The First-Round Summarization Experiments

In our experiments the system uses word bigrams to generate concepts. Word bigrams were chosen because they represent the right compromise between accuracy of extraction (higher level semantics would be too error prone) and specificity (simple words are not specific enough). In particular, word bigrams have resulted in the highest human-judged scores in the TAC'09 evaluations (Gillick *et al*., 2009).

In preprocessing, words are stemmed with the Porter stemmer (http://snowball.tartarus.org/algorithms/porter/stemmer.html); two consecutive stems then make for a bigram. We keep stop words in bigrams but remove stop-word-only bigrams. Concepts are weighted by their frequency in the source. Sentences that begin with pronouns are forbidden from entering the summary. Summaries are generated for each of the nine data sources separately, applying five different variations of the ICSI summarizer:

1. *Generic summaries*. Concepts are generated using word bigrams as described earlier. From this set of concepts we only keep the 200 most frequent bigrams that occur at least three times in the source. Then concepts are associated to sentences where they occur and we keep the 2,000 sentences that overlap most with concepts. Sentences of less than 10 words are also discarded. Using frequency as concept weight, we generate by the ILP approach the selection of sentences that maximizes the sum of selected concept weights (equation 1), while not exceeding a length of 400 words. The resulting sentences are ordered by source and by date.
2. *Question summaries*. Same as when generating the generic summaries, but only allows questions in the summary; all sentences that do not end with a question mark are discarded. For each question selected, we add up two to three non-question sentences that follow the question in the source, to hopefully cover the answer to the question. The size of these summaries is much larger because the length limit only applies to questions, not the answers.
3. *Query1 summaries*. Similar to generic summaries, but uses predefined queries to filter sentences prior to the selection. Sentences that share less than two words with the query are dropped and one summary is generated for each query. The Query1 variant extracts and weighs concepts using all source sentences, but performs sentence selection from the query-filtered sentences. This way, out-of-topic sentences are less likely to enter the summary even though generic concepts still lead the sentence selection.
4. *Query2 summaries*. Similar to Query1 summaries, but instead of filtering sentences before the selection, we filter sentences before concept extraction. Therefore, concepts extracted are more focused on the query, but the whole pool of sentences is available for extraction.
5. *Query3 summaries*. Combination of Query1 and Query2. Sentences are filtered by the query both for concept extraction and for sentence selection, putting more focus on the queries.

We used queries/concept-filters of seven categories: crisis, economic activity and related events, economy measures, financial institutions, impact of the crisis, solutions and government responses, and opinions. Each query is simply defined as a set of concept/topic/event terms of interest. These

DOI: 10.1002/isaf

topics/concepts/events/opinions were formulated to reflect the basic concerns in learning and understanding the crisis phenomena.

The experiments produced 23 summary outputs for each data source, so all in all resulted in 207 summary outputs from all the sources. Each output is around 400 words, presented nicely within one page. We designed a simple user interface to display the summaries and allow intuitive access to the source data with a single click. Summaries consist of the extracted sentences labeled with date and source information, ordered by source, then date, then post/article, then sequence within the post/article.

5.2. Results Assessment and Discussions

In this section we give our subjective assessment of the automatic summaries. It is mainly done by one of the authors reading through the summaries source by source and looking at the relevance, readability and potential usefulness of the extracted sentences. We also get a sense of the effect of different methods and sources on the outputs. The assessment, as such, is bound to be very biased, but it gives us important first-hand knowledge about the summaries and will be complemented by a survey evaluation conducted afterwards. Table 2 presents an overview of the summary outputs. From the examination of the different kinds of summaries we obtained some interesting observations: **T2**

- *Summary length*. Four hundred words fill about one page, which seems to be a very good length for a summary. It does not demand too much time from the reader, but is still able to deliver a meaningful amount of information, not just one or two pieces of information.
- *Summarization approaches*. The generic, Query1, Query2 and Query3 summary results share similar readability and complement each other in terms of content, but there is sometimes also much content overlapping. For example, in the case of Paul Krugman Blog (S1), Query1 and Query3 outputs have a considerable amount of overlapping sentences (differing a bit between the queries, but in most cases more than 50% overlapping). Query2 outputs, on the other hand, are almost always different from Query1 and Query3 outputs, with few exceptions. Overall, the results from Query1 and Query3 seem to deliver somewhat richer and more relevant information concerning the crisis than in the Query2 summaries.

The different concept filters did not seem to have a big impact on the extracted content. They seem to be unable to pick up the content needed; they do not differ in the way a user would have wanted. Not only may the extracted sentences be unrelated to the topics defined by the concept filters, but the outputs from applying different concept filters are also not necessarily complementary to each other in the topics and issues covered. There are similar sentences in the case of all the different concept filters, as well as in the generic summaries. This shows that concept filters need to be very carefully and skillfully designed to contribute to the summarization process effectively.

> 'Question summaries' are motivated from the observation that questions, especially those that contain key concepts, would reveal important issues of concern when trying to understand or discuss a topic. In addition, when questions are extracted together with the two or three following sentences, they form a natural local context for each other and, thus, would give much better readability than summaries that only list narrative sentences out of context. Questions and answers are also carriers of opinion-rich content. The question summaries are twice or even three times as long as other summaries owing to the way they are formulated. Readability is found to be much improved and much easier to understand compared with the generic summaries. The unique contribution of such an approach is also the capability to capture a lot of opinion content. It provides much richer information than the generic summary, and as a result is better able to sense and reflect the author's thinking and writing style.

