A System of Automatic Textual Abridgement

Antoinette Renouf
Alex Collier

Abstract

We have developed a system which automatically produces abstracts, or abridgements, from electronic texts. It can abridge any text *with* normal features, ranging from newspaper articles to complete books. The resultant abridgement consists of a set of core information-bearing sentences, the unique feature of which *is* that they together *also form a coherent and readable* mini-text. The system is fast and efficient. Various parameters may be set by the user, including length. Among additional facilities to improve performance is one designed to resolve pronominal or other ambiguity. The system works on other languages, and has applications in IT and beyond. This paper recounts the historical background to our current work.

Keywords

linguistics, NLP, software, text indexing, automatic abstracting, abridgement, information retrieval, translation aid, textual database interface, multilingual interface.

Domains

text indexing, automatic abstracting, information retrieval, writing aids, translation aids, multilingual interfaces.

1. Introduction

This paper will describe for the first time a project in automatic textual abridgement which in fact took place some years ago, based on the work of Hoey (1983, 1988, 1991, 1991). The objective was to take up Hoey's suggestion of automating the manual system of abridgement which he had described. The project was co-directed by Hoey, and carried out in the Research and Development Unit for English Studies when its members were at the University of Birmingham. It was funded by British Telecom.

Hoey had as his starting point an observation gained from teaching English language. This was that non native-speakers build more cohesion into written text than native-speakers, and in a different way. Whereas they tend to create overt links between one sentence and the next, native-speakers tend to make a greater variety of connections, and over significantly longer stretches of text.

In the course of studying the nature of this 'long-distance' cohesion, Hoey noticed some interesting facts about textual organisation: namely, that the most heavily cohesive sentences together incorporate the key points of the original text, that they are central to its thematic development and, moreover, that they together form a kind of coherent summary.

2. The Manual System

The central or key sentences in a given text are related, or 'bonded', together by a series of pairs of repeated words. Each pair of repeated items shared by two sentences is referred to as a 'link' between those sentences. Sentence 'bonds' are represented by the presence of a specific number of individual links. Hoey devised a matrix in which he could, in each cell, record the number of links that one sentence had with another. The matrix could then be used to identify highly bonded sentences, which could be combined to form an abridgement, and longer or shorter abridgements could be created on the basis of the degree of bonding recorded there.

The type of cohesion that Hoey focussed on was lexical rather than grammatical. Lexical cohesion is achieved by reiterating a word, appropriately contextualised, in two or more sentences in a text. There can be different degrees of identity between the first and subsequent instances of a word, so that 'government' can be exactly reiterated as 'government' (known as 'simple repetition'); varied inflectionally, for example as 'governmental' (known as 'complex repetition'); or substituted by a similar word or words such as 'parliamentary' (an instance of 'synonymy' or 'paraphrase'), or 'Tories' (an instance of 'co-reference').

It became clear that the process of identification of lexical patterning across text was capable of automation, particularly in the case of simple and complex lexical repetition, and indeed involves a matching task of the kind that computers are particularly suited to. Similarly, the matrix required for recording the lexical information could readily be constructed on a computer. In the course of the 1980's, therefore, a project was set up to automate Hoey's manual abridgement system.

In the first instance, simple and complex repetition only were taken into account; the lexical features of synonymy and paraphrase were postponed for consideration at a later stage. It was also decided to restrict the first stage of study to non-narrative text, since Hoey had observed that there appeared to be a different organisational principle underlying narrative writing.

One element of grammatical cohesion was made a priority for investigation. Pronominal reference was felt likely to inhibit the interpretability of the abstract wherever it was unresolved; that is to say, where a pronoun such as 'he' appeared in a sentence of the abstract, but the original referent, such as 'Mr Major', did not. This area was studied in the course of the project.

3. Stages Involved in the Creation of an Abridgement

The automation of Hoey's manual procedure was achieved in a number of stages. These were as follows:

3.1. Sentence Identification

The sentence is regarded as the central unit of information in this approach. Sentence boundaries therefore have to be identified accurately. Our experience was that by filtering out 'full-stops' used in decimal numbers or abbreviations a high degree of accuracy could be obtained in this regard simply on the basis of punctuation. Parsing the sentences was felt to be unnecessary and would have had the disadvantage of adding considerably to the processing time.

