# Ranking Text Units According to Textual Saliency, Connectivity and Topic Aptness

**Antonio Sanfilippo***
LINGLINK
Anite Systems
13 rue Robert Stumper
L-2557 Luxembourg

## Abstract

An efficient use of lexical cohesion is described for ranking text units according to their contribution in defining the meaning of a text (textual saliency), their ability to form a cohesive subtext (textual connectivity) and the extent and effectiveness to which they address the different topics which characterize the subject matter of the text (topic aptness). A specific application is also discussed where the method described is employed to build the indexing component of a summarization system to provide both generic and query-based indicative summaries.

## 1 Introduction

As information systems become a more integral part of personal computing, it appears clear that summarization technology must be able to address users' needs effectively if it is to meet the demands of a growing market in the area of document management. Minimally, the abridgement of a text according to a user's needs involves selecting the most *salient* portions of the text which are *topically* best suited to represent the user's interests. This selection must also take into consideration the degree of *connectivity* among the chosen text portions so as to minimize the danger of producing summaries which contain poorly linked sentences. In addition, the assessment of textual saliency, connectivity and topic aptness must be computed efficiently enough so that summarization can be conveniently performed on-line. The goal of this paper is to show how these objectives can be achieved through a conceptual indexing technique based on an efficient use of *lexical cohesion.*

## 2 Background

Lexical cohesion has been widely used in text analysis for the comparative assessment of saliency and connectivity of text fragments. Following Hoey (1991), a simple way of computing lexical cohesion in a text is to segment the text into units (e.g sentences) and to count *non-stop* words[1] which co-occur in each pair of distinct text units, as shown in Table 2 for the text in Table 1. Text units which contain a greater number of shared non-stop words are more likely to provide a better abridgement of the original text for two reasons:

- the more often a word with high informational content occurs in a text, the more topical and germane to the text the word is likely to be, and
- the greater the number of times two text units share a word, the more connected they are likely to be.

Text saliency and connectivity for each text unit is therefore established by summing the number of shared words associated with the text unit. According to Hoey, the number of *links* (e.g. shared words) across two text units must be above a certain threshold for the two text units to achieve a lexical cohesion rank. For example, if only individual scores greater than 2

* This work was carried out within the *Information Technology Group* at *SHARP Laboratories of Europe, Oxford, UK.* I am indebted to Julian Asquith, Jan IJdens, Ian Johnson and Victor Poznański for helpful comments on previous versions of this document. Many thanks also to Stephen Burns for internet programming support, Ralf Steinberger for assistance in dictionary conversion, and Charlotte Boynton for editorial help.

[1] Non-stop words can be intuitively thought of as words which have high informational content. They usually exclude words with a very high fequency of occurrence, especially closed class words such as determiners, prepositions and conjunctions (Fox, 1992).

#1# Apple Looking for a Partner
#2# NEW YORK (Reuter) - Apple is actively looking for a friendly merger partner, according to several executives close to the company, the New York Times said on Thursday.
#3# One executive who does business with Apple said Apple employees told him the company was again in talks with Sun Microsystems, the paper said.
#4# On Wednesday, Saudi Arabia's Prince Alwaleed Bin Talal Bin Abdulaziz Al Saud said he owned more than five percent of the computer maker's stock, recently buying shares on the open market for a total of $115 million.
#5# Oracle Corp Chairman Larry Ellison confirmed on March 27 he had formed an independent investor group to gauge interest in taking over Apple.
#6# The company was not immediately available to comment.

Table 1: Sample text with numbered text units

| *Text units* | | *Words shared* | *Score* |
|---|---|---|---|
| #1# | #2# | Apple, look, partner | 3 |
| #1# | #3# | Apple, Apple | 2 |
| #1# | #4# | | 0 |
| #1# | #5# | Apple | 1 |
| #1# | #6# | | 0 |
| #2# | #3# | Apple, Apple, executive, company | 4 |
| #2# | #4# | | 0 |
| #2# | #5# | Apple | 1 |
| #2# | #6# | company | 1 |
| #3# | #4# | | 0 |
| #3# | #5# | Apple, Apple | 2 |
| #3# | #6# | company | 1 |
| #4# | #5# | | 0 |
| #4# | #6# | | 0 |
| #5# | #6# | | 0 |

Table 2: Measuring lexical cohesion in text unit pairs.

are taken into account, the final scores and consequent ranking order computable from Table 2 are:

- first: text unit #2# (final score: 7);
- second: text unit #3# (final score: 4), and
- third: text unit #1# (final score: 3).