DOI: 10.1002/isaf

Table II. Summary outputs – first-round experiments

Source	Comment	Topics covered
Paul Krugman blog (S1)	Among the hardest to understand and prone to being misunderstood out of context when compared with the results from other sources. When reading the extracted sentences, one will be aware of the need to read more to find out what is really being talked about and what is the authors' true meaning and intention. Reading only the extracted content, one also clearly misses out on much of the author's sarcastic and humorous writing style, although one can still sense the succinct and critical style.	Regulations, money injection into the financial system, the dollar issue, mortgage crisis, policy failure, inflation, liabilities, recession, Asian financial crisis, social security, health care. Entities mentioned: the Fed, dollar, Alan Greenspan, Jim Hamilton, Jeff Frankel, Kucinich, Countrywide Financial Corporation, Obama, the government, Fannie and Freddie, Daniel Gross, The Great Depression.
Paul Krugman NYT column (S2)	Readability is much better than the summaries of the blog posts. However, the content is not all relevant because it covers a much longer period of time, not just the recent economic crisis. The relevant parts cover only very limited information about the crisis. One-third of the sentences are from the period before the current crisis.	Health care, social security, tax cuts, country's solvency, US economic policy, the world financial system, financial rescues in Europe and in USA, output and employment, causes of the crisis, public spending, inflation, US government debt rating, and so on.
Dean Parker Blog (S3)	The summaries pick up a lot of comments and critics from the author about the media (especially *The Washington Post* and NYT). The extracted sentences are in general easier to read compared with S1.	The relevant sentences about the crisis bring forward information concerning responsibility for the housing bubble, spending, dollar, trade deficit, run-up in house prices, wage, food price, gas price and so on.
Carmen Reinhart articles (S4)	Different from blogs, where wider economic, political and social topics are covered, this is a collection of articles explaining and debating the economic crisis using standard economic terminology, which benefits automated processing greatly. The extracted content is relatively easy to read and educating. Some very good questions are captured.	Debts, subprime lending, financial system, regulatory authority, real interest rate, economic growth, commodity prices, sovereign defaults, capital flow bonanzas, global market distress, global financial crisis, current account deficits, government securities, asset prices, academic crisis, recapitalize the banks, output and employment, domestic banking crises, inflation, governments default on external debt, housing price.
Barry Eichengreen articles (S5)	Big portions of sentences were about currency and exchange rate, and thus not directly relevant to the crisis (except the question summaries). Compared with other sources, there is more content about economic issues from a global perspective.	Currency issues, exchange rate, euro zone, EU, Asia, policy instruments, ECB, IMF, Fed, US Treasury, G7/8, G20, Great Depression
John Taylor articles (S6)	There were some problems in preprocessing which result in a large number of incomplete sentences. Readability is very low; thus, we discarded the results. This suggests that tools for users to quickly fix preprocessing errors will be necessary to help improve summarization results. When system outputs give hints that there are problems with, for example, sentence segmentation, intervention in the preprocessing is critical.	
Richard Baldwin articles (S7)	The extracted content is very relevant; this may mainly be attributed to the source text, which is relatively small in size and on focused topics.	An interesting set of questions on protectionism, G20, London Summit, global growth, world trade and investment, global crisis debate, US large current account deficit, WTO, IMF,

(*Continues*)

DOI: 10.1002/isaf

Table II. (Continued)

Source	Comment	Topics covered
	Good questions concerning global actions are captured. However, not as much insightful opinion as delivered in the summaries of Paul Krugman's blog and column.	ECB, dollar, euro area, government crisis-fighting measures, trade credit problem, fiscal stimulus.
Federal Reserve Beige Book (S8)	Various aspects of economic activities in different Federal Reserve districts. One interesting point is that no question was extracted from this source.	Signs of increased economic activity, sluggish conditions, residential and commercial real estate markets and construction activity, office rents; demand for mortgages, labour markets, wages, employment, manufacturing activity, pace of growth, consumer spending, retail sales, farming condition, auto sales, bank loans, tourism activity, housing activity, deposit growth.
Speech of Federal Reserve officials (S9)	This seems to be a very good official source of information to learn about the crisis. The speeches are professionally written and carefully phrased texts and have a certain structure to follow. They are official documents, thus do not show much "personality" in the text. Very straightforward and easy to read and understand.	Economic policy, monetary policy, Federal Reserve, inflation, banks and financial system, financial/economic stress/stability, policy tools, short-term interest rates, provision of liquidity, market unrest, structured-credit products, policy actions, central banks, housing sector, labour markets, credit and mortgage market.

DOI: 10.1002/isaf

- *Source text*. When examining the experiments and outputs we should note that it is very hard to overemphasize the importance of exploring and preparing the texts to be summarized. Much of the text summarization research is usually done with well-prepared source texts. However, in practical applications this is almost never the case. The users will need tools that can help to quickly and easily add, delete, configure, preprocess and explore the source text as necessary. This is as important as a careful design of the summarization method in practical applications. In many cases, in order to be able to create good summaries to serve a specific purpose, users need to be selective with the source text and understand the source text better in terms of what summaries it can deliver.

So, how useful is the extracted content? Can the summaries replace the original text? Hardly. Although there are relevant pieces of information in the extracted sentences, the summaries, as such, are informative, but reading a summary of multiple documents, especially blog posts, gives you a sense of incompleteness. It is thus not as enjoyable as reading a single blog post since it cannot provide a sense of thorough understanding of one particular topic. However, this is perhaps a limitation for any summary which only is manifested more strongly in texts written with a distinct writing personality and thinking style.

5.3. Further Experiments: Creating Event-based Summaries

In this section we look at how useful these summaries are with regard to helping acknowledge and understand some major events during the crisis.

The US Council on Foreign Relations maintained a Timeline of the Global Economic Crisis (http://www.cfr.org/publication/18709/) where a comprehensive list of important events and factors of the crisis are presented and described. In 2008, *Time* magazine also named the Top 10 Financial Collapses of the year (http://www.time.com/time/specials/packages/completelist/0,29569,1855948,00.html#ixzz0jqzkej6D), which are significant events during the economic crisis (Losing Lehman, AIG's Credit Default Swaps, The Detroit Three, Citigroup, Freddie Mac and Fannie Mae, Supposedly 'Safe' Securities, Rating Agencies' Credibility, Exploding Hedge Funds, Greenspan's Reputation, Iceland Goes Belly Up).

So, how have the various summaries helped to reveal and understand all these crisis events? To see how the extracted content has covered the major events during the crisis, we looked at the number of different sentences that contain the main entity of the events in the summaries for each source. We chose to examine only summaries from the following six sources: the Beige Book, speeches of Federal Reserve Officials, Carmen Reinhart articles, Dean Parker blog, Paul Krugman blog and Paul Krugman's column in NYT. This helps to greatly reduce the amount of work needed to go through summaries of all the sources. Also, judging from the summary outputs, other data sources may have less relevant content on these events than these six sources.

