A Financial News Summarisation System based on Lexical Cohesion

Paulo Cesar Fernandes de Oliveira, Khurshid Ahmad, Lee Gillam

Department of Computing
School of Electronics, Computing and Mathematics
University of Surrey
Guildford, Surrey, England, GU2 7XH
{p.fernandes-de-oliveir, k.ahmad, l.gillam}@surrey.ac.uk

Abstract

Summarisation is important for the dissemination of information and forms an integral part of work in information extraction. A collection of keywords can be treated as a summary of a text, an abstract is essentially a summary of a (learned) text, and a headline can also be deemed as a summary. The automatic generation of summaries typically relies on the statistical distribution of terms in a given text or on the linguistic/pragmatic behaviour of key terms in the text. The latter is the subject of this paper: how can we use linguistic devices that are supposed to engender *cohesion* in a text for producing summary of the text. Michael Hoey (1991) proposed lexical cohesion as a basis for producing summaries of non-narrative texts. We have used lexical cohesion to produce summaries of narrative texts, namely financial news. SummariserPort is a system that produces summaries of financial texts and is currently being evaluated.

1. Introduction

The rapid growth of information due to the World Wide Web and on-line information services, have increased demands for effective and efficient methods of accessing and condensing texts. The growth in the volumes of financial news available has also grown in line with these other information sources. Company reports, analysts' reports, on-line briefings and a host of discussions are readily available on financial websites and can be delivered to your email. According to America's number-one money manager Peter Lynch (Lynch, p.20) "Stock market news has gone from hard to find (in the 1970s and early 1980s), then easy to find (in the late 1980s), then hard to get away from". We can argue that for the 00s, it is hard to find the news in the constant commentary.

It is possible to retrieve vast quantities of materials using Information Retrieval (IR) as evidenced through a variety of search engines. One further consequence of this growth is the need for some form of text summarization such that the retrieved content can be reduced to a readable volume. In this respect, automatic text summarization has increasingly stimulated interest in both research and commercial sectors.

The main goal of automatic summarization is to take an information source, extract content from it, and show or present to the user the most important part in a condensed form according to the user's needs, and in a sensitive manner (Mani, 2000). A summary generated by a computer program that can mimic the summarizing skills of a human being would have a big commercial and scientific value (e.g. financial market) (Benbrahim, 1996).

In short, a summary acts as a filter, indicating the major content of the original story. A good summary will provide the reader with an indication of whether the whole document is worth reading. A wide variety of texts can benefit from summarization, including newspapers, journals, press releases, scientific reports, organisational memos, etc.

Automatic summarization has been investigated since the 1950's in psychology, linguistics and information science (e.g. Kieras (1985); van Dijk (1980)). Current research tends to be carried out in AI where linguistic approaches are combined.

Our approach is based on a linguistic theory of text organisation called lexical cohesion (Hoey, 1991). This approach uses the more frequent words of the text, their variants and conceptual relationships to establish links between sentences in the text. Frequency information (repetition) is used to compute the strength of association between sentences to identify the sentences which best represent the message of the original text – an indicative summary. The sentences that associate most strongly are then selected for the summary. As this approach is based on frequency of recurrence, the process can be applied to various languages. Our system has summarized English documents and also Brazilian Portuguese texts.

2. Lexical Cohesion

One important reference on cohesion is Halliday and Hasan (1976). They define lexical cohesion as 'selecting the same lexical item twice, or selecting two that are closely related' (p.12). This observation suggests two interlinked conclusions. First, that cohesion focuses on repetition across sentences; and second, that it is possible to study lexical cohesion by studying repetition.

In their study, two important concepts were created: tie (p.3) which is 'a single instance of cohesion'; and texture (p.2) which is defined as the property of 'being a text'. In other words, for Halliday and Hasan, the organization of text (which is texture) is composed of relationships amongst items in the text; some are semantic, some are grammatical and they refer to these as cohesive ties. Table 1 below show the taxonomy.