A text abridgement can be obtained by selecting text units in ranking order according to the text percentage specified by the user. For example, a 35% abridgement of the text in Table 2 would result in the selection of text units #2# and #3#.

As Hoey points out, additional techniques can be used to refine the assessment of lexical cohesion. A typical example is the use of thesaurus functions such as synonymy and hyponymy to extend the notion of word sharing across text units, as exemplified in Hirst and St-Onge (1997) and Barzilay and Elhadad (1997) with reference to WordNet (Miller *et al.*, 1990). Such an extension may improve on the assessment of textual saliency and connectivity thus providing better generic summaries, as argued in Barzilay and Elhadad (1997).

There are basically two problems with the uses of lexical cohesion for summarization reviewed above. First, the basic algorithm requires that (*i*) all unique pairwise permutations of distinct text units be processed, and (*ii*) all cross-sentence word combinations be evaluated for each such text unit pair. The complexity of this algorithm will therefore be $O(n^2 * m^2)$ for $n$ text units in a text and $m$ words in a text unit of average length in the text at hand. This estimate may get worse as conditions such as synonymy and hyponymy are checked for each word pair to extend the notion of lexical cohesion, e.g. using WordNet as in Barzilay and Elhadad (1997). Consequently, the approach may not be suitable for on-line use with longer input texts. Secondly, the use of thesauri envisaged in both Hirst and St-Onge (1997) and Barzilay and Elhadad (1997) does not address the question of topical aptness. Thesaural relations such as synonymy and hyponymy are meant to capture word similarity in order to assess lexical cohesion among text units, and not to provide a thematic characterization of text units.[2] Consequently, it will not be possible to index and retrieve text units in term of topic aptness according to users' needs. In the remaining part of the paper, we will show how these concerns of efficiency and thematic characterization can be addressed with specific reference to a system performing generic and query-based indicative

[2] Notice incidentally that such thematic characterization could not be achieved using thesauri such as WordNet since since WordNet does not provide an arrangement of synonym sets into classes of discourse topics (e.g. finance, sport, health).

summaries.

## 3 An Efficient Method for Computing Lexical Cohesion

The method we are about to describe comprises three phases:

- a **preparatory phase** where the input text undergoes a number of normalizations so as to facilitate the process of assessing lexical cohesion;
- an **indexing phase** where the sharing of elements indicative of lexical cohesion is assessed for each text unit, and
- a **ranking phase** where the assessment of lexical cohesion carried out in the indexing phase is used to rank text units.

### 3.1 Preparatory Phase

During the preparatory phase, the text undergoes a number of normalizations which have the purpose of facilitating the process of computing lexical cohesion, including:

- removal of formatting commands
- text segmentation, i.e. breaking the input text into text units
- part-of-speech tagging
- recognition of proper names
- recognition of multi-word expressions
- removal of stop words
- word tokenization, e.g. lemmatization.

### 3.2 Indexing Phase

In providing a solution for the efficiency problem, our aim is to compute lexical cohesion for all text units in a text without having to process all cross-sentence word combinations for all unique and distinct pair-wise text unit permutations. To achieve this objective, we index each text unit with reference to each word occurring in it and reverse-index each such word with reference to all other text units in which the word occurs, as shown in Table 3 for text unit **#2#**. The sharing of words can then be measured by counting all occurrences of identical text units linked to the words associated with the "head" text unit (**#2#** in Table 3), as shown in Table 4. By repeating the two operations described above for each text unit in the text shown in Table 1, we will obtain a table of lexical cohesion links equivalent to that shown on Table 2.