It is quickly noticed that even in these six sources only little content concerning the list of events can be found in the various summaries. One reason could be that the source texts do not contain much relevant content in the first place – the best sources of such information would probably be news (and blog) collections of rich origins. Or even if they do, the focus of the texts are often very much on larger and deeper issues (government economic policy, markets, financial systems, regulations and so on) related to or behind these crisis events, so they are not necessarily the best place to look for event- related information. Another reason would be that the generic summaries as well as all the query-based summaries were not particularly directed or adjusted for extracting information related to these events.

Among the different types of text data we have here, the blog and commentary column (S1, S2, S3) appears to be the only one that has contributed limited information about some of the events. In the blog

DOI: 10.1002/isaf

and NYT column of Paul Krugman, we found comments concerning Alan Greenspan, and Freddie Mac and Fannie Mae, but to a very limited extent; the blog of Dean Parker contains one comment about Citigroup. However, we should not make the conclusion that the summarizer generally does not work well in generating summaries of crisis events.

Considering that neither the generic summaries nor the query-based summaries were specifically targeted for content related to the major crisis events listed, we set out on another set of experiments. In this second round of experiments, we first apply specific event-queries to the original source texts. The retrieved documents and blog posts are then fed into the generic summarizer for further processing.[2] We used the Lemur search engine (http://www.lemurproject.org/) to index the collection of documents and allow the submission of text queries.[3] For each query, the summarization system generates a 400-word summary from the 20 most relevant documents retrieved by the search engine. Using this setup, queries can be quickly crafted and modified in order to fit a user's information need, leading to deeper exploration of the events and facts described in the source texts.

Some initial tests quickly indicate that the system outputs are very sensitive to the way the queries are processed. When using strict phrase matching, many empty summaries are the result. It works much better when applying more relaxed word matching. All the extracted sentences are not always relevant to the specific crisis event; however, at least some of the sentences are always related to the event concerned, and almost all the sentences are relevant to the economic crisis in general. In this sense, this approach produced better summary extracts than in the first round of experiments.

For the same event-queries, we tested retrieving a different number of documents; for example, 20 and 100. This has a definite effect on the set of extracted sentences; however, there is always overlapping content. Even if one can give subjective opinions about which summary is preferred, it is hard to tell if that suggests any performance pattern. The experiments also notice that blog posts, newswire columns and official speeches tend to be richer sources of crisis event information than the Beige Book and research articles, and understandably so.

To examine the effect of query terms, we experimented with using the following alternative queries: (i) event title phrases (e.g. "February 2007 US housing downturn"); (ii) named entity in the event title (e.g. Bears Stern, Lehman Brothers); and (iii) expanded query term set, manually created based on event descriptions; for example:

> "September 15 2008 Lehman Brothers collapses Lehman collapse series moves on Wall Street Lehman Brothers investment bank US financial sector Lehman bankruptcy largest bankruptcy in US history Lehman failing Bank of America buys Merrill Lynch AIG downgrade AIG largest US insurer Federal Reserve loans to AIG".

The three different types of queries were all fed into a word-matching-based retrieval process. It turns out that simply using event titles or named entity terms can already bring out much relevant and useful content concerning the crisis. Sometimes the extracted sentences do not contain the event information explicitly, but when linked back to the original source text the relevance of the text become obvious. The expanded query term sets also work, but it is not clear that the results are better than the simpler alternatives.

[2]Comparing our work with the guided summarization task, basically we adopted the approach of query-focused summaries rather than using the list of categories and aspects. However, such a list of categories and aspects as used in TAC 2011 provided a very interesting framework for us to investigate in our future work.

[3]Lemur implements a language-modelling-based information retrieval model described in Strohman *et al*. It is an off-the-shelf search engine that could perform as well as many others. We used Lemur as a search facility to our corpus mainly out of convenience because it is not important in our work to actually try different search engines. What we need is just a search engine that performs as well as a baseline.

DOI: 10.1002/isaf

Depending on the source and the event, sometimes using the first two simpler alternatives may even obtain better results than using the expanded list of related words as queries.

5.4. Extended Experiments and Evaluation Survey

So far, we have tried to understand the system and the summary outputs through the lenses of us as researchers. In order to get a more objective evaluation of the results, we conducted a small survey to find out how other users perceive the usefulness and quality of the summaries. To make the evaluation practically manageable, we again downsized the text sources to include only the four listed in Table 3. The text collection in this dataset is different from the previous one. It is newly collected and has a different time span and should be more complete. T3

In this third round of experiments we adopted a number of different strategies for summarization. First, each data source is summarized separately. Second, each individual collection is split into smaller collections of different years and summarized year by year. For each collection, one 400-word generic summary, six 400-word query-based summaries and one question-based summary are produced. Six query-filters are used:

1. housing, bubble, subprime, mortgage, foreclosures, housing price;
2. downturn, bankruptcy, collapse, Lehman Brothers, Bear Stearns, AIG, Fannie Mae, Freddie Mac;
3. Alan Greenspan, Bernanke;
4. unemployment, jobs;
5. Wall Street, stock, stock market, financial risk, catastrophe, credit ratings, hedge funds, banks, banking system
6. effect, rescue, rescue plan, responses, fiscal stimulus, stimulus, bailout.

The summarizer is similar to the one used before: concepts are generated using word bigrams. From this set of concepts, only the 200 most frequent bigrams that occur at least three times in the source are kept. The remaining concepts are associated with the sentences where they appear, and the 2000 sentences that overlap most with the concepts are kept and form the base for selection; sentences of less than 10 words are discarded. Question-based summaries are also generated in the same way as described earlier.

