Cohesion and Collocation: Using Context Vectors in Text Segmentation

Stefan Kaufmann
CSLI, Stanford University
Linguistics Dept., Bldg. 460
Stanford, CA 94305-2150, U.S.A.
kaufmann@csli.stanford.edu

Abstract

Collocational word similarity is considered a source of text cohesion that is hard to measure and quantify. The work presented here explores the use of information from a training corpus in measuring word similarity and evaluates the method in the text segmentation task. An implementation, the `VecTile` system, produces similarity curves over texts using pre-compiled vector representations of the contextual behavior of words. The performance of this system is shown to improve over that of the purely string-based `TextTiling` algorithm (Hearst, 1997).

1 Background

The notion of text cohesion rests on the intuition that a text is "held together" by a variety of internal forces. Much of the relevant linguistic literature is indebted to Halliday and Hasan (1976), where cohesion is defined as a network of relationships between locations in the text, arising from (i) grammatical factors (co-reference, use of pro-forms, ellipsis and sentential connectives), and (ii) lexical factors (reiteration and collocation). Subsequent work has further developed this taxonomy (Hoey, 1991) and explored its implications in such areas as paragraphing (Longacre, 1979; Bond and Hayes, 1984; Stark, 1988), relevance (Sperber and Wilson, 1995) and discourse structure (Grosz and Sidner, 1986).

The lexical variety of cohesion is semantically defined, invoking a measure of word similarity. But this is hard to measure objectively, especially in the case of collocational relationships, which hold between words primarily because they "regularly co-occur." Halliday and Hasan refrained from a deeper analysis, but hinted at a notion of "degrees of proximity in the lexical system, a function of the probability with which one tends to co-occur with another." (p. 290)

The `VecTile` system presented here is designed to utilize precisely this kind of lexical relationship, relying on observations on a large training corpus to derive a measure of similarity between words and text passages.

2 Related Work

Previous approaches to calculating cohesion differ in the kind of lexical relationship they quantify and in the amount of semantic knowledge they rely on. *Topic parsing* (Hahn, 1990) utilizes both grammatical cues and semantic inference based on pre-coded domain-specific knowledge. More general approaches assess word similarity based on thesauri (Morris and Hirst, 1991) or dictionary definitions (Kozima, 1994).

Methods that solely use observations of patterns in vocabulary use include *vocabulary management* (Youmans, 1991) and the *blocks* algorithm implemented in the `TextTiling` system (Hearst, 1997). The latter is compared below with the system introduced here.

A good recent overview of previous approaches can be found in Chapters 4 and 5 of (Reynar, 1998).

3 The Method

3.1 Context Vectors

The `VecTile` system is based on the `WordSpace` model of (Schütze, 1997; Schütze, 1998). The idea is to represent words by encoding the environments in which they typically occur in texts. Such a representation can be obtained automatically and often provides sufficient information to make deep linguistic analysis unnecessary. This has led to promising results in information retrieval and related areas (Flournoy et al., 1998a; Flournoy et al., 1998b).

Given a dictionary W and a relatively small set C of meaningful "content" words, for each pair in $W \times C$, the number of times is recorded that the two co-occur within some measure of distance in a training corpus. This yields a $|C|$-dimensional vector for each $w \in W$. The direction that the vector has in the resulting $|C|$-dimensional space then represents the collocational behavior of w in the training corpus. In the present implementation, $|W| = 20,500$ and $|C| = 1000$. For computational efficiency and to avoid the high number of zero values in the resulting matrix, the matrix is reduced to 100 dimensions using Singular-Value Decomposition (Golub and van Loan, 1989).

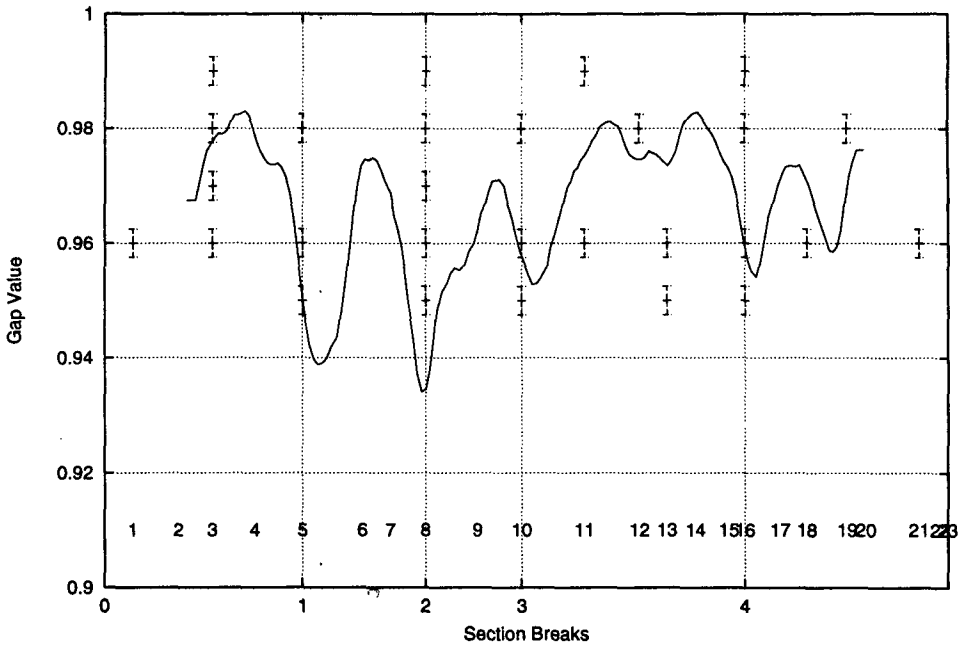

Figure 1: Example of a `VecTile` similarity plot

As a measure of similarity in collocational behavior between two words, the cosine between their vectors is computed: Given two n-dimensional vectors $\vec{v}, \vec{w}$,

$$cos(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2 \sum_{i=1}^{n} w_i^2}} \quad (1)$$

3.2 Comparing Window Vectors

In order to represent pieces of text larger than single words, the vectors of the constituent words are added up. This yields new vectors in the same space, which can again be compared against each other and word vectors. If the word vectors in two adjacent portions of text are added up, then the cosine between the two resulting vectors is a measure of the lexical similarity between the two portions of text.

The `VecTile` system uses word vectors based on co-occurrence counts on a corpus of New York Times articles. Two adjacent windows (200 words each in this experiment) move over the input text, and at pre-determined intervals (every 10 words), the vectors associated with the words in each window are added up, and the cosine between the resulting window vectors is assigned to the gap between the windows in the text. High values indicate lexical closeness. Troughs in the resulting similarity curve mark spots with low cohesion.

3.3 Text Segmentation

To evaluate the performance of the system and facilitate comparison with other approaches, it was used in text segmentation. The motivating assumption behind this test is that cohesion reinforces the topical unity of subparts of text and lack of it correlates with their boundaries, hence if a system correctly predicts segment boundaries, it is indeed measuring cohesion. For want of a way of observing cohesion directly, this indirect relationship is commonly used for purposes of evaluation.

4 Implementation

The implementation of the text segmenter resembles that of the `TextTiling` system (Hearst, 1997). The words from the input are stemmed and associated with their context vectors. The similarity curve over the text, obtained as described above, is smoothed out by a simple low-pass filter, and low points are assigned *depth scores* according to the difference between their values and those of the surrounding peaks. The mean and standard deviation of those depth scores are used to calculate a cutoff below which a trough is judged to be near a section break. The nearest paragraph boundary is then marked as a section break in the output.

An example of a text similarity curve is given in Figure 1. Paragraph numbers are inside the plot at the bottom. Speaker judgments by five subjects are inserted in five rows in the upper half.

Table 1: Precision and recall on the text segmentation task

	TextTiling		VecTile		Subjects	
Text #	Prec	Rec	Prec	Rec	Prec	Rec
1	60	50	60	50	75	77
2	14	20	100	80	76	76
3	50	50	50	50	72	73
4	25	50	10	25	70	75
5	10	25	40	50	70	74
avg	32	40	52	51	73	75

The crucial difference between this and the **TextTiling** system is that the latter builds window vectors solely by counting the occurrences of strings in the windows. Repetition is rewarded by the present approach, too, as identical words contribute most to the similarity between the block vectors. However, similarity scores can be high even in the absence of pure string repetition, as long as the adjacent windows contain words that co-occur frequently in the training corpus. Thus what a direct comparison between the systems will show is whether the addition of collocational information gleaned from the training corpus sharpens or blunts the judgment.

For comparison, the **TextTiling** algorithm was implemented and run with the same window size (200) and gap interval (10).

5 Evaluation

5.1 The Task

In a pilot study, five subjects were presented with five texts from a popular-science magazine, all between 2,000 and 3,400 words, or between 20 and 35 paragraphs, in length. Section headings and any other clues were removed from the layout. Paragraph breaks were left in place. Thus the task was not to find paragraph breaks, but breaks between multi-paragraph passages that according to the the subject's judgment marked topic shifts. All subjects were native speakers of English.[1]

[1] The instructions read:
"You will be given five magazine articles of roughly equal length with section breaks removed. Please mark the places where the topic seems to change (draw a line between paragraphs). Read at normal speed, do not take much longer than you normally would. But do feel free to go back and reconsider your decisions (even change your markings) as you go along.
Also, for each section, suggest a headline of a few words that captures its main content.
If you find it hard to decide between two places, mark both, giving preference to one and indicating that the other was a close rival."

5.2 Results

To obtain an "expert opinion" against which to compare the algorithms, those paragraph boundaries were marked as "correct" section breaks which at least three out of the five subjects had marked. (Three out of seven (Litman and Passonneau, 1995; Hearst, 1997) or 30% (Kozima, 1994) are also sometimes deemed sufficient.) For the two systems as well as the subjects, precision and recall with respect to the set of "correct" section breaks were calculated. The results are listed in Table 1.

The context vectors clearly led to an improved performance over the counting of pure string repetitions.

The simple assignment of section breaks to the nearest paragraph boundary may have led to noise in some cases; moreover, it is not really part of the task of measuring cohesion. Therefore the texts were processed again, this time moving the windows over whole paragraphs at a time, calculating gap-values at the paragraph gaps. For each paragraph break, the number of subjects who had marked it as a section break was taken as an indicator of the "strength" of the boundary. There was a significant negative correlation between the values calculated by both systems and that measure of strength, with $r = -.338 (p = .0002)$ for the **VecTile** system and $r = -.220 (p = .0172)$ for **TextTiling**. In other words, deep gaps in the similarity measure are associated with strong agreement between subjects that the spot marks a section boundary. Although r^2 is low both cases, the **VecTile** system yields more significant results.

5.3 Discussion and Further Work

The results discussed above need further support with a larger subject pool, as the level of agreement among the judges was at the low end of what can be considered significant. This is shown by the *Kappa* coefficients, measured against the expert opinion and listed in Table 2. The overall average was .594.

Despite this caveat, the results clearly show that adding collocational information from the training

Table 2: *Kappa* coefficients

Text#	Subject# 1	2	3	4	5	$\overline{K}$
1	.775	.629	.596	.444	.642	.617
2	.723	.649	.491	.753	.557	.635
3	.859	.121	.173	.538	.738	.486
4	.870	.532	.635	.299	.870	.641
5	.833	.500	.625	.423	.500	.576
All texts	.814	.491	.508	481	.675	.594

corpus improves the prediction of section breaks, hence, under common assumptions, the measurement of lexical cohesion. It is likely that these encouraging results can be further improved. Following are a few suggestions of ways to do so.

Some factors work against the context vector method. For instance, the system currently has no mechanism to handle words that it has no context vectors for. Often it is precisely the co-occurrence of uncommon words not in the training corpus (personal names, rare terminology etc.) that ties text together. Such cases pose no challenge to the string-based system, but the `VecTile` system cannot utilize them. The best solution might be a hybrid system with a backup procedure for unknown words.