$$\left\langle \texttt{\#2\#} \left\{ \begin{array}{l} \texttt{< Apple \{\#1\#,\#3\#,\#3\#,\#5\#\} >} \\ \texttt{< company \{\#3\#,\#6\#\} >} \\ \texttt{< executive \{\#3\#\} >} \\ \texttt{< look \{\#1\#\} >} \\ \texttt{< partner \{\#1\#\} >} \end{array} \right\} \right\rangle$$

Table 3: Text unit **#2#** and its words with pointers to the other text units in which they occur.

| | #1# | #3# | #5# | #6# |
|---|---|---|---|---|
| #2# | 3 | 4 | 1 | 1 |

Table 4: Total number of lexical cohesion links which text unit **#2#** has with all other text units

According to this method, we are still processing pair-wise permutations of text units to collect lexical cohesion links as shown in Table 4. However, there are two important differences with the original algorithm. First, non-cohesive text units are not taken into account (e.g. the pair **#2#-#4#** in the example under analysis); therefore, on average the number of text unit permutations will be significantly smaller than that processed in the original algorithm. With reference to the text in Table 1, for example, we would be processing 7 text unit permutations less which is over 41% of the number of text unit permutations which need computing according to the original algorithm, as shown in Table 2. Secondly, although pair-wise text unit combinations are still processed, we avoid doing so for all cross-sentence word permutations. Consequently, the complexity of the algorithm is $O(n^2 * m)$ for $n$ text units in a text and $m$ words in a text unit of average length in the text as compared to $O(n^2 * m^2)$ for the original algorithm.[3]

[3] A further improvement yet would be to avoid counting lexical cohesion links per text unit as in Table 4, and just sum all text unit occurrences associated with reversed-indexed words in structures such as those in Table 3, e.g. the lexical cohesion score for text unit **#2#** would simply be 9. This would remove the need of processing pair-wise text unit permutations for the assessment of lexical cohesion links, thus bringing the complexity down to $O(n * m)$. Such further step, however, would preempt the possibility of excluding lexical cohesion scores for text unit pairs which are below a given threshold.

```
• Let
    – TRSH be the lexical cohesion threshold
    – TU be the current text unit
    – $LC^{TU}$ be the current lexical cohesion score of $TU$ (i.e. $LC^{TU}$ is the count of tokenized words $TU$ shares with some other text unit)
    – CLevel be the level of the current lexical cohesion score calculated as the difference between $LC^{TU}$ and $TRSH$
    – Score be the lexical cohesion score previously assigned $TU$ (if any)
    – Level be the level for the lexical cohesion score previously assigned to $TU$ (if any)
•   – if $LC^{TU} = 0$, then do nothing
    – else, if the scoring structure $\langle Level, TU, Score \rangle$ exists, then
        * if $Level > CLevel$, then do nothing
        * else, if $Level = CLevel$, then the new scoring structure is $\langle Level, TU, Score + LC^{TU} \rangle$
        * else, if $CLevel > 0$, then
            · if $Level > 0$, then the new scoring structure is $\langle 1, TU, Score + LC^{TU} \rangle$
            · else, if $Level \leq 0$, then the new scoring structure is $\langle 1, TU, LC^{TU} \rangle$
        * else the new scoring structure is $\langle CLevel, TU, LC^{TU} \rangle$
    – else
        * if $CLevel > 0$, then create the scoring structure $\langle 1, TU, LC^{TU} \rangle$
        * else create the scoring structure $\langle CLevel, TU, LC^{TU} \rangle$
```

Table 5: Method for ranking text units according to lexical cohesion scores.

## 3.3 Ranking Phase

Each text unit is ranked with reference to the total number of lexical cohesion scores collected, such as those shown in Table 4. The objective of such a ranking process is to assess the import of each score and combine all scores into a rank for each text unit. In performing this assessment, provisions are made for a threshold which specifies the minimal number of links required for text units to be lexically cohesive, following Hoey's approach (see §1). The procedure outlined in Table 5 describes the scoring methodology adopted. Ranking a text unit according to this procedure involves adding the lexical cohesion scores associated with the text unit which are either

- above the threshold, or
- below the threshold and of the same magnitude.

If the threshold is 0, then there is a single level and the final score is the sum of all scores. Suppose for example, we are ranking text units #2# with reference to the scores in Table 4 with a lexical cohesion threshold of 2. In this case we apply the ranking procedure in Table 5 to each score in Table 4, as shown in Table 6. Following this procedure for all text units in Table 1, we will obtain the ranking in Table 7.

```
• Costant values
    – $TRSH = 2$
    – $TU = #2#$
• Scoring text unit #2#
    – Lexical cohesion with text unit #6#
        * $LC^{TU} = 1$
        * $CLevel = -1$ (i.e. $LC^{TU} - TRSH$)
        * no previous scoring structure
        * current scoring structure: $\langle -1, #2#, 1 \rangle$
    – Lexical cohesion with text unit #5#
        * $LC^{TU} = 1$
        * $CLevel = -1$
        * previous scoring structure: $\langle -1, #2#, 1 \rangle$
        * current scoring structure: $\langle -1, #2#, 2 \rangle$
    – Lexical cohesion with text unit #3#
        * $LC^{TU} = 4$
        * $CLevel = 2$
        * previous scoring structure: $\langle -1, #2#, 2 \rangle$
        * current scoring structure: $\langle 0, #2#, 4 \rangle$
    – Lexical cohesion with text unit #1#
        * $LC^{TU} = 3$
        * $CLevel = 1$
        * previous scoring structure: $\langle 1, #2#, 4 \rangle$
        * final scoring structure: $\langle 1, #2#, 7 \rangle$
```

Table 6: Ranking text unit #2# for lexical cohesion.

## 4 Assessing Topic Aptness

When used with a dictionary database providing information about the thematic domain of words (e.g. *business, politics, sport*), the same method can be slightly modified to compute lexical cohesion with reference to discourse topics rather than words. Such an application makes

| *Rank* | *Text unit* | *Level* | *Score* |
|---|---|---|---|
| 1st | #2# | 1 | 7 |
| 2nd | #3# | 1 | 4 |
| 3rd | #1# | 1 | 3 |
| 4th | #5# | 0 | 2 |
| 5th | #6# | −1 | 2 |
| 6th | #4# | — | 0 |

Table 7: Ranking for all text units in the text shown on Table 1.

| WORD_POS | CODE | EXPLANATION |
|---|---|---|
| company_n | F | Finance & Business |
| | MI | Military (the armed forces) |
| | SCG | Scouting & Girl Guides |
| | TH | Theatre |
| partner_n | DA | Dance & Choreography |
| | F | Finance & Business |
| | MGE | Marriage, Divorce, Relationships & Infidelity |
| | TG | Team Games |

Table 8: Fragment of dictionary database providing subject domain information.

```
              ⎧ < DA    {#2#-partner} >        ⎫
              ⎪ < F     {#2#-partner,          ⎪
⟨ #1#-partner ⎨          #3#-company,          ⎬ ⟩
              ⎪          #6#-company} >        ⎪
              ⎪ < MGE   {#2#-partner} >        ⎪
              ⎩ < TG    {#2#-partner} >        ⎭
```

Table 9: Text unit #1# and its subject domain codes with pointers to the other text units in which they occur.

| | #3# | #6# |
|---|---|---|
| #1#-partner | 1<br>F<br>company | 1<br>F<br>company |

Table 10: Total number of lexical cohesion links induced by subject domain codes for text unit #1#.

it possible to detect the major topics of a document automatically and to assess how well each text unit represents these topics.

In our implementation, we used the "subject domain codes" provided in the machine readable version of CIDE (*Cambridge International Dictionary of English* (Procter, 1995)). Table 8 provides an illustrative example of the information used. Both the indexing and ranking phases are carried out with reference to subject domain codes rather than words.