With such a careful structuring and blending of sources we hope to be able to achieve a balance of summary content both in terms of the topic coverage (represented by queries) and temporal coverage. However, the number of summaries is too large to be evaluated with reasonable resources. In addition, from the end-user's point of view, the final product of the summarization process should not only be many different summaries. Therefore, we chose to feed the intermediate summaries from the first batch

Table III. Dataset for survey evaluation

Source and author	Description	No. of words	No. of sentences
Dean Barker blog	Blog posts during 12 April 2006 to 13 April 2010	560,074	26,268
Paul Krugman blog	Blog posts during 26 September 2005 to 13 April 2010	488,821	26,499
Paul Krugman column	Commentary column for the NYT related to crisis (2000 to 12 April 2010)	716,894	34,786
Speeches of Federal Reserve officials	Official speeches (13 June 1996 to 13 April 2010)	2,834,009	111,519

DOI: 10.1002/isaf

processing into the summarizer again as input documents, hopefully to reduce redundancy and retain only the most relevant and shared content in a merged summary. In the end, three summaries from each source are included as the content to be evaluated in the survey: merged-generic-summaries, merged-queries-summaries and merged-questions-summaries, each about 400 words, all from year 2006 on. So, all in all, it is a $1200 \times 4 = 4800$-word summary that is presented to the evaluators. This is still a heavy load of information to any evaluator, but considering the size of the source collection, the compression rate is already very low compared with a normal compression rate of 10%. The extracted sentences are presented together with the title of the post or column or speech they belong to, source by source, then in chronicle order, arranged in two columns in order to retain the fragmented nature of the extracted sentences.

In the survey, we asked participants to answer 12 questions concerning the summary. All the answers are based on a five-point Likert scale. On a scale of 1–5 (strongly disagree, somewhat disagree, not sure, somewhat agree, strongly disagree), the summary was evaluated with regard to relevance, informativeness, proportion of important observations of interest, usefulness, whether it is easy to understand or confusing, if it meets one's expectations and whether they like it. In addition, on a scale of 1–5 (very poor, not so good, ok, good, very good), the summary was rated in terms of proportion of important information, readability and overall quality. We also had open questions for the evaluators to give free-text comment on the advantages and disadvantages of such a summary. Our evaluation, as such, is not so much about whether the summary really catches the gist of the original documents, but about how useful the summary is in helping to learn about the crisis.

Eight subjects were recruited to participate in the evaluation. They were nonexperts in economics crisis, but they are all researchers in universities. Six of them completed the survey. The results are shown in Figure 1. **F1**

As indicated in Figure 1, the readability and overall quality of the summaries are rated as ok, but not so good, although a good proportion of useful information can be found in the summaries. The extracted content is very relevant and informative about the crisis but not so useful after all. There are important observations of interest contained in the summaries, but not many. The summaries are confusing and hard to understand. Still, the summaries seem to meet the judges' expectations of automatic summaries. Considering very few actually like the summaries, the general expectations on automatic summaries are really not so high.

Overall the opinions of the judges vary to a certain extent, with a Fleiss's kappa of 0.398, indicating only a fair to moderate agreement in the evaluation. Some of the comments they made provide more detailed opinions and insights into the content, readability and quality of the summaries:

S1. "Observations of interest can help in searching for more information on the topic. Information out of context; abbreviations not explained, missing information (e.g. numbers, statistics); content is misaligned with the title; very difficult to comprehend."
S2. "Quick glance of an issue perhaps can help on deciding if I would like to read the complete article. Many of the summaries seem to have either a question only or a result without any hint of how it was reached; perhaps it is not a very good to read summaries which span many years all at once. Q8 I believe this can affect the outcome of this study."
S3. "The quality of the syntax in the text is surprisingly good. I see the usefulness of summarization more in getting single piece of information rather than to have a short version of the all (whole) article."
S4. "Potentially summaries can generate new sources of information. With this particular case they did not seem to do that so well. The two columns per page are distracting when you try to read

