Text Segmentation Using Reiteration and Collocation

Amanda C. Jobbins
Department of Computing
Nottingham Trent University
Nottingham NG1 4BU, UK
ajobbins@resumix.com

Lindsay J. Evett
Department of Computing
Nottingham Trent University
Nottingham NG1 4BU, UK
lje@doc.ntu.ac.uk

Abstract

A method is presented for segmenting text into subtopic areas. The proportion of related pairwise words is calculated between adjacent windows of text to determine their lexical similarity. The lexical cohesion relations of reiteration and collocation are used to identify related words. These relations are automatically located using a combination of three linguistic features: word repetition, collocation and relation weights. This method is shown to successfully detect known subject changes in text and corresponds well to the segmentations placed by test subjects.

Introduction

Many examples of heterogeneous data can be found in daily life. The Wall Street Journal archives, for example, consist of a series of articles about different subject areas. Segmenting such data into distinct topics is useful for information retrieval, where only those segments relevant to a user's query can be retrieved. Text segmentation could also be used as a pre-processing step in automatic summarisation. Each segment could be summarised individually and then combined to provide an abstract for a document.

Previous work on text segmentation has used term matching to identify clusters of related text. Salton and Buckley (1992) and later, Hearst (1994) extracted related text portions by matching high frequency terms. Yaari (1997) segmented text into a hierarchical structure, identifying sub-segments of larger segments. Ponte and Croft (1997) used word co-occurrences to expand the number of terms for matching. Reynar (1994) compared all words across a text rather than the more usual nearest neighbours. A problem with using word repetition is that inappropriate matches can be made because of the lack of contextual information (Salton et al., 1994). Another approach to text segmentation is the detection of semantically related words.

Hearst (1993) incorporated semantic information derived from WordNet but in later work reported that this information actually degraded word repetition results (Hearst, 1994). Related words have been located using spreading activation on a semantic network (Kozima, 1993), although only one text was segmented. Another approach extracted semantic information from Roget's Thesaurus (RT). Lexical cohesion relations (Halliday and Hasan, 1976) between words were identified in RT and used to construct lexical chains of related words in five texts (Morris and Hirst, 1991). It was reported that the lexical chains closely correlated to the intentional structure (Grosz and Sidner, 1986) of the texts, where the start and end of chains coincided with the intention ranges. However, RT does not capture all types of lexical cohesion relations. In previous work, it was found that collocation (a lexical cohesion relation) was under-represented in the thesaurus. Furthermore, this process was not automated and relied on subjective decision making.

Following Morris and Hirst's work, a segmentation algorithm was developed based on identifying lexical cohesion relations across a text. The proposed algorithm is fully automated, and a quantitative measure of the association between words is calculated. This algorithm utilises linguistic features additional to those captured in the thesaurus to identify the other types of lexical cohesion relations that can exist in text.

1 Background Theory: Lexical Cohesion

Cohesion concerns how words in a text are related. The major work on cohesion in English was conducted by Halliday and Hasan (1976). An instance of cohesion between a pair of elements is referred to as a *tie*. Ties can be anaphoric or cataphoric, and located at both the sentential and supra-sentential level. Halliday and Hasan classified cohesion under two types: grammatical and lexical. *Grammatical cohesion* is expressed through the grammatical relations in text such as ellipsis and conjunction. *Lexical cohesion* is expressed through the vocabulary used in text and the semantic relations between those words. Identifying semantic relations in a text can be a useful indicator of its conceptual structure.

Lexical cohesion is divided into three classes: general noun, reiteration and collocation. General noun's cohesive function is both grammatical and lexical, although Halliday and Hasan's analysis showed that this class plays a minor cohesive role. Consequently, it was not further considered. Reiteration is subdivided into four cohesive effects: word repetition (e.g. *ascent* and *ascent*), synonym (e.g. *ascent* and *climb*) which includes near-synonym and hyponym, superordinate (e.g. *ascent* and *task*) and general word (e.g. *ascent* and *thing*). The effect of general word is difficult to automatically identify because no common referent exists between the general word and the word to which it refers. A *collocation* is a predisposed combination of words, typically pairwise words, that tend to regularly co-occur (e.g. *orange* and *peel*). All semantic relations not classified under the class of reiteration are attributed to the class of collocation.