Grammatical Cohesion	Reference	Implies that the information is to be retrieved through the reference item is the referential meaning (pronouns and determiners – personal, demonstrative and demonstrative)	
	Substitution	Refers to the process of replacing one expression by another (e.g. Did Susie finish her homework in time? – I don't think *so*.	
	Ellipsis	Designates the process of substituting something by nothing (e.g. How many potatoes do you want, Sir? – Four [], please.	
	Conjunction	Covers the use of adjunct-like elements to mark the semantic relationships (e.g. He didn't study much. Nevertheless, he passed all his exams.	
Lexical Cohesion	Reiteration General*	Repetition	Suggests the lexical item is repeated
		Synonymy	Relates lexical items which have the same meaning
		Antonymy	Constrasts between a term which is the opposite of another
		Hyponymy	Relates specific and general lexical items, such that the former is included in the latter
		Meronymy	Relates parts and wholes

Table 1: Classes according to Halliday and Hasan study (1976)
*Hasan (1984) systematized her later study (1976) with a new categorization of these classes to compensate the loss of clarity. There is another device for lexical cohesion called collocation which involves equivalence, naming and semblance, but has not been developed to the same extent as reiteration.

Michael Hoey (1991), in his *Patterns of Lexis in Text*, made a claim about the way text is organized. He demonstrated that lexical repetition is the principal means of explicitly marking cohesion in a text and illustrates that lexical cohesion forms clusters among sentences. He stressed, using Halliday and Hasan's example texts, that the most dominant type of cohesion in English is lexical cohesion (over of 40% of the ties are lexical).

Hoey (1991) categorizes repetition into different lexical types. These includes:

- simple repetition – two identical items (e.g. bear – bear) or two similar items whose difference is 'entirely explicable in terms of a closed grammatical paradigm' (e.g. bears (N) – bears (N)) (p.53);
- complex repetition – which results from two items sharing a lexical morpheme but differing with respect to other morphemes or grammatical function (e.g. human (N) – human (Adj.), dampness – damp);
- simple paraphrase – two different items of the same grammatical class which are 'interchangeable in the context' (p.69) and 'whenever a lexical item may substitute for another without loss or gain in specificity and with no discernible change in meaning'. (p.62). (e.g. sedated – tranquillised).
- complex paraphrase – Two different items of the same or different grammatical class; this is restricted to three situations:
  (i) antonyms which do not share a lexical morpheme (e.g. *hot – cold*);
  (ii) two items one of which 'is a complex repetition of the other, and also a simple paraphrase (or antonym) of a third' (p.64). (e.g. a complex paraphrase is recorded for 'record' and 'discotheque' if a simple paraphrase has been recorded for 'record' and 'disc', and a complex repetition has been recorded for 'disc' and 'discotheque';
  (iii) when there is the possibility of substituting an item for another (for instance, a complex paraphrase is recorded between 'record' and 'discotheque' if 'record' can be replaced with 'disc'.

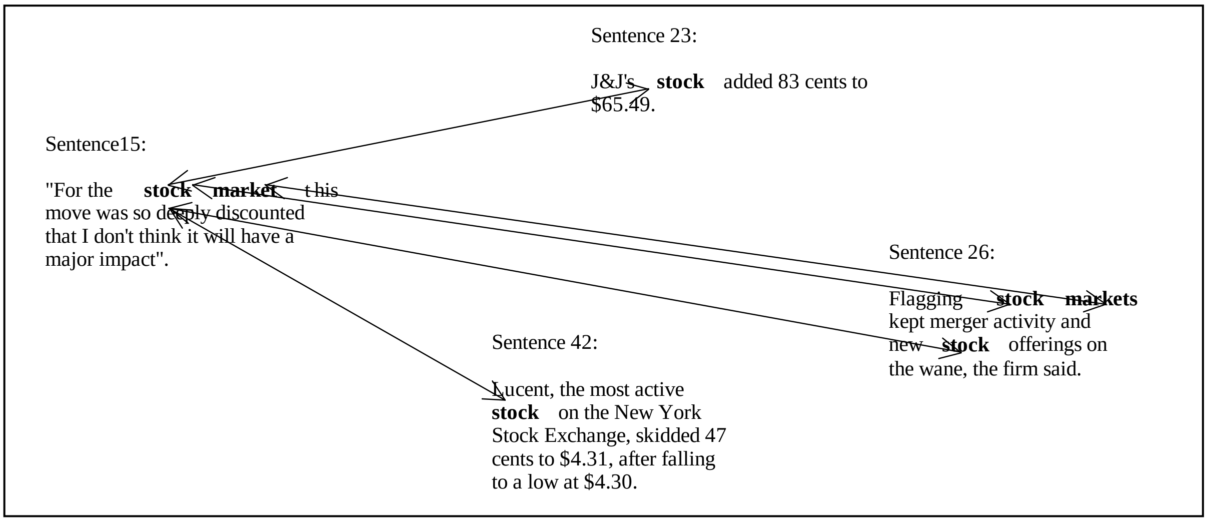
Hoey proposed two key notions, links and bonds, and repetition matrices, which are used to establish the number of connections between sentences. We explain these in the following sections.