DOI: 10.1002/isaf

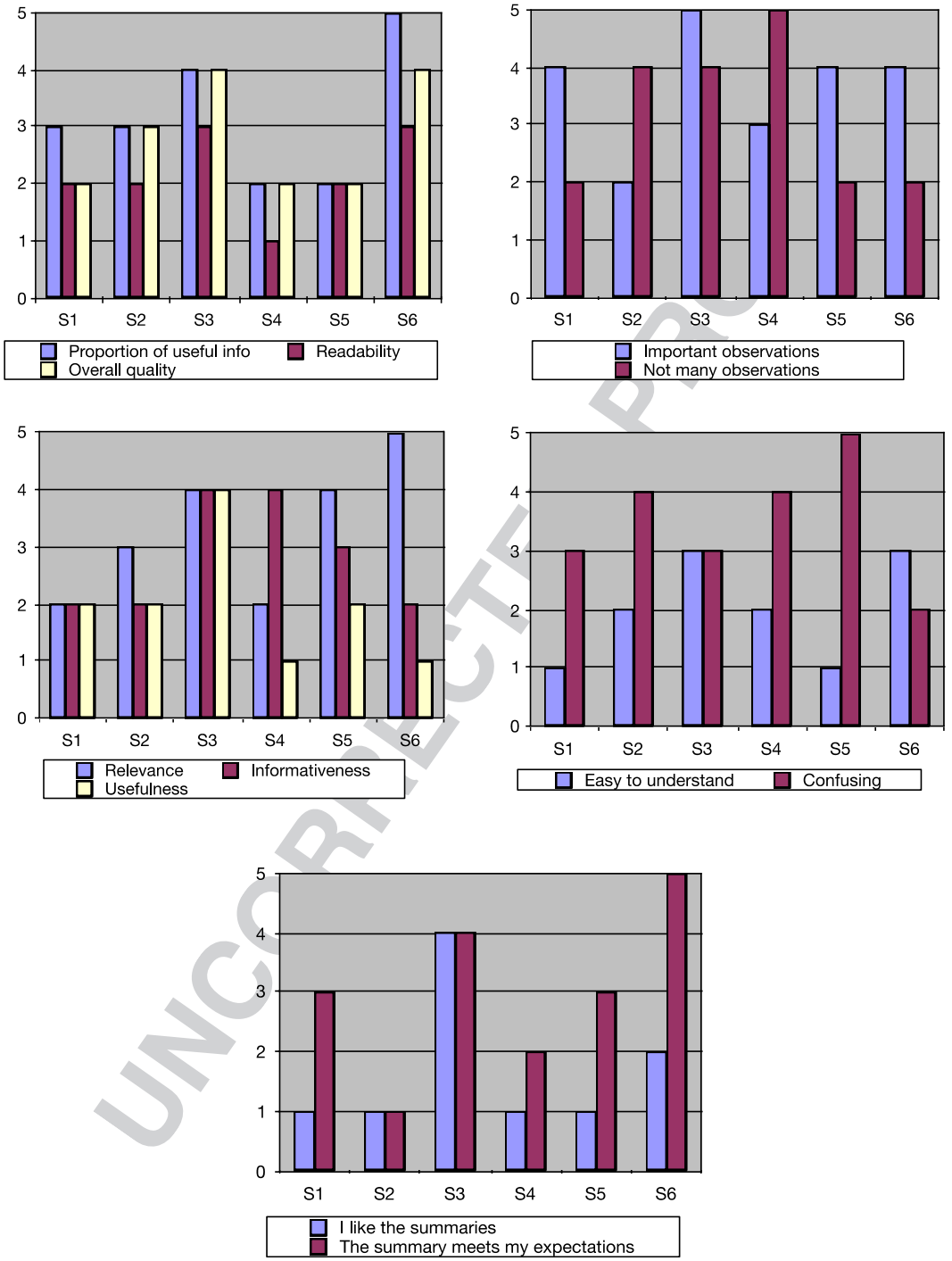


Figure 1. ▪▪

Q7

DOI: 10.1002/isaf

summaries in chronological order. . . . I really dislike the two columns and the fact that they are of different width. . .”

S5. “You get an idea about what issues are discussed and what is problematic. If you regularly follow the discussion the summaries can be practical. Too short, personal opinions hang loose. Q9 You easily hooked to catchy key words without understanding or evaluating the content. Summaries might work for people who are aware of the subject discussed, know the basis and the people summarized. However, if you are not familiar with the area before, they are confusing and annoying.”

S6. “They sound like the most central sentences from the sources; likely useful for people already educated on the subject. If one wants to learn/get introduced to the topic, such disconnected and specific statements without proper context are very overwhelming. A high level conclusion of the different sides/aspects in this heated debate would be a better introduction for a novice in this topic, but it’s obvious that it would be very hard to create automatically.”

The survey results and comments from respondents confirm a number of important characteristics of automatic summaries:

1. The lack of context and sense of heavily fragmented and incomplete information highlights the bottleneck issue of readability. However, it should be noted that in the survey it is the summaries instead of the system that was evaluated by the judges. With the interactive feature of the system and easy link to source text, the readability problem could be much improved.
2. The note on the importance of the amount of information in the summaries and the way to present summaries to readers was a surprising finding. Not much attention has been paid to these aspects in text summarization research.
3. Users’ familiarity with the topic should be an important a factor to consider when evaluating summaries, as well as when generating summaries.

6. DISCUSSION

As an attempt to explore and understand the capability of text summarization techniques in a rather complicated practical application, this study has been demanding and challenging in nature, especially in terms of the summary evaluations. The analysis of the summary outputs is done first in a subjective manner and then complemented by a questionnaire survey. Such an evaluation approach has many limitations, and it is hard to draw firm conclusions. However, our study so far indicates a number of things to be highlighted:

1. Concerning the text sources, in this study we included several different types of texts; some are shorter (blog posts) and some are longer. The results of summarizing all kinds of texts together usually tend to lean on longer texts such as the Beige Book and official speeches. In addition, the failure in processing the John Taylor text reminds us about the importance of other components of an effective system than simply a comprehensive summarization method, especially flexible preprocessing facilities and output evaluation instruments. Mechanisms for users to easily select and reorganize the source texts would be very helpful.
2. Results on different data sources vary very much for the same summarizer. Blogs written by leading economists carry rich personal, professional or political opinion and wisdom. Some parts of the blog

DOI: 10.1002/isaf

texts are often written in a candid way using sarcastic language. The question-based content extraction approach seems more suitable for blog-type texts than for the Beige Book. Question-based content extraction could also be a good entry point for extracting opinion and fact information.

3. Regarding the summary output presentation, appropriate length is important; contextual information (local and global) is very important to include into the summaries or make easily accessible within a click. The summaries are usually only information pieces out of context. An easy link to the source text brings up at least a vague context. Date and source, and title (and first sentence) are very useful meta-information for the extracted content. Not only will they make the extracted sentences more tractable, but they also provide a general semantic context that helps comprehending the fragmented sentences, although our study also found that such mechanisms prove to have only limited effect. More alternate representations should be tackled; for example, timelines, topic terms and opinion maps. In developing applications, an environment to test and analyse results is much more important than one single method.
4. With regard to the system per se, the task of summarizing real events is very different from the tasks studied by international summarization evaluations because it intrinsically binds summarization and information retrieval. This makes the estimation of relevant information less reliable because of retrieval noise and because the sources are not focused on particular themes. An additional challenge is to deal with different genres (blog, news, official articles) in the same collection, with large differences in quantity, quality and focus, conditions that a system crafted for evaluations is not prepared to deal with. Finally, the interactive scenario prevents the use of more advanced models that are too slow to run in reasonable time.
5. In terms of evaluation, we found that evaluating the summaries was made difficult due to the infeasibility of preparing some gold standards given the input sources. We had to factor out the coverage of our sources and compare against external references, such as lists of known important facts, which would only allow for relative benchmarks. Therefore, we performed a preliminary survey to obtain human judgments of the system outputs. The evaluation indicates that there is a long way to go for extractive summaries to satisfy nonexpert users. Although extractive summarization systems have been successful, in that they are generic in nature and are applicable to any domain, extracting sentences out of their context is still a major limiting factor in terms of accurate understanding and leads to serious acceptation issues.

In most evaluation campaigns related to summarization systems, nonexpert users are used instead of human expert judges in order to remove bias in summary evaluation. Evaluation campaigns also provide topic-focused documents containing mostly relevant sentences when they should introduce nonrelevant content to reflect more realistic scenarios. The ICSI-ILP system was very successful in the TAC evaluations, but it is not as successful in our experiments. To fix such a mismatch, it would be helpful for evaluation campaigns to adapt towards richer, more realistic application settings. From a system development point of view, one possible way to deal with the problem is to introduce certain domain-specific facilities, like storylines, to automatically organize the extracted content into better structures to facilitate the evaluation.

1. The summaries are, in general, helpful as a way to help further understand the topic of an economic crisis for people who are already familiar with it to a certain extent, but are not so useful in terms of being a trusted source for a good summary of the crisis, especially for people who are new to the topic. In general, event-based queries are able to extract content that is more relevant to the economic crisis than the generic and query-based summaries in the first batch of experiments.