2 Identifying Lexical Cohesion

To automatically detect lexical cohesion ties between pairwise words, three linguistic features were considered: word repetition, collocation and relation weights. The first two methods represent lexical cohesion relations. Word repetition is a component of the lexical cohesion class of reiteration, and collocation is a lexical cohesion class in its entirety. The remaining types of lexical cohesion considered, include synonym and superordinate (the cohesive effect of general word was not included). These types can be identified using relation weights (Jobbins and Evett, 1998).

Word repetition: Word repetition ties in lexical cohesion are identified by same word matches and matches on inflections derived from the same stem. An inflected word was reduced to its stem by look-up in a lexicon (Keenan and Evett, 1989) comprising inflection and stem word pair records (e.g. "orange oranges").

Collocation: Collocations were extracted from a seven million word sample of the Longman English Language Corpus using the association ratio (Church and Hanks, 1990) and outputted to a lexicon. Collocations were automatically located in a text by looking up pairwise words in this lexicon. Figure 1 shows the record for the headword *orange* followed by its collocates. For example, the pairwise words *orange* and *peel* form a collocation.

```
orange free green lemon peel red
state yellow
```

Figure 1. Excerpt from the collocation lexicon.

Relation Weights: Relation weights quantify the amount of semantic relation between words based on the lexical organisation of RT (Jobbins and Evett, 1995). A thesaurus is a collection of synonym groups, indicating that synonym relations are captured, and the hierarchical structure of RT implies that superordinate relations are also captured. An alphabetically-ordered index of RT was generated, referred to as the Thesaurus Lexicon (TLex). Relation weights for pairwise words are calculated based on the satisfaction of one or more of four possible connections in TLex.

3 Proposed Segmentation Algorithm

The proposed segmentation algorithm compares adjacent windows of sentences and determines their lexical similarity. A window size of three sentences was found to produce the best results. Multiple sentences were compared because calculating lexical similarity between words is too fine (Rotondo, 1984) and between individual sentences is unreliable (Salton and Buckley, 1991).

Lexical similarity is calculated for each window comparison based on the proportion of related words, and is given as a normalised score. Word repetitions are identified between identical words and words derived from the same stem. Collocations are located by looking up word pairs in the collocation lexicon. Relation weights are calculated between pairwise words according to their location in RT. The lexical similarity score indicates the amount of lexical cohesion demonstrated by two windows. Scores plotted on a graph show a series of peaks (high scores) and troughs (low scores). Low scores indicate a weak level of cohesion. Hence, a trough signals a potential subject change and texts can be segmented at these points.

4 Experiment 1: Locating Subject Change

An investigation was conducted to determine whether the segmentation algorithm could reliably locate subject change in text.

Method: Seven topical articles of between 250 to 450 words in length were extracted from the World Wide Web. A total of 42 texts for test data were generated by concatenating pairs of these articles. Hence, each generated text consisted of two articles. The transition from the first article to the second represented a known subject change point. Previous work has identified the breaks between concatenated texts to evaluate the performance of text segmentation algorithms (Reynar, 1994; Stairmand, 1997). For each text, the troughs placed by the segmentation algorithm were compared to the location of the known subject change point in that text. An error margin of one sentence either side of this point, determined by empirical analysis, was allowed.