3.2. The Word Index

Once the sentences had been delimited, a sentence number was assigned to each word, multiple occurrences of the same word were clustered together and their sentence numbers were conglomerated, as in the following example:

apple 3 12 34 53
apples 9 29
banana 4 12 38 53
cherry 3 12 53
.
.
.

Figure 1: *Sample Word Index*

3.3. Stopword Removal

The very frequent words of the language consist primarily of closed sets of grammatical items, such as pronouns, prepositions and articles. We can expect these to occur in nearly every sentence, with the result that they are not the best indicators of cohesive patterns which could uniquely characterise a text. For this reason they were removed from the Word Index at this stage.

3.4. Lemmatisation

In order to be able to detect complex repetition all words were reduced to their base form. Thus, in the word index above, the two lines for 'apple' and 'apples' would be conflated into:

apple 3 912 29 34 53

It was this information which was used to create the connectivity matrix, which is described next.

3.5. The Connectivity Matrix

The main component of the manual analysis is the 'Connectivity Matrix', which holds and displays the number of links that exist between the sentences of the text being processed.

Since a square matrix would be symmetrical about its diagonal, we just used one half of it, as can be seen in the following figure, which is a matrix for a text with twelve sentences:

```
1 1
2 0 2
3 1 0 3
4 0 0 4 4
5 0 0 1 4 5
6 0 0 3 3 1 6
7 1 0 3 5 1 2 7
8 0 0 0 0 0 0 0 8
9 1 0 3 3 1 1 5 0 9
101 0 2 1 1 1 1 0 3 10
11 1 0 2 1 1 1 1 0 2 4 11
12 0 0 3 3 2 1 3 0 3 1 3 12
```

Figure 2: Sample Matrix

Here the numbers down the left-hand side and along the top of the diagonal refer to sentences in the text. Compiling this matrix manually for a long text (ie more than about 60 sentences) can be extremely time-consuming because of the explosion in the number of cells which occurs as the sentence count rises. A text with 100 sentences would therefore create a matrix with around 5,000 cells, making it difficult for the abridger to locate and update the correct cells in the matrix, especially when they lie towards the centre, far distant from the scales at the sides.

As said earlier, the complexity of the task of creating these matrices had made this a prime candidate for automation. The information which sprawls over several sheets of paper could be easily contained and manipulated in the memory of a computer by placing it in a two-dimensional array. If we call this array links, then recording the fact that a link exists between two sentences a and b was merely a case of incrementing the value of links[a,b] . Once the whole text had been processed, the entire contents of the array could be printed out, resulting in a human-readable connectivity matrix for the text. The link information was derived from the lines of the Word Index. If we go back to the example Word Index entry:

apple 3 9 12 29 34 53

it tells us that the occurrence of the lemma 'apple' forms a link between sentences 3 and 9, 3 and 12,3 and 29 .., and between 9 and 12,9 and 29... and between 12 and 29, 12 and 34 and so on, up to and including 34 and 53. Each of these pairs was entered into the matrix according to its subscript or 'coordinates' as given by the two sentence numbers. Thus we would increment links[3,9], links[3,12] links[34,53].

To create the matrix manually for less than half the chapter took one of us many days; the computer created the matrix for the whole chapter in a matter of seconds. However, automating the task of matrix creation did not simplify the task of verifying the results. The abridgement of one chapter of a book, containing some two hundred sentences, resulted in a matrix measuring nearly two metres square, and containing approximately 20,000 cells. The escalation in scale brought about by computerising matrix creation rendered verification nearly impossible. It therefore proved necessary to display the connectivity information in another, less bulky, form. It can be seen in the matrix shown above that many (nearly half in this case) of the cells are empty, ie O. This redundancy was reduced by changing the way the link information was stored within the machine to ignore empty cells. This meant that a small text would now only use a small amount of memory. It also meant that much larger texts could be processed, since not storing empty cells allowed memory to be allocated for many more sentences. Thereafter, the system was run on an entire novel (containing over 4000 sentences) and it produced a matrix of over 8 megabytes in size.