DOI: 10.1002/isaf

The summaries are more useful as an informative and indicative aid for obtaining a quick grasp of content than as an accurate source of knowledge about the crisis.

7. CONCLUSIONS

In this study we explored the possibilities of applying text summarization tools to learn from text documents the various discussions surrounding the current economic crisis. Different from most research work on text summarization that deals with controlled environment and laboratory problems, we applied a state-of-the-art text summarization system, the ICSI-ILP summarizer, to a relevant real-world problem. A large number of experiments were carried out and the summary outputs were manually examined and also evaluated by several human judges in a small survey.

Our study probably generated more questions than conclusions. However, it revealed new knowledge and highlighted many important things to consider when developing and deploying text summarization technology for complicated real-world applications, although it is not yet possible to make reliable and validated conclusions about the summaries and the system. Our contribution lies more in the insights into the potential and limitations of the summarization methods, which other types of studies have not addressed. Taking these factors into consideration in the summarization process could have a big impact on the evaluation result.

There are naturally many different ways for us to develop a good understanding of a complicated phenomenon such as an economic crisis. Text summarization is only one of the many natural-language technology applications that could be helpful. Other text analysis applications, such as entity recognition, topic and event detection, sentiment analysis and opinion mining (opinion and opinion cluster identification, contradictory opinion detection), would be very relevant technologies to explore individually as well as to incorporate into the summary generation process to help with faster, richer and better understanding of the economic issues concerned. This will be the focus of our future studies.

ACKNOWLEDGEMENTS

Financial support from the Academy of Finland (grant number 111692) is gratefully acknowledged.

REFERENCES

Brandow R, Mitze K, Rau L. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management* **31**(5): 675–685.

Cecchetti SG, Kohler M, Upper C. 2009. Financial crises and economic activity. NBER working paper 15379. Bank of International Settlement.

Dang H, Owczarzak K. 2009. Overview of the TAC 2009 summarization track. In *Proceedings of the Text Analysis Conference, Gaithersburg, Maryland, USA*.

Edmundson HP. 1999. New methods in automatic extracting. In Advances in Automatic Text Summarization, Mani I, Maybury MT (eds). MIT Press: Cambridge, MA; 23–42. (Article originally published in 1968.)

Erkan G, Radev DR. 2004. LexRank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* **22**: 457–479.

Filatova E, Hatzivassiloglou V. 2004. Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization (ACL 2004)*; 104–111. Q10

Galley M. 2006. Automatic summarization of conversational multi-party speech. In *Proceedings of the AAAI SIGART Doctoral Consortium*.

DOI: 10.1002/isaf

Gillick D, Favre B, Hakkani-Tür D. 2008. The ICSI summarization system at TAC 2008. In *Text Analysis Conference*, Gaithersburg, Maryland, USA.

Gillick D, Favre B, Hakkani-Tür D, Bohnet B, Liu Y, Xie S. 2009. The ICSI/UTD summarization system at TAC 2009. In *Text Analysis Conference*, Gaithersburg, Maryland, USA.

Gupta S, Mulas-Granados C. 2009. The effectiveness of fiscal policy during banking crises: an empirical assessment of 118 cases. IMF working paper.

Hennig L, Umbrath W, Wetzker R. 2008. An ontology-based approach to text summarization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08*; vol. 3, 291–294.

Hovy E, Lin C. 1999. Automated text summarization in SUMMARIST. In Advances in Automatic Text Summarization, Mani I, Maybury MT (eds). MIT Press: Cambridge, MA; 81–94.

Hovy E, Lin CY, Zhou L, Fukumoto J. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*; 604–611.

Li WJ, Wei FR, Ouyang Y, Lu Q, He YX. 2008. Exploiting the role of named entities in query-oriented document summarization. In *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence (PRICAI 08)*, Hanoi, Vietnam, 15–19 December; 740–749.

Lin C-Y. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop 2004*, Moens M-F, Szpakowicz S (eds); 74–81.

Liu S. 2009. Experience with and reflections on text summarization tools. *International Journal of Computational Intelligence Systems* **2**(3): 202–218.

Liu S, Lindroos J. 2006. Towards fast digestion of IMF staff reports with automated text summarization systems. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)*, Nishida T, Shi Z, Visser U, Wu X, Liu J, Wah B, Cheung W, Cheung Y-M (eds). IEEE Computer Society: Los Alamitos, CA; 978–982.

Luhn HP. 1999. The automatic creation of literature abstracts. In Advances in Automatic Text Summarization, Mani I, Maybury MT (eds). MIT Press: Cambridge, MA; 15–22. (Article originally published in 1958.)

Mani I, Maybury MT (eds). 1999. A*dvances in Automatic Text Summarization*. MIT Press: Cambridge, MA.

McKeown K, Radev DR. 1999. Generating summaries of multiple news articles. In Advances in Automatic Text Summarization, Mani I, Maybury MT (eds). MIT Press: Cambridge, MA; 381–390.

Nenkova A, McKeown K. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval* **5**(2–3): 103–233. Q11

Nenkova A, Passonneau R. 2004. Evaluating content selection in summarization: the pyramid method. In *HLT-NAACL 2004: Main Proceedings*, Dumais S, Marcu D, Roukos S (eds). Association for Computational Linguistics; 145–152.

Nenkova A, Passonneau R, McKeown K. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing* **4**(2): article no. 4. Q12

Otterbacher J, Radev DR, Kareem O. 2008. Hierarchical summarization for delivering information to mobile devices. *Information Processing and Management: An International Journal* **44**(2): 931–947.

Ouyang Y, Li WJ, Zhang RX, Lu Q. 2010. A Study on position information in document summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Beijing, China, 23–27 August; 919–927.

Paice CD. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management: An International Journal* **26** (1): 171–186.

Papineni K, Roukos S, Ward T, Zhu W. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA; 311–318.

Radev D, Allison T, Blair-Goldensohn S, Blitzer J, Celebi A, Drabek E, Lam W, Liu D, Qi H, Saggion H, Teufel S, Topper M, Winkel A. 2003. The MEAD Multi Document Summarizer, MEAD Documentation v3.08.

Radev D, Jing H, Stys M, Tam D. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management* **40**: 919–938.

Reinhart C, Rogoff K. 2008a. This time is different: a panoramic view of eight centuries of financial crises. NBER working paper 13882.