Another point to note is how well the much simpler `TextTile` system compares. Indeed, a close look at the figures in Table 1 reveals that the better results of the `VecTile` system are due in large part to one of the texts, viz. #2. Considering the additional effort and resources involved in using context vectors, the modest boost in performance might often not be worth the effort in practice. This suggests that pure string repetition is a particularly strong indicator of similarity, and the vector-based system might benefit from a mechanism to give those vectors a higher weight than co-occurrences of merely similar words.

Another potentially important parameter is the nature of the training corpus. In this case, it consisted mainly of news texts, while the texts in the experiment were scientific expository texts. A more homogeneous setting might have further improved the results.

Finally, the evaluation of results in this task is complicated by the fact that "near-hits" (cases in which a section break is off by one paragraph) do not have any positive effect on the score. This problem has been dealt with in the Topic Detection and Tracking (TDT) project by a more flexible score that becomes gradually worse as the distance between hypothesized and "real" boundaries increases (TDT, 1997a; TDT, 1997b).

Acknowledgements

Thanks to Stanley Peters, Yasuhiro Takayama, Hinrich Schütze, David Beaver, Edward Flemming and three anonymous reviewers for helpful discussion and comments, to Stanley Peters for office space and computational infrastructure, and to Raymond Flournoy for assistance with the vector space.

References

S.J. Bond and J.R. Hayes. 1984. Cues people use to paragraph text. *Research in the Teaching of English*, 18:147–167.

Raymond Flournoy, Ryan Ginstrom, Kenichi Imai, Stefan Kaufmann, Genichiro Kikui, Stanley Peters, Hinrich Schütze, and Yasuhiro Takayama. 1998a. Personalization and users' semantic expectations. ACM SIGIR'98 Workshop on Query Input and User Expectations, Melbourne, Australia.

Raymond Flournoy, Hiroshi Masuichi, and Stanley Peters. 1998b. Cross-language information retrieval: Some methods and tools. In D. Hiemstra, F. de Jong, and K. Netter, editors, *TWLT 14 Language Technology in Multimedia Information Retrieval*, pages 79–83.

Talmy Givòn, editor. 1979. *Discourse and Syntax*. Academic Press.

G. H. Golub and C. F. van Loan. 1989. *Matrix Computations*. Johns Hopkins University Press.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Udo Hahn. 1990. Topic parsing: Accounting for text macro structures in full-text analysis. *Information Processing and Management*, 26:135–170.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Marti Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press.

Hideki Kozima. 1994. *Computing Lexical Cohesion as a Tool for Text Analysis*. Ph.D. thesis, University of Electro-Communications.

Chin-Yew Lin. 1997. *Robust Automatic Topic Identification*. Ph.D. thesis, University of Southern California. [Online] `http://www.isi.edu/~cyl/thesis/thesis.html` [1999, April 24].

Diane J. Litman and Rebecca J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd ACL*, pages 108–115.

L.E. Longacre. 1979. The paragraph as a grammatical unit. In Givòn (Givòn, 1979), pages 115–134.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indication of the structure of text. *Computational Linguistics*, 17(1):21–48.

Jeffrey C. Reynar. 1998. *Topic Segmentation: Algorithms and Applications.* Ph.D. thesis, University of Pennsylvania. [Online] `http://www.cis.edu/~jcreynar/research.html` [1999, April 24].

K. Richmond, A. Smith, and E. Amitay. 1997. Detecting subject boundaries within text: A language independent statistical approach. In *Proceedings of The Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2).*

Hinrich Schütze. 1997. *Ambiguity Resolution in Language Learning.* CSLI.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Dan Sperber and Deidre Wilson. 1995. *Relevance: Communication and Cognition.* Harvard University Press, 2nd edition.

Heather Stark. 1988. What do paragraph markings do? *Discourse Processes*, 11(3):275–304.

1997a. *The* TDT *Pilot Study Corpus Documentation version 1.3*, 10. Distributed by the Linguistic Data Consortium.

1997b. *The Topic Detection and Tracking (*TDT*) Pilot Study Evaluation Plan*, 10. Distributed by the Linguistic Data Consortium.

Gilbert Youmans. 1991. A new tool for discourse analysis: The vocabulary-management profile. *Language*, 47(4):763–789.