3.6. Building the Abridgement

Once a count of the bonds for each sentence had been established by the software, the user could decide on a suitable threshold for the sentences in the abridgement. Any sentence which had at least this number of bonds would be included in the abridgement. The bond threshold was generally set at three; that is, the software examined the matrix and noted those sentences with three or more bonds. Another pass was then made through the source text and the required sentences were printed out.

If the resulting abridgement was felt to be too long or too short, then the threshold could be raised or lowered accordingly, a new list of sentences created by re-examining the matrix and a new abridgement extracted. This adjustment could be automated to some extent, so that a user of the system could just specify that they wanted an abridgement that is ten per cent of the size of the original, and the software would automatically adjust the threshold until a text of the required size is produced.

4. The Complexity of Manual Summarisation

It is difficult and time-consuming to summarise a text, particularly a long one, without automatic means. The problem is to decide what the text is actually about; 'aboutness' has been shown to be a complex matter (e.g. Van Dijk, 1977; Phillips, 1985). An article will have informative and evaluative strands, which may be lexically interwoven; in addition, the information that is given may relate to more than one aspect of the main topic, if indeed there is a single one. There are multiple issues inherent in any topic, and the writer will invoke those which s/he deems to be salient or attractive to the readership. The summarising process entails making a judgement about which is the primary topic, and how far to attempt to account for its various aspects and sub-topics.

Hoey has carried out studies, with postgraduate student groups, that illustrate the difficulties of summarisation. In testing his own method of abridgement, his subjects have been requested to select a set of key sentences that seem to them to summarise adequately the main conceptual concerns of a given text. His results have revealed very little concensus among individual readers.

5. The Source Text

The sample text that we have selected to demonstrate the automated system in action is taken from the Times Newspaper, 7th November, 1991, and consists of forty-four sentences. It is one of the lead articles and concerns the affairs of the late Robert Maxwell. It reads as follows (sentences are numbered, for ease of reference):

[1] Shares back on market as sons reassure bankers; Death of Robert Maxwell THE HEIRS - SHARES in Maxwell Communication Corporation and Mirror Group Newspapers will recommence trading this morning after a one-and-a-half day suspension triggered by the disappearance at sea and the death of Robert Maxwell.

[2] Kevin Maxwell, Mr Maxwell's youngest son and the new chairman of MCC, said last night that a full statement would be made before the opening of trade.

[3]He assured employees and shareholders that the companies were robust and needed no new refinancing agreements.

[4] He added that reports of crisis meetings throughout the day had been exaggerated.

[5] "I would like to categorically ensure employees and our shareholders that there have been no such meetings," he said.

[6] "The banking arrangements for both public companies are robust."

[7] Shares remained suspended on the Stock Exchange yesterday at the request of the companies, as executives and advisers worked to shore up bank support, and to enable the heavily indebted companies to consult their financial advisers, Samuel Montagu and Rothschild.

[8] Kevin Maxwell spent the afternoon talking to bankers to fend off a possible run on the shares.

[9] Executives are concerned that those banks and brokers who have taken shares in the listed companies as collateral against loans to private Maxwell companies, may sell as soon as trading restarts.

[10] That could push the shares sharply lower and breach loan covenants with other banks.

[11] Estimates of the debts within both the private and public Maxwell companies exceed Pounds 2 billion despite asset sales of almost Pounds I billion this year.

[12] When they were suspended, shares in both companies were at their low for the year.

[13] Moreover, MCC ceased trading at 121p, down 18p on the day after the American investment bank, Goldman Sachs, sold part of its stake.

[14] The Stock Exchange said it was making routine inquiries into such a sharp fall.

[15] Kevin and his brother Ian were confirmed as chairmen of the two listed companies at board meetings yesterday, after being appointed acting chairmen immediately after their father's disappearance.

[16] Ian is chairman and publisher of Mirror Group Newspapers.

[17] Kevin Maxwell is now chairman and chief executive of Maxwell Communication Corporation, and assumes the chairmanship of all MCC subsidiaries previously held by Robert Maxwell.

[18] He was also appointed chairman of the New York Daily News.

[19] Peter Laister, already a non-executive director of MCC, was appointed deputy chairman.

[20] Bankers said the MCC board must increase the number of non-executive directors as soon as possible to keep within general principles of corporate governance and throw open the company to outside scrutiny as never before.

[21] Mr Maxwell's concern about keeping his personal finances away from public scrutiny is one of the reasons bankers are skittish.

[22] In Mr Maxwell's case, his family finances are inextricably bound to the financial make-up of the public companies.

[23] Family holding companies control 51 per cent of Mirror Group and more than 60 per cent ofMCC.

[24] Stockbrokers consider that Mirror Group, which was floated earlier this year, should fare better than MCC when they are requoted.

[25] Mirror Group's debts stand at about Pounds 280 million but at the time of the float, Samuel Montagu said it had "ring-fenced" the new company.

[26] Strategically, it is likely the company's advisers will suggest the Maxwell brothers wind down their stake in Mirror Group and seek ways of selling more of MCC's assets in America.

[27] These include Macmillan the publishers, Official Airline Guide, Berlitz and Que Software.

[28] Reports from New York said Kohlberg Kravis Roberts, the leveraged buyout specialist, was reported to be taking another look at the language publisher Berlitz.

[29] KKR looked at adding Berlitz to its rapidly expanding publishing business last spring, and is still believed to be a potential buyer.

[30] MCC floated 44 per cent of Berlitz two years ago at around $16 per share.

[31] Trading in those shares also remained suspended yesterday at $19.50.

[32] lan Maxwell, referring to legal actions arising out of allegations by Seymour Hersh, the American author, that his father had in some way been linked with Mossad, the Israeli intelligence agency, said: "It is absolutely clear that the legal steps taken by MirrorGroup Newspapers will be pursued vigorously."

[33] He and his brother said they recognised the immensity of the task of stepping into their father's shoes.

[34] lan said: "It has been an incredibly busy time.

[35] I have not had time to reflect or deal with emotions other than to recognise the responsibility to thousands of employees, shareholders, advertisers and suppliers.

[36] For MGN, it will be business as usual."

[37] Kevin Maxwell emphatically denied suggestions that his father's company was a complex and tangled web.

[38] He said: "This is far more complex than a corner shop.

[39] It is simply a very large organisation built up over many years."

[40] Neither brother could explain why the shares in the companies had dropped significantly before their father died.

[41] lan Maxwell said their mother was treating her loss in a dignified and extremely composed manner.

[42] They were a large family and the strength of the love they had for one another was helping them through.

[43] He confirmed that samples of body tissue from Mr Maxwell were being sent to Oxford for analysis.

[44] He added: "I have absolutely no suspicion about his death."

The main topic of the above text can probably best be described as 'the consequences of Robert Maxwell's death'. There have been many consequences, and some immediate ones are focussed on in the text. Two of them concern the effect of Maxwell's death on the financial stability of the Maxwell groups and of his family. Within these two areas, a series of details or sub-topics are introduced: on the company front, individual companies within the larger groups and their possible fates are mentioned; on the family front, the recent moves and new roles of the Maxwell brothers are outlined. As stated in sentence 22, '...his family finances are inextricably bound to the financial make-up of the public companies', and this relationship is reflected in the way information about the two is presented.

A third theme running through the text concerns the suspicion in which Maxwell and his sons were held. Various manifestations of this suspicion are detailed, including the suspension of shares and the concern of executives. Somewhat incongruously, the suspicion of Maxwell's links with Mossad is also raised. At this point in the text, there appears to be a shift into a fourth theme, which concerns the personal attitudes and feelings of the Maxwell family.

This brief and superficial analysis of the text is intended as a preface to the commentary on the adequacy of the abridgements produced by the automated system that follow. The aim has been to identify the main components of the source text, and to give some idea of the task facing the human or the machine in attempting to produce a fair representation of it.