Reinhart C, Rogoff K. 2008b. The aftermath of financial crises. Paper presented at the Meetings of the American Economic Association in San Francisco, December.

Salton G, Allen J, Buckley C, Singhal A. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science* **264**(5164): 1421–1426.

DOI: 10.1002/isaf

Takamura H, Okumura M. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*; 781–789.

Wei FR, Li WJ, Lu Q, He YX. 2010. A document-sensitive graph model for multi-document summarization. *Knowledge and Information Systems* **22**(2): 245–259.

Wong KF, Wu ML, Li WJ. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, Manchester, UK, 18–22 August; 985–992.

DOI: 10.1002/isaf

Author Query Form

Journal: Intelligent Systems in Accounting, Finance and Management

Article: isaf_1340

Dear Author,

During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions.

- If you intend to annotate your proof electronically, please refer to the E-annotation guidelines.
- If you intend to annotate your proof by means of hard-copy mark-up, please refer to the proof mark-up symbols guidelines. If manually writing corrections on your proof and returning it by fax, do not write too close to the edge of the paper. Please remember that illegible mark-ups may delay publication.

Whether you opt for hard-copy or electronic annotation of your proofs, we recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

Query No.	Query	Remark
Q1	AUTHOR: Ref not listed. Ouyang et al 2010 or Li et al 2008?	
Q2	AUTHOR: "Kupiec et al., 1995" is cited in text but not given in the reference list. Please provide details in the list or delete the citation from the text.	
Q3	AUTHOR: The citation "Nenkova and Passoneau, 2004" (original) has been changed to "Nenkova and Passonneau, 2004". Please check if appropriate.	
Q4	AUTHOR: "Hovy et al., 2005" is cited in text but not given in the reference list. Please provide details in the list or delete the citation from the text.	
Q5	AUTHOR: "Nenkova et al., 2004" is cited in text but not given in the reference list. Please provide details in the list or delete the citation from the text.	
Q6	AUTHOR: "Lin et al., 2005" is cited in text but not given in the reference list. Please provide details in the list or delete the citation from the text.	
Q7	AUTHOR: Please supply a suitable caption for fig. 1.	

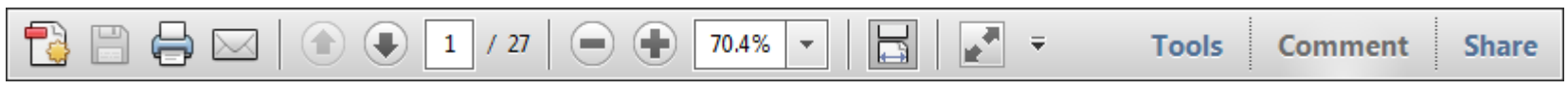
Query No.	Query	Remark
Q8	AUTHOR: "not a very good to read " doesn’t make sense. Is quote correct? Delete "a"?	
Q9	AUTHOR: Change to "loose" from "lose" OK? Also, You easily hooked to catchy" doesn’t make sense	
Q10	AUTHOR: Please check pubyear if captured correctly.	
Q11	AUTHOR: Reference “Nenkova & McKeown (2011)” is not cited in the text. Please indicate where it should be cited; or delete from the reference list.	
Q12	AUTHOR: Please provide page range for reference Nenkova et al. 2007.	

WILEY-BLACKWELL

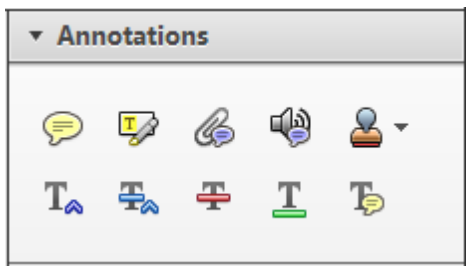
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 7.0 or above). (Note that this document uses screenshots from Adobe Reader X)
The latest version of Acrobat Reader can be downloaded for free at: http://get.adobe.com/uk/reader/

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:

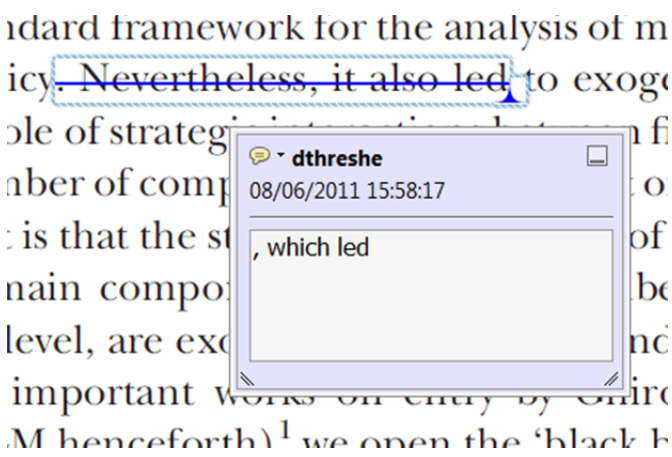


1. Replace (Ins) Tool – for replacing text.

Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the Replace (Ins) icon in the Annotations section.
- Type the replacement text into the blue box that appears.

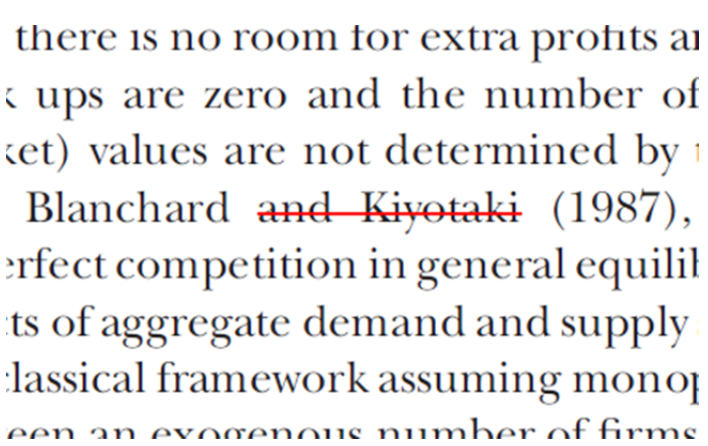


2. Strikethrough (Del) Tool – for deleting text.

Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.