2.1 Links and Bonds

Links occur whenever there is a repetition of an item in two separate sentences. Hoey used 'link' as a version of Halliday and Hasan's concept of tie. He argued that he does not use 'tie' because it includes certain kinds of cohesion devices which do not count towards links (e.g. conjunctions, collocation). To illustrate the concept of links, we show an excerpt of a financial news file[1] collected from Reuters' Website[2] (see Figure 1).

[1] Text title: U.S. stocks hold some gains. Collected from Reuters' Website on 20 March 2002.

[2] Reuters's Website: see **http://www.reuters.co.uk**



Sentence 23:

J&J's **stock** added 83 cents to $65.49.

Sentence15:

"For the **stock market** this move was so deeply discounted that I don't think it will have a major impact".

Sentence 26:

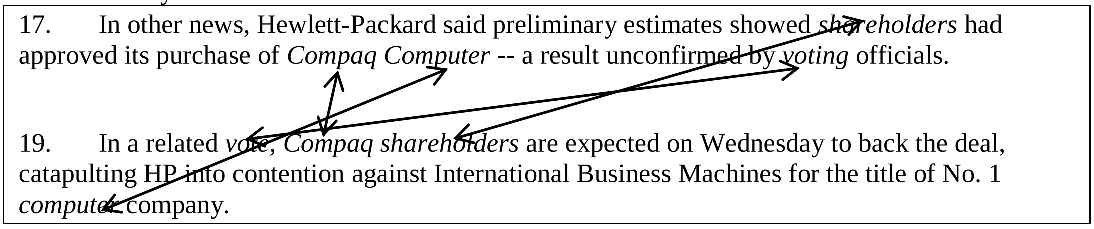
Flagging **stock markets** kept merger activity and new **stock** offerings on the wane, the firm said.

Sentence 42:

Lucent, the most active **stock** on the New York Stock Exchange, skidded 47 cents to $4.31, after falling to a low at $4.30.

Figure 1: Links between 'stock' and 'markets' across sentences

Hoey also proposes ‘bonding’ to account for relations between sentences. A bond is established whenever there is an above-average degree of linkage between two sentences. It can be defined as ‘a connection between any two sentences by virtue of there being a sufficient number of links between them’ (p.91). Normally, three links constitute a bond. Hoey stresses that the number of links which constitute a bond is relative to the type of text and to the average number of links in the text (p.91), but the least number of links is three ‘because of the greater likelihood of two repetitions occurring in a pair of sentences by chance’ (p.190). For example, the two sentences in the Figure 2 are bonded by four links.



17. In other news, Hewlett-Packard said preliminary estimates showed *shareholders* had approved its purchase of *Compaq Computer* -- a result unconfirmed by *voting* officials.

19. In a related *vote*, *Compaq shareholders* are expected on Wednesday to back the deal, catapulting HP into contention against International Business Machines for the title of No. 1 *computer* company.