Results: Table 1 gives the results for the comparison of the troughs placed by the segmentation algorithm to the known subject change points.

linguistic feature	troughs placed		subject change points located (out of 42 poss.)
	average	std. dev.	
word repetition collocation	7.1	3.16	41 (97.6%)
word repetition relation weights	7.3	5.22	41 (97.6%)
word repetition	8.5	3.62	41 (97.6%)
collocation relation weights	5.8	3.70	40 (95.2%)
word repetition collocation relation weights	6.4	4.72	40 (95.2%)
relation weights	7	4.23	39 (92.9%)
collocation	6.3	3.83	35 (83.3%)

Table 1. Comparison of segmentation algorithm using different linguistic features.

Discussion: The segmentation algorithm using the linguistic features word repetition and collocation in combination achieved the best result. A total of 41 out of a possible 42 known subject change points were identified from the least number of troughs placed per text (7.1). For the text where the known subject change point went undetected, a total of three troughs were placed at sentences 6, 11 and 18. The subject change point occurred at sentence 13, just two sentences after a predicted subject change at sentence 11.

In this investigation, word repetition alone achieved better results than using either collocation or relation weights individually. The combination of word repetition with another linguistic feature improved on its individual result, where less troughs were placed per text.

5 Experiment 2: Test Subject Evaluation

The objective of the current investigation was to determine whether all troughs coincide with a subject change. The troughs placed by the algorithm were compared to the segmentations identified by test subjects for the same texts.

Method: Twenty texts were randomly selected for test data each consisting of approximately 500 words. These texts were presented to seven test subjects who were instructed to identify the sentences at which a new subject area commenced. No restriction was placed on the number of subject changes that could be identified. Segmentation points, indicating a change of subject, were determined by the agreement of three or more test subjects (Litman and Passonneau, 1996). Adjacent segmentation points were treated as one point because it is likely that they refer to the same subject change.

The troughs placed by the segmentation algorithm were compared to the segmentation points identified by the test subjects. In Experiment 1, the top five approaches investigated identified at least 40 out of 42 known subject change points. Due to that success, these five approaches were applied in this experiment. To evaluate the results, the information retrieval metrics precision and recall were used. These metrics have tended to be adopted for the assessment of text segmentation algorithms, but they do not provide a scale of correctness (Beeferman et al., 1997). The degree to which a segmentation point was 'missed' by a trough, for instance, is not considered. Allowing an error margin provides some degree of flexibility. An error margin of two sentences either side of a segmentation point was used by Hearst (1993) and Reynar (1994) allowed three sentences. In this investigation, an error margin of two sentences was considered.

Results: Table 2 gives the mean values for the comparison of troughs placed by the segmentation algorithm to the segmentation points identified by the test subjects for all the texts.

Discussion: The segmentation algorithm using word repetition and relation weights in combination achieved mean precision and recall rates of 0.80 and 0.69, respectively. For 9 out of the 20 texts segmented, all troughs were relevant. Therefore, many of the troughs placed by the segmentation algorithm represented valid subject changes. Both word repetition in combination with collocation and all three features in combination also achieved a precision rate of 0.80 but attained a lower recall rate of 0.62. These results demonstrate that supplementing word repetition with other linguistic features can improve text segmentation. As an example, a text segmentation algorithm developed by Hearst (1994) based on word repetition alone attained inferior precision and recall rates of 0.66 and 0.61.

linguistic feature	mean values for all texts				
	relevant	relevant found	nonrel. found	prec.	rec.
word repetition relation weights	4.50	3.10	1.00	0.80	0.69
word repetition collocation	4.50	2.80	0.85	0.80	0.62
word repetition collocation relation weights	4.50	2.80	0.85	0.80	0.62
collocation relation weights	4.50	2.75	0.90	0.80	0.60
word repetition	4.50	2.50	0.95	0.78	0.56

Table 2. Comparison of troughs to segmentation points placed by the test subjects.

In this investigation, recall rates tended to be lower than precision rates because the algorithm identified fewer segments (4.1 per text) than the test subjects (4.5). Each text was only 500 words in length and was related to a specific subject area. These factors limited the degree of subject change that occurred. Consequently, the test subjects tended to identify subject changes that were more subtle than the algorithm could detect.

Conclusion

The text segmentation algorithm developed used three linguistic features to automatically detect lexical cohesion relations across windows. The combination of features word repetition and relation weights produced the best precision and recall rates of 0.80 and 0.69. When used in isolation, the performance of each feature was inferior to a combined approach. This fact provides evidence that different lexical relations are detected by each linguistic feature considered.

Areas for improving the segmentation algorithm include incorporation of a threshold for troughs. Currently, all troughs indicate a subject change, however, minor fluctuations in scores may be discounted. Future work with this algorithm should include application to longer documents. With trough thresholding the segments identified in longer documents could detect significant subject changes. Having located the related segments in text, a method of determining the subject of each segment could be developed, for example, for information retrieval purposes.

References

Beeferman D., Berger A. and Lafferty J. (1997) *Text segmentation using exponential models*, Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing

Church K. W. and Hanks P. (1990) *Word association norms, mutual information and lexicography*, Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, pp. 76-83

Grosz, B. J. and Sidner, C. L. (1986) *Attention, intentions and the structure of discourse*, Computational Linguistics, 12(3), pp. 175-204

Halliday M. A. K. and Hasan R. (1976) *Cohesion in English*, Longman Group

Hearst M. A. (1993) *TextTiling: A quantitative approach to discourse segmentation*, Technical Report 93/24, Sequoia 2000, University of California, Berkeley

Hearst M. A. (1994) *Multi-paragraph segmentation of expository texts*, Report No. UCB/CSD 94/790, University of California, Berkeley

Jobbins A. C and Evett L. J. (1995) *Automatic identification of cohesion in texts: Exploiting the lexical organisation of Roget's Thesaurus*, Proceedings of ROCLING VIII, Taipei, Taiwan

Jobbins A. C. and Evett L. J. (1998) *Semantic Information from Roget's Thesaurus: Applied to the Correction of Cursive Script Recognition Output*, Proceedings of the International Conference on Computational Linguistics, Speech and Document Processing, India, pp. 65-70

Keenan F. G and Evett L. J. (1989) *Lexical structure for natural language processing*, Proceedings of the 1st International Lexical Acquisition Workshop at IJCAI

Kozima H. (1993) *Text segmentation based on similarity between words*, Proceedings of the 31st Annual Meeting on the Association for Computational Linguistics, pp. 286-288

Litman D. J. and Passonneau R. J. (1996) *Combining knowledge sources for discourse segmentation*, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics

Morris J. and Hirst G. (1991) *Lexical cohesion computed by thesaural relations as an indicator of the structure of text*, Computational Linguistics, 17(1), pp. 21-48

Ponte J. M. and Croft W. B. (1997) *Text Segmentation by Topic*, 1st European Conference on Research and Advanced Technology for Digital Libraries (ECDL'97), pp. 113-125

Reynar J. C. (1994) *An automatic method of finding topic boundaries*, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (Student Session), pp. 331-333

Rotondo J. A. (1984) *Clustering analysis of subjective partitions of text*, Discourse Processes, 7, pp. 69-88

Salton G. and Buckley C. (1991) *Global text matching for information retrieval*, Science, 253, pp. 1012-1015

Salton G. and Buckley C. (1992) *Automatic text structuring experiments* in "Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval," P. S. Jacobs, ed, Lawrence Earlbaum Associates, New Jersey, pp. 199-210

Salton G., Allen J. and Buckley C. (1994) *Automatic structuring and retrieval of large text files*, Communications of the Association for Computing Machinery, 37(2), pp. 97-108

Stairmand M. A. (1997) *Textual context analysis for information retrieval*, Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, pp. 140-147

Yaari Y. (1997) *Segmentation of expository texts by hierarchical agglomerative clustering*, RANLP'97, Bulgaria