As shown in Table 9 for text unit #1#, the indexing procedure provides a record of the subject domain codes occurring in each text unit; each such subject code is reverse-indexed with reference to all other text units in which the subject code occurs. In addition, a record of which word originates which cohesion link is kept for each text unit index. The main function of keeping track of this information is to avoid counting lexical cohesion links generated by overlapping domain codes which relate to the same word — for words associated with more than one code. Such provision is required in order to avoid, or at least reduce the chances of, counting codes which are out of context, that is codes which relate to senses of the word other than the intended sense. For example, the word *partner* occurring in the first two text units of the text in Table 1 is associated with four different subject codes pertaining to the domains of Dance (DA), Finance (F), Marriage (M) and Team Games (TG), as shown in Table 8. However, only the Finance reading is appropriate in the given context. If we count the cohesion links generated by *partner* we would therefore count three incorrect cohesion links. By excluding all four cohesion links, the inclusion of contextually inappropriate cohesion links is avoided. Needless to say, we will also throw away the correct cohesion link (F in this case). However, this loss can be redressed if we also compute lexical cohesion links generated from shared words across text units as discussed in §2, and combine the results with the lexical cohesion ranks obtained with subject domain codes.

The lexical cohesion links for text unit #1# will therefore be scored as shown in Table 10, where associations between link scores and relevant codes as well as the words generating them are maintained. As can be observed, only the appropriate code expansion F (Finance) for the words *partner* and *company* is taken into account. This is simply because F is the only code shared by the two words (see Table 8).

As mentioned earlier, lexical cohesion links induced by subject domain scores can be used to rank text units using the procedure shown in Table 5. Other uses include providing a topic profile of the text and an indication of how well each text unit represents a given topic. For example, the code BZ (Business & Commerce) is associated with the words:

|  | #2 | #3# | #4# | #5# |
|---|---|---|---|---|
| #2#-executive |  | 1<br>BZ<br>business | 1<br>BZ<br>market | 1<br>BZ<br>interest |
| #3#-executive |  |  | 1<br>BZ<br>market | 1<br>BZ<br>interest |
| #3#-business | 1<br>BZ<br>execut. |  | 1<br>BZ<br>market | 1<br>BZ<br>interest |
| #4#-market | 1<br>BZ<br>execut. | 2<br>BZ<br>execut.<br>business |  | 1<br>BZ<br>interest |
| #5#-interest | 1<br>BZ<br>execut. | 2<br>BZ<br>execut.<br>business | 1<br>BZ<br>market |  |

Table 11: Lexical cohesion links relating to code BZ.

| CODES | TEXT UNIT PAIRS |
|---|---|
| BZ | 2-3 2-4 2-5 3-4 3-5 3-2 3-4 3-5<br>4-2 4-3 4-3 4-5 5-2 5-3 5-3 5-4 |
| F | 1-2 1-3 1-6 2-1 2-3 2-6 3-1 3-2 6-1 6-2 |
| FA | 2-5 5-2 |
| IV | 4-5 5-4 |
| CN | 3-4 4-3 |

Table 12: Subject domain codes and the text units pairs they relate.

- *executive* occurring once in text units #2# and #3#;
- *business* occurring once in text unit #3#;
- *market* occurring once in text unit #4#, and
- *interest* occurring once in text unit #5#.

After calculating the lexical cohesion links for all text units following the method illustrated in Tables 9-10 for text unit #1#, the links scored for the code BZ will be as shown in Table 11. By repeating this operation for all codes for which there are lexical cohesion scores — F, FA, IV and CN for the text under analysis — we could then count all text unit pairs which each code relates, as shown in Table 12. The relations between subject domain codes and text unit pairs in Table 12 can subsequently be turned into percentage ratios to provide a topic/theme profile of the text as shown in Table 13.

By keeping track of the links among text units, relevant codes and their originating words, it is also possible to retrieve text units on the basis of specific subject domain codes or specific words. When retrieving on specific words, there is also the option of expanding the word into subject domain codes and using these to retrieve text units. The retrieved text units can then be ordered according to the ranking order previously computed.

|  |  |  |
|---|---|---|
| 50% | BZ | Business & Commerce |
| 31.25% | F | Finance & Business |
| 6.25% | IV | Investment & Stock Markets |
| 6.25% | FA | Overseas Politics &<br>International Relations |
| 6.25% | CN | Communications |

Table 13: Topic profile of document in Table 1, according to the distribution of subject domain codes across text units shown in Table 12.