Figure 2: Example of bonded sentences

2.2 Repetition Matrices

Hoey suggested representation of the summary links using a ‘repetition matrix’. The links between pairs of sentences due to repetition of a specific item can be represented in the form of a matrix, where the rows and columns represent the sentence numbers, and the elements (cells) show the number of links between the sentences. The rows represent links with subsequent sentences, the columns links with previous sentences. An extract of the link matrix for the news-wire text mentioned above is presented in Table1. It shows that, for instance, sentence 2 has 1 link with sentences 15, 17 and 20; and shares no links with sentences 3 to 14. In the other hand, sentence 18 has 5 links with sentence 19, 2 links with sentence 20, and so forth.

i j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
1		4	5	1	2	0	2	0	0	0	0	2	0	0	2	0	1	0	0	1	
2			0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	
3				2	1	1	2	0	0	0	0	0	0	2	0	0	0	0	0	0	
4					1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
5						0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
6							1	0	0	0	0	0	0	1	0	0	0	0	0	0	
7								0	0	0	0	0	0	2	0	0	0	0	0	0	
8									0	0	0	0	0	0	0	0	0	0	0	0	
9										1	1	0	0	0	0	0	0	0	0	0	
10											2	0	0	0	0	0	0	0	0	1	
11												0	0	0	0	0	0	0	0	0	
12													0	0	0	0	0	0	0	0	
13														0	0	0	0	0	0	0	
14															0	0	0	0	0	0	
15																0	0	0	0	0	
16																	0	2	0	0	
17																		3	4	1	
18																			5	2	
19																				2	
20																					

…

Table 1: An extract (one quarter) from a 43x43 link matrix of a text entitled ‘U.S. stocks hold some gains’ (20th March 2002). Only the upper part is represented because the matrix is symmetric

There is a variation in the number of links between sentences. The majority of elements of this matrix demonstrate an absence of common terms between the sentences. There are, however, some sentences connected by an appreciable number of links and these are the ones we want to retain for an abridgement of the text.

When two sentences have an above-average number of links, they are said to form a bond. The cut-off point for bond information, the link threshold, depends on the length and type of text but should never be less than 3 to avoid accidental repetition. A link matrix can therefore give rise to a bond matrix, a table of 1s and 0s, denoting either an existence or an absence of bonds between the sentences. Table 2 is the bond matrix correspondent to the link matrix of Table 1 for a link threshold of 3.

i j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
1		**1**	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3				0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5						0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6							0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7								0	0	0	0	0	0	0	0	0	0	0	0	0	
8									0	0	0	0	0	0	0	0	0	0	0	0	
9										0	0	0	0	0	0	0	0	0	0	0	
10											0	0	0	0	0	0	0	0	0	0	
11												0	0	0	0	0	0	0	0	0	
12													0	0	0	0	0	0	0	0	
13														0	0	0	0	0	0	0	
14															0	0	0	0	0	0	
15																0	0	0	0	0	
16																	0	0	0	0	
17																		**1**	**1**	0	
18																			**1**	0	
19																				0	
20																					
…																					

Table 2: The bond matrix corresponding to the link matrix of Figure3 with link threshold = 3

2.3 Central, Marginal, Topic-opening and Topic-closing Sentences

Central sentences are those which have a high number of bonds, according to Hoey,'the most bonded sentences' in the text (p.265). Marginal sentences are 'sentences that form no bonds or, for some texts, few bonds' (p.267). A sentence is topic-opening if it bonds with more subsequent than preceding sentences, and it is topic-closing if it bonds more times with preceding sentences. The combination of topic-opening, topic-closing and 'most bonded' sentences can be used to automatically produce an indicative summary. The example provided above has a majority of marginal sentences and a few topic opening and topic closing sentences. These can be used to produce an indicative summary of this text.

3. SummariserPort – a Summariser System

Based on the principles of lexical cohesion outlined above, we have developed a system for automatic text summarisation that we shall describe in this section.

3.1 Technical Description

SummariserPort has been developed from a prototype text-processing program called Tele-Pattan (Benbrahim, 1996). It was written in Java and represents a computer implementation of two of Hoey's four categories of lexical repetition, Simple Repetition and Complex Repetition. The architecture of SummariserPort is shown in Figure 3.

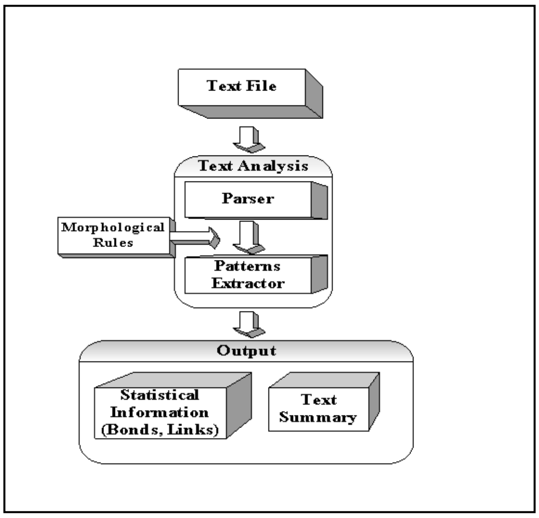


Figure 3: The Architecture of SummariserPort

The description of this architechture is provided below:

Parser: this module starts by reading the text file and segmenting it into sentences. We used a Java class designed specifically to parse natural language into words and sentences. It is called BreakIterator. Some features are (1) it has a built-in knowledge of punctuation rules; (2) it does not require any special mark-up.

Morphological Rules: In this module the second category of Hoey's (1991) approach is performed. The instances of complex repetition are looked up by means of a list of derivational suffixes encoded into the program. For the English language, it contains 75 morphology conditions that lead to approximately 2500 possible relations among words. Such repetition (complex) is counted as links.

Patterns Extractor: performs analysis of simple repetition, counting lexical links between sentences. It uses an optional stop list containing closed class words and other non-lexical items, such as pronouns, prepositions, determiners, articles, conjunctions, some adverbs, etc., which are excluded from forming links.

Output: produces the results. One result of this module is the summary itself. Other outputs include: link matrix; bond matrix; a histogram with number of occurrences of each link; total number of sentences, words and links; list of sentences grouped into categories of topic-opening, topic-closing and central and the word frequency list.

3.2 Summary Generation

Our program was set to produce summaries consisting of sentences from categories mentioned in the section 2.3, namely topic-opening, topic-closing and most-bonded. Within each category the sentences are ranked according to the number of bonds each sentence has with others sentences in the text. The indicative summary generated is approximately 30% of the size of the original text, which is produced by taking 10% of the number of sentences for each category.

4. Evaluation

The program itself produces summaries in parts of a second once it is initialised and hence is suitable for high-volume throughput such as that from a news-feed. The process of evaluating a summary remains the most controversial part of such research. Authors witing about evaluation stress and state such problems (e.g. Karen Sparck Jones (1994), Edmunson (1969); Paice (1990) and Hand (1997)).

For us, the following aspects are relevant to the evaluation process for a summary: (a) the inclusion of the essential information; (b) the exclusion of non-essential information and (c) the readability of the summaries.

We considered that the inclusion of the relevant information's criterion forms the basis of our evaluation procedure.

Our corpus contains 623 news-wire financial texts collected from Reuter's Website on January 2002. For evaluation purposes, we chose randomly 5 different texts files. Each of these files were summarised by SummariserPort. In other words, a summary for each text was produced and its quality was evaluated against the summaries of the other texts.

A trial questionnaire was created for evaluating the quality of summary produced from these files that were evaluated by two test groups, PhD students from the University of Surrey and Financial Traders from JRC, Berlin.