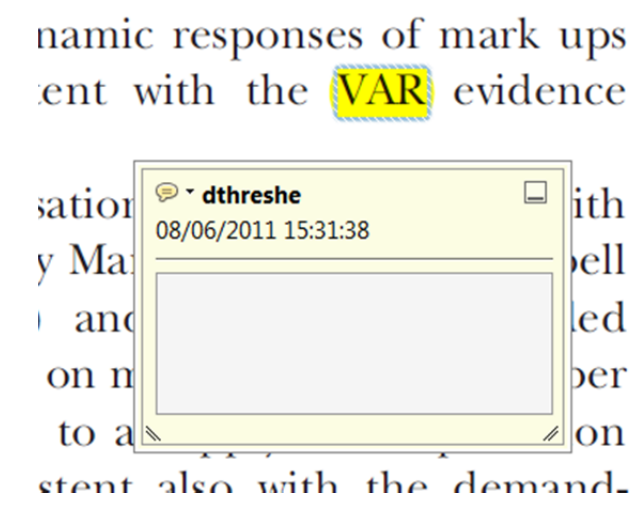


3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.

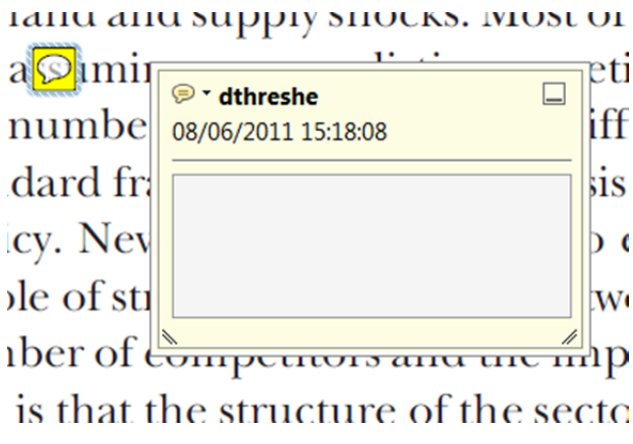


4. Add sticky note Tool – for making notes at specific points in the text.

Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.



WILEY-BLACKWELL

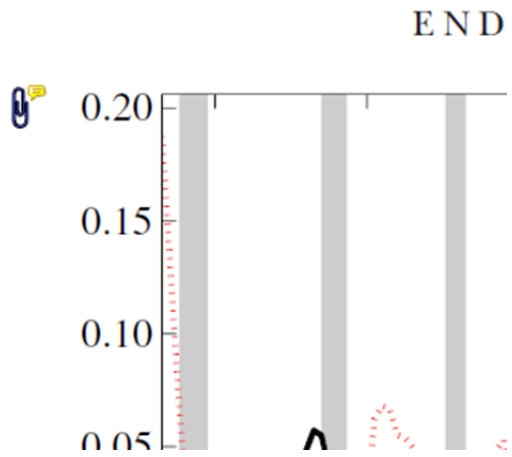
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

5. Attach File Tool – for inserting large amounts of text or replacement figures.

Inserts an icon linking to the attached file in the appropriate pace in the text.

How to use it

- Click on the Attach File icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.

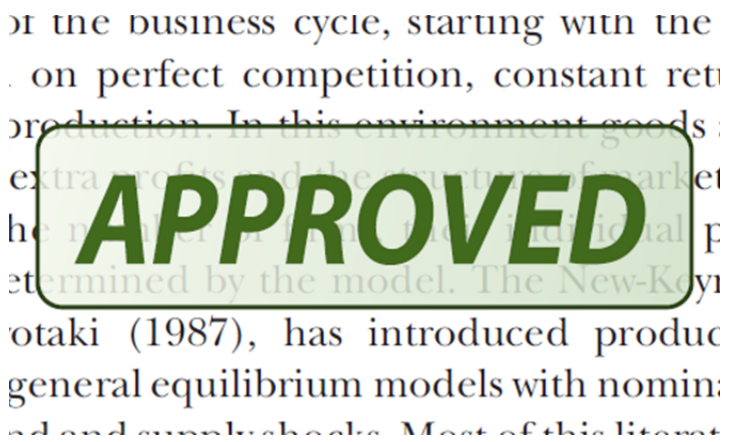


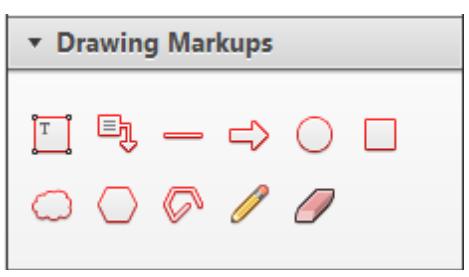
6. Add stamp Tool – for approving a proof if no corrections are required.

Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the Add stamp icon in the Annotations section.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).



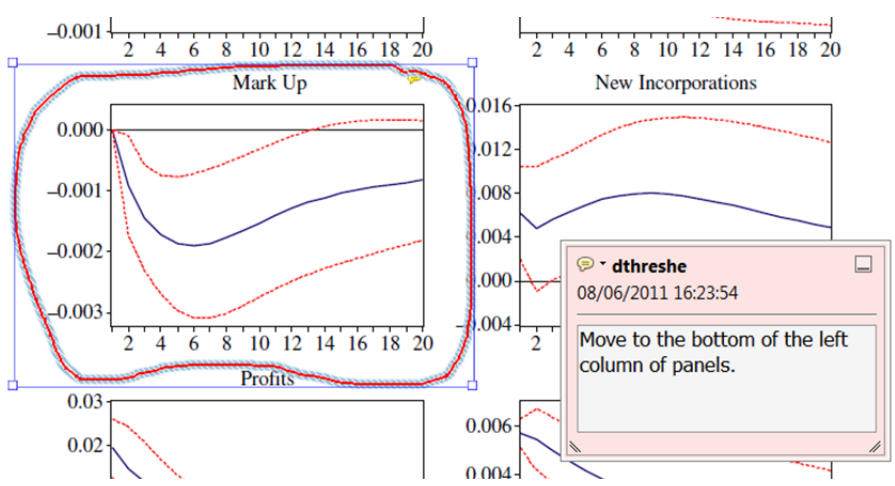


7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

How to use it

- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.



For further information on how to annotate proofs, click on the Help menu to reveal a list of further options: