DOCUMENT RESUME

ED 378 808 FL 022 688

AUTHOR Collier, Alex
TITLE A System for Automating Concordance Line Selection.
PUB DATE Sep 94
NOTE 7p.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Abstracts; *Access to Information; *Automatic Indexing; Cohesion (Written Composition); Foreign Countries; *Indexes; *Information Retrieval; Information Utilization; Lexicography; *Reference Materials; *Subject Index Terms

ABSTRACT

This paper argues that as the number of concordance lines presented to researchers increases in line with ever-growing corpus size, the automatic selection of the most central lines will become more important. A method will be presented for selecting the most representative members from a set of concordance lines on the basis of repeated lexical features. This follows from previous work on lexical cohesion that was utilized in a system for automatic abridgement generation. Such a system would grant full accessibility to corpus material while avoiding the presentation of so much data to researchers that they become overloaded. (Contains 7 references.) (Author/CK)

A System for Automating Concordance Line Selection

Alex COLLIER

Department of English
University of Liverpool
Liverpool
UK
acollier@liv.ac.uk

Abstract

In this paper it will be argued that as the number of concordance lines presented to researchers increases in line with ever-growing corpus size, the automatic selection of the most central lines will become more important. A method will be presented of selecting the most representative members from a set of concordance lines on the basis of repeated lexical features. This follows on from work on lexical cohesion done by Hoey which was utilised in a system for automatic abridgement generation developed by the author and Hoey. Such a system would grant full accessibility to corpus material while avoiding the presentation of so much data to researchers that they become overloaded.

Keywords:
Corpus-based NLP – Hybrid Approaches

Manual Abstra ing

For several years now, Michael Hoey has been investigating the phenomenon of repeated sequences of lexical items in text and the relationship of these to the core sentences of a text. The phenomenon on which he focussed is known as 'lexical cohesion', which is one of the forms of textual cohesion described in Halliday and Hasan (1976). Hoey noticed that the recurrence of certain items across the sentences of a text could be used as an aid to identifying the 'key' sentences of that text, where a 'key' sentence can be characterised by its centrality to the text and the expressive power which it has in standing in for several other sentences.

In Hoey (1991) the 'connection by repetition' is termed a *link*, which defines a subset of the relations which Halliday and Hasan termed *ties*. Hoey describes several different kinds of repetition, which range in complexity but all contribute to the links present between the sentences of the text. These repetition types can be categorised as follows:

Simple (Lexical) Repetition
: The link exists between two instances of the same word ('dream' ↔ 'dream').

Complex Lexical Repetition
: Here a link is made between members of a lemma, eg ('dream' ↔ 'dreaming'), which also includes morphologically related antonyms of the 'happy ↔ unhappy' variety.

Mutual Substitutability
: This instantiates a link between synonyms or expressions which are coreferential within the text ('clever' ↔ 'intelligent', 'John Major' ↔ 'the Prime Minister').

Complex Paraphrase
: This covers some cases of repetition by antonym not covered by complex lexical repetition, such as 'light ↔ dark'.

Hyponymy
: Here, superordinates are included as repetitions, eg 'tulip ↔ flower'.

Pronominal Repetition
: In this case, reference is made to a lexical item by means of a pronoun and a link is therefore made between the pronoun and the lexical item, eg 'dream' ↔ 'it'.

Substitution
: Where a word such as 'this' or 'the above' is used to refer back to a previous clause, sentence or even larger section of the text, this must also be considered as a link.

On the basis of this analysis Hoey was able to calculate the frequency of recurrent lexical items for

each sentence. This information was then used to build up a measure of connectivity between the sentences of the text. This information can be represented in a matrix as seen below, where the number down the left-hand side and the diagonal indicate the sentence number and the other digits show the number of links between the sentences.

```
1
2 0 2
3 2 0 3
4 1 0 0 4
```

Thus the matrix shows that there are two links between sentences (1) and (3) and just one link between sentences (1) and (4).

Whilst looking at non-narrative texts, Hoey had noticed that those sentences in a text which had the greatest degree of cohesion could be used to build an abridgement of the text. In order to ascertain the level of cohesion, Hoey introduced an additional measure which he called the *bond*, defining it as 'a connection which exists between a pair of sentences by virtue of there being an above-average number of links relating them'. In Hoey's experience the sentences having the greatest bond count tended to cohere together, thus making the abridgement itself seem more like a 'text' with internal structure of its own.

The bond strength of a sentence was arrived at by establishing a link threshold, which Hoey set at a minimum of three, whereby any sentence with at least that number of links with another sentence could be said to have one bond with that sentence. Thus in the matrix above, if we were to set a threshold of two links then sentences (1) and (3) would share one bond, whereas sentence (4) disappears from the equation. This is an important step, since links are always symmetrical; bonds, however, describe a wider relationship between the sentences of a text and enable two sentences which share the same lexical items to obtain differing cohesion scores, thanks to the fact that their links with *other* sentences can be taken into consideration. This can be seen by applying a threshold of just one link: this assigns a score of two bonds to sentence (1) (two links with (3) + one link with (4)), but only one bond to sentences (3) and (4), both of which link with (1). Sentence (2), since it has no links at all, fails to register any bonds.

In the course of his research Hoey analysed many texts by this method and in the majority of cases coherent abridgements were created. The link-bond ratio exemplified above is, as stated earlier, far lower than that employed by Hoey and is set at this level for illustrative purposes only.

Automating the Abridgement

In a project at the University of Birmingham carried out in collaboration with British Telecom (see Collier et al (1991)) the approach outlined above was developed into a working software suite for the abridgement of electronic text. Within the limits of the project it was only possible to incorporate the two simplest forms of repetition, simple and complex lexical repetition, into the software, as well as offering the user the opportunity of opening a 'window' of sentences onto the text whenever a pronoun was encountered in the abridgement; insufficient time was available to us to incorporate the real-world knowledge which would link 'John Major' and 'Prime Minister' and the inclusion (manually) of the more simple thesaural links brought only a minor improvement in the results. What the software lacked in sophistication it may be said to make up for in speed. In a matter of seconds the software created a matrix like the one shown above for a complete newspaper article of thirty to forty sentences; Hoey's manual approach required many hours of painstaking and error-prone examination of the text.

Once the matrix has been generated it can be manipulated by selecting a link threshold and calculating the number of bonds for each sentence. Creating an abridgement is then simply a case of setting a *bond* threshold which results in the required number of sentences. Thanks to the fully-automated nature of the process, adjusting the link or bond thresholds is a straightforward task, enabling different combinations of parameters to be tried until the optimal result is obtained.

Despite the cruder nature of this method, satisfactory results were produced for a wide range of texts, in that the key sentences identified by the system still tended to reflect a large proportion of the rest of the text.

The Need for Automatic Selection

In the past decade, the size of corpora has increased significantly. Until relatively recently a typical corpus (based on the LOB and Brown type corpus) amounted to around a million tokens. The task of generating concordances might well have been performed non-interactively, possibly

overnight, resulting in a printout representing all the contexts of the type selected, sorted according to the desired criteria. For all but the most frequent types this printout would amount to no more than a few pages, which researchers could easily spread out for inspection and thus gain an overview of the grammatical and collocational behaviour of the word under scrutiny.

There are two main reasons why this approach is not feasible for today's corpora.

Firstly, the size of the corpus, which is now measured in tens or hundreds of millions of tokens, has greatly increased the frequency of each type.

Secondly, the working methods of the researchers have evolved to take advantage of the technological benefits now available, such as instant concordance generation and immediate re-sorting of the concordance lines based on any position relative to the node word. It is now expected that these operations take place online at the computer screen.

These factors combine to reduce the overview of a type's behaviour available to the corpus researcher: the amount of data to be presented increases constantly, whereas, paradoxically it may seem, the area in which it can be displayed has shrunk to the size of a computer screen. As corpus size constantly increases the first dimension will continue to increase; by comparison the second dimension, unless we are soon to witness astounding advances in VDU technology, will continue to be a limiting factor to the accessibility of corpus data. This is the point at which the work done on textual abridgement becomes of relevance to the problem.

Applying Lexical Cohesion

As the number of concordance lines continues to increase, some method is required which restores the overview of the patterns present. Otherwise, the full benefit will never be reaped from the continued increase in corpus size. The development of a system which automatically selects concordance lines which carry information redolent of the information carried by other lines would be very valuable in combating the problems associated with large corpora. Such a system would in effect extend the notion of centrality from sentences in a text to individual concordance lines within a set of lines. As well as changing in extent, centrality would no longer be defined in terms of propositional or semantic ties but rather in terms of the linguistic ties which exist between members of a concordance set.

The way in which the abridgement methodology can be applied to this problem lies in treating the set of concordance lines for a type as a *text* in which each line is equivalent to a sentence. This is a novel approach for several reasons:

- In Hoey's approach grammatical items were largely excluded from the analysis; here it is essential to include these, since they play a highly significant role in the definition of recurrent patterns.
- We are not dealing with 'normal' text, but rather with a composite set of contexts whose only connection, in terms of the way that they are extracted from the corpus, is that they all exemplify the same word.
- A corollary of this is that we cannot expect concordance lines to be cohesive together in the way that the sentences of a text are, yet some relationship clearly does exist between the members of a concordance set. It is this relationship, and the way in which it impinges on the centrality of particular lines, which we wish to define.

From an information retrieval perspective, we could talk of a central line as being one which is informationally rich and useful in providing an 'abridgement' of other lines with similar but possibly less 'concentrated' features. Put in lexicographical terms, such lines would be candidates for use as examples in a dictionary definition, since they would display some important aspect of the headword that the lexicographer wished to convey to the dictionary user.

The approach which has been taken therefore is to process the set of concordances and build a connectivity matrix by means of the abridgement software. A bond score is then assigned to each concordance line, the hypothesis being that the more central members of the concordance set will score more highly.

There are a number of procedural differences between this approach and that used for creating textual abridgements:

- In the current configuration of the software no lemmatisation takes place. Since there are already a large number of variables in place it was felt that lemmatisation would add further complexity which might hide important information about the links between individual concordance lines.

Naturally, this does not preclude its inclusion at a latter stage, once the selection process is better understood.

- Since the underlying system was originally designed to look at *lexical* cohesion, a certain number of grammatical words are automatically disregarded by the system, but the number and nature of these stopwords is user-definable when the software is run. One interesting contrast is obtained by comparing the output when first a functionally-defined list (eg all pronouns, articles and prepositions) and then a frequency-based list (the most frequent 100 words from the corpus) are supplied as stopwords. Of course, this flexibility introduces yet another parameter to the system, which at present offers over a hundred different configurations based on the stopword list, the size of the span and the positional notation employed.

The various parameters involved in the process are described below, but first it may be useful to introduce an example. Let us take a concordance set for the node 'kin' (examples courtesy of (BCET)):

```
(1)  barrier excuses. As for the "kith and kin" appeal, to quote the Reverend Arth
(2) id. "And I'm sure the cattle's next of kin have been informed but is that quit
(3) ted backing of France for her kith and kin in Algeria and for her Army protec
(4) earted commitment towards our kith and kin overseas." Identity of "race, langu
(5) f they do lecture their white kith and kin rather than the guerrillas, it is p
(6) e a narrow view of who is our kith and kin. Religion very properly tends to em
(7)  been seeing to that. The only next of kin seems to be a cousin in Droitwich.
(8) l kindness toward him, they're not his kin.... That's exactly the feeling. Old
(9) and her property passes to her next of kin under the  intestacy rules. That me
```

If this is provided as input to the system with the links-per-bond threshold set to 1 the following is obtained:

```
4 (1)  barrier excuses. As for the "kith and kin" appeal, to quote the Reverend Arth
4 (3) ted backing of France for her kith and kin in Algeria and for her Army protec
4 (4) earted commitment towards our kith and kin overseas." Identity of "race, langu
4 (5) f they do lecture their white kith and kin rather than the guerrillas, it is p
4 (6) e a narrow view of who is our kith and kin. Religion very properly tends to em
2 (2) id. "And I'm sure the cattle's next of kin have been informed but is that quit
2 (7)  been seeing to that. The only next of kin seems to be a cousin in Droitwich.
2 (9) and her property passes to her next of kin under the intestacy rules. That me
```

Here the first number on the line indicates the number of bonds that have been identified, whereby any line which fails to bond is omitted. The second number is simply the line number as shown in the original lines above. From this simple example the functionality of the system becomes apparent:

1. The somewhat untypical line (8) has been dropped because it obtained a bond score of zero and any line failing to form any bonds is excluded from the output.
2. The most common pattern of 'kith and kin' has been raised to prominence.
3. The slightly less common pattern 'next of kin' follows closely behind.