6. The Automatic Abridgements

We have applied the abridgement software to the source text and produced three sample abridgements. The first is generated using a link: bond ratio of 3, and a bond threshold of 4, and consists of eight sentences. The second has the same link: bond ratio, but a higher bond threshold, of 6, and is therefore shorter, at four sentences. The third abridgement differs from the first two, in having a link: bond ratio of 4, and a low bond threshold, of 2. It contains six sentences. We shall refer to these abridged texts by their parameters: namely as '3/4', '3/6' and '4/2' respectively.

They look as follows:

Abridgement 3/4

[I] Shares back on market as sons reassure bankers; Death of Robert Maxwell THE HEIRS – SHARES in Maxwell Communication Corporation and Mirror Group Newspapers will recommence trading this morning after a one-and-a-half day suspension triggered by the disappearance at sea and the death of Robert Maxwell.

[7] Shares remained suspended on the Stock Exchange yesterday at the request of the companies, as executives and advisers worked to shore up bank support, and to enable the heavily indebted companies to consult their financial advisers, Samuel Montagu and Rothschild.

[9] Executives are concerned that those banks and brokers who have taken shares in the listed companies as collateral against loans to private Maxwell companies, may sell as soon as trading restarts.

[11] Estimates of the debts within both the private and public Maxwell companies exceed Pounds 2 billion despite asset sales of almost Pounds 1 billion this year.

[24] Stockbrokers consider that Mirror Group, which was floated earlier this year, should fare better than MCC when they are requoted.

[25] Mirror Group's debts stand at about Pounds 280 million but at the time of the float, Samuel Montagu said it had "ring-fenced" the new company.

[26] Strategically, it is likely the company's advisers will suggest the Maxwell brothers wind down their stake in Mirror Group and seek ways of selling more of MCC's assets in America.

[32] Ian Maxwell, referring to legal actions arising out of allegations by Seymour Hersh, the American author, that his father had in some way been linked with Mossad, the Israeli intelligence agency, said: "It is absolutely clear that the legal steps taken by Mirror Group Newspapers will be pursued vigorously."

Abridgement 3/4 focuses on the financial upheavals and uncertainties caused by Maxwell's death, and is generally satisfactory, at least as a summary of that aspect of the text. It does not touch on family matters except in its final sentence. This sentence has been characterised earlier by us as being somewhat incongruous within the text, but it appears here in the abridgement on the strength of certain of its lexical items, such as 'Maxwell', 'legal', 'Mirror', 'Group' and 'father.

Abridgement 3/6

[1] Shares back on market as sons reassure bankers;Death of Robert Maxwell THE HEIRS - SHARES in Maxwell Communication Corporation and Mirror Group Newspapers will recommence trading this morning after a one-and-a-half day suspension triggered by the disappearance at sea and the death of Robert Maxwell.

[9] Executives are concerned that those banks and brokers who have taken shares in the listed companies as collateral against loans to private Maxwell companies, may sell as soon as trading restarts.

[25] Mirror Group's debts stand at about Pounds 280 million but at the time of the float, Samuel Montagu said it had "ring-fenced" the new company.

[26] Strategically, it is likely the company's advisers will suggest the Maxwell brothers wind down their stake in Mirror Group and seek ways of selling more of MCC's assets in America.

Abridgement 3/6, with its higher bond threshold, is reduced to four sentences. It is also satisfactory as a summary of the financial aspects of the main topic, although in losing sentence [24] it has lost the reference to MCC that was a counterpart to the reference (sentence [25]) to Mirror Group, the other Maxwell company. On the other hand, it has also discarded the extraneous reference to Mossad that occurs in Abridgement 3/4.

Abridgement 4/2

[I] Shares back on market as sons reassure bankers;Death of Robert Maxwell THE HEIRS - SHARES in Maxwell Communication Corporation and Mirror Group Newspapers will recommence trading this morning after a one-and-a-half day suspension triggered by the disappearance at sea and the death of Robert Maxwell.

[7] Shares remained suspended on the Stock Exchange yesterday at the request of the companies, as executives and advisers worked to shore up bank support, and to enable the heavily indebted companies to consult their financial advisers, Samuel Montagu and Rothschild.

[9] Executives are concerned that those banks and brokers who have taken shares in the listed companies as collateral against loans to private Maxwell companies, may sell as soon as trading restarts.

[17] Kevin Maxwell is now chairman and chief executive of Maxwell Communication Corporation, and assumes the chairmanship of all MCC subsidiaries previously held by Robert Maxwell.

[26] Strategically, it is likely the company's advisers will suggest the Maxwell brothers wind down their stake in Mirror Group and seek ways of selling more ofMCC's assets in America.

[32] Ian Maxwell, referring to legal actions arising out of allegations by Seymour Hersh, the American author, that his father had in some way been linked with Mossad, the Israeli intelligence agency, said: "It is absolutely clear that the legal steps taken by Mirror Group Newspapers will be pursued vigorously."

Abridgement 4/2 presents a slightly different account. It touches on two aspects of the main topic: the financial and also something of the familial. The hitherto problematic sentence [32] referring to Mossad also mentions Ian Maxwell and so is here to some extent integrated thematically into the abridged text, through the counterbalance provided by the reference to Kevin Maxwell in sentence [17].

7. General Comments on the Automatic Abridgements

The three sample abridgements differ in length and kind. However, they share three sentences: [I], [9] and [26].

The fact that they all contain the initial sentence of the text is not surprising. Firstly, the headline has been incorporated into the first sentence, which is therefore favoured by the methodology, in that it becomes a particularly lexically-rich sentence. The incorporation is not universal but is the result of inconsistent coding in the newspaper text. But the accident is a happy one, because it coincides with an important fact about the composition of newspaper articles: that the text as a whole is typically summarised in its first sentences. This journalistic convention is confirmed by Gerry Kreibich (1991):

'One oft-repeated training hint (for junior journalists) goes like this: 'Imagine someone shouts across the room to ask you what the story is that you are working on at that moment.

Whatever you instinctively shout back will probably be the right intro and will simply need tidying up."

Knowledge of the special status of the opening sentence or sentences of a newspaper text has for some time also been exploited by a number of companies producing text retrieval software, such as BRS Software and Profile, who, in their training sessions, point the user to the fact that first paragraph search might be a useful alternative to simple keyword search in the retrieval of relevant texts.

The three shared sentences must be considered as particularly key ones in the text. It is interesting to note that neither [9] nor [26] is as long as other sentences that have not been singled out, since this indicates that it is the nature of their lexical content and not simply the number of words that they contain that marks them out as key sentences.

It is clear from the sample abridgements above that each adjustment to the parameters will bring a different focus and slant to the abridged text. Response is almost instantaneous, so a preferred version can easily be found. Sometimes an abridgement with a high evaluative content is more appropriate, or one picking up on a secondary theme.

The flexibility of the system at this stage of development is restricted in that it takes just simple and lexical repetition into account. Results would obviously be different once thesaural and instantial relations had been taken into account. The beauty of the system as outlined, however, resides in its simplicity.

8. Conclusion

Hoey's abstracting system has many applications in the field of text retrieval, some of which are being pursued by the Research Unit. Obvious areas of interest include the adaptation of the system to accommodate other text types and text sources. The system has been found to work on other European languages, and the multi-lingual aspects are being further investigated.

Acknowledgment

We would like to express our gratitude to British Telecom for their fmancial support, without which the early phases of this project could not have been carried out.

Bibliography

Hoey, M P (1983) On the Surface of Discourse, London: George Allen and Unwin; reprinted 1991, Nottingham: English Language Studies, University of Nottingham.

Hoey, M P (1988) The Clustering of Lexical Cohesion in Non-narrative Text, in Trondheim Papers in Applied Linguistics 4, pp 154-180.

Hoey, M P (1991) Another Perspective on Coherence and Cohesive Harmony, in Functional and Systemic Linguistics: Approaches and Uses (ed. E Ventola), Berlin: Mouton de Gruyter, pp 385-414.

Hoey, M P (1991) Patterns of Lexis in Text, Oxford: OUP.

Kreibich, G (1991) Tell it like it is, in Police Review, 13 December, pp 2476-2477.

Phillips, M K Aspects of Text Structure: An investigation of the lexical organisation (1985) of text, Amsterdam, North Holland.

Van Dijk, T A Text and Context: Explorations in the Semantics and Pragmatics of (1977) Discourse, London: Longman.