## 5 Applications, Extensions and Evaluation

An implementation of this approach to lexical cohesion has been used as the driving engine of a summarization system developed at SHARP Laboratories of Europe. The system is designed to handle requests for both generic and query-based indicative summaries. The level-based differentiation of text units obtained through the ranking procedure discussed in §3.3, is used to select the most salient and better connected portion of text units in a text corresponding to the summary ratio requested by the user. In addition, the user can display a topic profile of the input text, as shown in Table 13 and choose whichever code(s) s/he is interested in, specify a summary ratio and retrieve the wanted portion of the text which best represents the topic(s) selected. Query-based summaries can also be issued by entering keywords; in this case there is the option of expanding key-words into codes and use these to issue a summary query.

The method described can also be used to develop a conceptal indexing component for information retrieval, following Dobrov *et al.* (1997). Because an attempt is made to prune contextually inappropriate sense expansions of words, the present method may help reducing the ambiguity problem.

Possible improvements of this approach can be implemented taking into account additional ways of assessing lexical cohesion such as:

- the presence of synonyms or hyponyms across text units (Hoey, 1991; Hirst and St-Onge, 1997; Barzilay and Elhadad 1997);

- the presence of lexical cohesion established with reference to lexical databases offering a semantic classification of words other than synonyms, hyponyms and subject domain codes;
- the presence of near-synonymous words across text units established by using a method for estimating the degree of semantic similarity between word pairs such as the one proposed by Resnik (1995);
- the presence of anaphoric links across text units (Hoey, 1991; Boguraev & Kennedy, 1997), and
- the presence of formatting commands as indicators of the relevance of particular types of text fragments.

To evaluate the utility of the approach to lexical cohesion developed for summarization, a testsuite was created using 41 Reuter's news stories and related summaries (available at `http://www.yahoo.com/headlines/news/`), by annotating each story with best summary lines. In one evaluation experiment, summary ratio was set at 20% and generic summaries were obtained for the 41 texts. On average, 60% of each summary contained best summary lines. The ranking method used in this evaluation was based on combined lexical cohesion scores based on lemmas and their associated subject domain codes given in CIDE. Summary results obtained with the Autosummarize facility in Microsoft Word 97 were used as baseline for comparison. On average, only 30% of each summary in Word 97 contained best summary lines. In future work, we hope to corroborate these results and to extend their validity with reference to query-based indicative summaries using the evaluation framework set within the context of SUMMAC (*Automatic Text Summarization Conference*, see `http://www.tipster.org/`).

## References

Barzilay, R. and M. Elhadad (1997) Using Lexical Chains for Text Summarization. In I. Mani and M. Maybury (eds) *Intelligent Scalable Text Summarization, Proceedings of a Workshop Sponsored by the Association for Computational Linguistics*, Madrid, Spain.

Boguraev, B. & C. Kennedy (1997) Salience-based Content Characterization of Text Documents. In I. Mani and M. Maybury (eds) *Intelligent Scalable Text Summarization, Prooceedings of a Workshop Sponsored by the Association for Computational Linguistics*, Madrid, Spain.

Dobrov, B., N. Loukachevitch and T. Yudina (1997) Conceptual Indexing Using Thematic Representation of Texts. In *The 6th Text Retrieval Conference (TREC-6).*

Fox, C. (1992) Lexical Analysis and Stoplists. In Frakes W and Baeza-Yates R (eds) Information Retrieval: Data Structures & Algorithms. Prentice Hall, Upper Saddle River, NJ, USA, pp. 102-130.

Hirst, G. and D. St-Onge (1997) Lexical Chains as Representation of context for the detection and correction of malapropism. In C. Fellbaum (ed) *WordNet: An electronic lexical database and some of its applications.* MIT Press, Cambridge, Mass.

Hoey, M. (1991) Patterns of Lexis in Text. OUP, Oford, UK.

Miller, G., Beckwith, R., C. Fellbaum, D. Gross and K. Miller (1990) Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235-312.

Procter, P. (1995) *Cambridge International Dictionary of English*, CUP, London.

Philip Resnik (1995) Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI-95.*