As noted earlier, there is one major difference in the central lines in the set of concordances and the central sentences in a conventional text: the central sentences selected by the textual abridgement system are expected to be coherent, ie to make sense together; this does not apply in the concordance line scenario, where the expectation is rather that they will have particular linguistic features in common. Naturally, each line in a set of concordances for a given type will contain at least one occurrence of the node word and, not surprisingly, this is actually discarded by the system, since it would introduce one extra link in every line with all other lines. This leads us to the question of what is actually detected as a 'feature'.

Lexical Cohesion meets Collocation

A number of statistical measures have been developed or adapted to rank the collocates of a node word in order of significance (Z-score, MI-score and T-score for example) and systems now exist

which apply these measures to a set of concordances to produce an overview of the most significant collocates on a position by position basis for each slot relative to the node (one to the left, one to the right, two to the left, two to the right .), enabling a picture of the typical environment of the node *in terms of a particular position.* While there is nothing inherently wrong with these measures, the method by which they are applied to the data tends to tell us little about the more general patterns which occur in concordance lines because they measure the significance of finding a particular word in a particular position. Some attempts have been made to tie the most significant collocates back to the line from whence they came, but this tends to become statistically and computationally complex. The method put forward here avoids statistical measures completely and relies upon the raw frequency of occurrence of the recurrent features.

A recurrent feature can be defined in a number of ways depending on the parameters passed to the software.

Simple Lexical Repetition

This is exemplified in the above example of the 'kin' concordances, where the collocate 'kith' is identified as occurring somewhere in the context. No attempt is made to pinpoint the location of the collocate relative to the node word.

Position Relative to the Node

A slightly more strict definition of the position of the collocate, in which it is merely specified whether it occurs before or after the node. This would detect features such as phrasal verbs, where the constituent verb and particle do not always maintain a fixed positional relationship.

Absolute Position

Here the exact location of the collocate is recorded. In order to count as a link between two lines it must therefore occur in precisely the same location relative to the node in the two lines. By opting for this positional notation and setting a higher link threshold it becomes possible to identify, for example, lines which contain instances of fixed phrases.

In Conclusion

The work described above has grown out of the author's interest in collocational patterns and the concordancing of large corpora. The combination of this interest with work undertaken on the creation of automatic abridgements has led to this attempt to investigate the link between abridgements of real texts and the selection of 'central' or 'representative' concordance lines. Many different sets of concordances have been processed and the results are, if nothing else, always indicative of the recurrent patterns in the environment of the node word.

To some extent, the correspondence between a line's bond score and its 'real' centrality remains a matter of interpretation, for few humans agree on what is a 'good' concordance line. On the whole, however, the results accord with linguistic intuitions and blind tests on a groups of Birmingham lexicographers demonstrated a high degree of overlap between those lines which scored highly and those considered to be good candidates for inclusion in a dictionary as illustrative citations. It has proved difficult to validate the results at anything beyond this purely intuitive evaluation, since there are no accepted and reliable automatic means of carrying out the analysis described here. It can only be added that output from software tools describeds earlier, which are designed to point the lexicographer towards the frequently recurring features of a set of concordances, serves to reinforce the centrality of the lines flagged as 'central' by this system.

References

BCET "The Birmingham Collection of English Texts", a corpus of around twenty million tokens of fiction, non-fiction and journalism built at the University of Birmingham in th. Mid-1980's. See Renouf (1987).

Bibliography

Collier, A.J., Hoey M.P. & Renouf, A.J. 1991. *British Telecom Project in Automatic Text Abstraction: Final Report*

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

Hoey, M.P. 1991. *Patterns of Lexis in Text.* London: OUP.

Johansson, S. 1980. 'The LOB Corpus of British English Texts: Presentation and Comments', *ALLC Journal.*

Kucera, H. and Francis, W. Nelson 1967. *Computational Analysis of Present Day American*

English, Brown University Press, Providence, Rhode Island.

Renouf, A.J. 1987. 'Corpus Development' in *Looking Up*, Ed. J M Sinclair, Harper-Collins, London.