Both the students and the traders were asked to rank the summary of each text. The results are tabulated in the tables below (Table 3 represents student results, Table 4 represents Trader results):

Rank the summaries from the worst to the best. (Obs: a rating scale from 1 = best to 5 = worst was used)					
	S_1	S_2	S_3	S_4	S_5
Student 1	1	3	2	4	5
Student 2	1	3	2	4	5
Student 3	3	1	2	4	5
Student 4	4	2	3	5	1
Student 5	1	3	2	4	5
Average	2	2.4	2.2	4.2	4.2

Table 3: Student Evaluation of summaries

Rank the summaries from the worst to the best (Obs: a rating scale from 1 = best to 5 = worst was used)					
	S_1	S_2	S_3	S_4	S_5
Trader 1	1	3	5	4	2
Trader 2	3	4	5	1	2
Trader 3	5	2	1	3	4
Trader 4	1	3	5	4	2
Average	2.5	3	4	3	2.5

Table 5:Trader evaluation of summaries

The tables above show the subjectivity inherent in such an evaluation. Individuals judge summaries differently depending on their expectations. However, in the majority of cases, Summary S_1 was the most popular. A question arises here. Why S_1 was considered the best among the others? We believe that the answer is because human evaluation is subjective and normally depends on the taste and motivation of the evaluator.

During the evaluation process, the Traders identified sentences they wanted to appear in the summary, but which were not included. This result is shown in Table 6.

	S_1	S_2	S_3	S_4	S_5
Trader 1	-7				
Trader 2		-12, -17		-2	
Trader 3	-2, -7, -10				
Trader 4	-7	-17		-1, -2	

Table 6: Missing Sentences (-)

We are currently investigating the reasons these sentences, while important to the traders, were missing from the summary.

The evaluation work is in progress and will be reported elsewhere[3] (Oliveira and Ahmad, in prep.).

5. Conclusions

According to the outcomes provided by the experiments results, we can now draw some conclusions:

- SummariserPort proved to be robust, fast, reliable and is able to produce summaries of very high quality;
- Our system empirically supports Hoey's theory about the organisation of the text, i.e. lexical cohesion;
- The evaluation process does not stop here. We are evaluating our system continuously. We are also trying another kinds of evaluations which are reported on the literature in order to obtain better results (e.g . task-based evaluation, target-based evaluation and automatic evaluation).

Acknowledgements

The PhD student evaluators and JRC Berlin, in conjunction with the authors, acknowledge the support of the EU-cofunded project Generic Information-based Decision Assistant (GIDA IST-2000-31123).

References

Benbrahim, M., 1996. *Automatic text summarisation through lexical cohesion analysis*. PhD thesis. Artificial Intelligence Group. Department of Computing. University of Surrey. Guilford.

Edmunson, H.P., 1969. New methods in automatic abstracting. *Journal of ACM*. 16(2):264-285.

Halliday, M.A.K. and Hasan, R., 1976. *Cohesion in English*. London and New York: Longman.

Hand, T.F., 1997. A proposal for task-based evaluation of text summarization systems. In *ACL/EACL-97 summarization workshop*. p.31-36.

Hasan, R., 1984. Coherence and cohesive harmony. In J. Flood (ed.), *Understanding reading comprehension: Cognition, language and the structure of prose*. Newark: International Reading Association. p.181-219.

Hoey, M., 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Jones, K.S., 1994. Towards better NLP system evaluation. In *Proceedings of the human language technology workshop*. p.102-107. San Francisco: ARPA.

Kieras, D.E., 1985. Thematic processes in the comprehension of technical prose. In B.K. Britton and J.B.Black (eds.) *Understanding expository text: A*

[3] The document will appear on our website: (http://www.computing.ac.uk/ai/gida)

theoretical and practical handbook for analysing explanatory text. Hillsdale, NJ: Lawrence Erlbaum, p.89-107.

Lynch, P. 2000 *One up on Wall Street*. Simon & Schuster, New York

Mani, I., 2000. *Automatic Summarization*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Oliveira, P.C.F. and Ahmad, K. (in prep.). Evaluating summaries: methods and techniques.

Paice, C.D., 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*. 26(1):171-186.

van Dijk, T.A., 1980. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction and cognition*. Hillsdale, NJ: Lawrence Erlbaum.