Contextually–Mediated Semantic Similarity Graphs for Topic Segmentation

Geetu Ambwani
StreamSage/Comcast
Washington, DC, USA
ambwani@streamsage.com

Anthony R. Davis
StreamSage/Comcast
Washington, DC, USA
davis@streamsage.com

Abstract

We present a representation of documents as directed, weighted graphs, modeling the range of influence of terms within the document as well as contextually determined semantic relatedness among terms. We then show the usefulness of this kind of representation in topic segmentation. Our boundary detection algorithm uses this graph to determine topical coherence and potential topic shifts, and does not require labeled data or training of parameters. We show that this method yields improved results on both concatenated pseudo-documents and on closed-captions for television programs.

1 Introduction

We present in this paper a graph-based representation of documents that models both the long-range scope "influence" of terms and the semantic relatedness of terms in a local context. In these graphs, each term is represented by a series of nodes. Each node in the series corresponds to a sentence within the span of that term's influence, and the weights of the edges are proportional to the semantic relatedness among terms in the context. Semantic relatedness between terms is reinforced by the presence of nearby, closely related terms, reflected in increased connection strength between their nodes.

We demonstrate the usefulness of our representation by applying it to partitioning of documents into topically coherent segments. Our segmentation method finds locations in the graph of a document where one group of strongly connected nodes ends and another begins, signaling a shift in topicality. We test this method both on concatenated news articles, and on a more realistic segmentation task, closed-captions from commercial television programs, in which topic transitions are more subjective and less distinct. Our methods are unsupervised and require no training; thus they do not require any labeled instances of segment boundaries. Our method attains results significantly superior to that of Choi (2000), and approaches human performance on segmentation of television closed-captions, where inter-annotator disagreement is high.

2 Graphs of lexical influence

2.1 Summary of the approach

Successful topic segmentation requires some representation of semantic and discourse cohesion, and the ability to detect where such cohesion is weakest. The underlying assumption of segmentation algorithms based on lexical chains or other term similarity measures between portions of a document is that continuity in vocabulary reflects topic continuity. Two short examples illustrating topic shifts in television news programs, with accompanying shift in vocabulary, appear in Figure 1.

We model this continuity by first modeling what the extent of a term's influence is. This differs from a lexical chain approach in that we do not model text cohesion through recurrence of terms. Rather, we determine, for each occurrence of a term in the document (excluding terms generally treated as stopwords), what interval of sentences surrounding that occurrence is the best estimate of the extent of its relevance. This idea stems from work in Davis, et al. (2004), who describe the use of *relevance intervals* in multimedia information retrieval. We summarize their procedure for constructing relevance intervals in

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 60–68,
Uppsala, Sweden, 16 July 2010.

section 2.2. Next, we calculate the relatedness of these terms to one another. We use pointwise mutual information (PMI) as a similarity measure between terms, but other measures, such as WordNet-based similarity or Wikipedia Miner similarity (Milne and Witten, 2009), could augment or replace it.

S_44 Gatorade has discontinued a drink with his image but that was planned before the company has said and they have issued a statement in support of tiger woods.
S_45 And at t says that while it supports tiger woods personally, it is evaluating their ongoing business relationship.
S_46 I'm sure, alex, in the near future we're going to see more of this as companies weigh the short term difficulties of staying with tiger woods versus the long term gains of supporting him fully.
S_47 Okay.
S_48 Mark potter, miami.
S_49 Thanks for the wrapup of that.
S_50 We'll go now to deep freeze that's blanketing the great lakes all the way to right here on the east coast.

S_190 We've got to get this addressed and hold down health care costs.
S_191 Senator ron wyden the optimist from oregon, we appreciate your time tonight.
S_192 Thank you.
S_193 Coming up, the final day of free health clinic in kansas city, missouri.

Figure 1. Two short closed-caption excerpts from television news programs, each containing a topic shift

The next step is to construct a graphical representation of the influence of terms throughout a document. When constructing topically coherent segments, we wish to assess coherence from one sentence to the next. We model similarity between successive sentences as a graph, in which each node represents both a term and a sentence that lies within its influence (that is, a sentence belonging to a relevance interval for that term). For example, if the term "drink" occurs in sentence 44, and its relevance interval extends to sentence 47, four nodes will be created for "drink", each corresponding to one sentence in that interval. The edges in the graph connect nodes in successive sentences. The weight of an edge between two terms *t* and *t'* consists not only of their relatedness, but is reinforced by the presence of other nodes in each sentence associated with terms related to *t* and *t'*.

The resulting graph thus consists of cohorts of nodes, one cohort associated with each sentence, and edges connecting nodes in one cohort to those in the next. Edges with a low weight are pruned from the graph. The algorithm for determining topic segment boundaries then seeks locations in which a relatively large number of relevance intervals for terms with relatively high relatedness end or begin.

In sum, we introduce two innovations here in computing topical coherence. One is that we use the extent of each term's relevance intervals to model the influence of that term, which thus extends beyond the sentences it occurs in. Second, we amplify the semantic relatedness of a term *t* to a term *t'* when there are other nearby terms related to *t* and *t'*. Related terms thereby reinforce one another in establishing coherence from one sentence to the next.

2.2 Constructing relevance intervals

As noted, the scope of a term's influence is captured through relevance intervals (RIs). We describe here how RIs are created. A corpus—in this case, seven years of *New York Times* text totaling approximately 325 million words—is run through a part-of-speech tagger. The pointwise mutual information between each pair of terms is computed using a 50-word window.[1]

PMI values provide a mechanism to measure relatedness between a term and terms occurring in nearby sentences of a document. When processing a document for segmentation, we first calculate RIs for all the terms in that document. An RI for a term *t* is built sentence-by-sentence, beginning with a sentence where *t* occurs. A sentence immediately succeeding or preceding the sentences already in the RI is added to that RI if it contains terms with sufficiently high PMI values with *t*. An adjacent sentence is also added to an RI if there is a pronominal believed to refer to *t*; the algorithm for determining pronominal reference is closely based on Kennedy and Boguraev (1996). Expansion of an RI is terminated if there are no motivations for expanding it further. Additional termination conditions can be included as well. For example, if large local voca-

[1] PMI values are constructed for all words other than those in a list of stopwords. They are also constructed for a limited set of about 100,000 frequent multi-word expressions. In our segmentation system, we use only the RIs for nouns and for multiword expressions.

bulary shifts or discourse cues signaling the start of end of a section are detected, RIs can be forced to end at those points. In one version of our system, we set these "hard" boundaries using an algorithm based on Choi (2000). In this paper we report segmentation results with and without this limited use of Choi's algorithm. Lastly, if two RIs for *t* are sufficiently close (i.e., the end of one lies within 150 words of the start of another), then the two RIs are merged.

The aim of constructing RIs is to determine which portions of a document are relevant to a particular term. While this is related to the goal of finding topically coherent segments, it is of course distinct, as a topic typically is determined by the influence of multiple terms. However, RIs do provide a rough indication of how far a term's influence extends or, put another way, of "smearing out" the occurrence of a term over an extended region.

2.3 From relevance intervals to graphs

Consider a sentence S_i, and its immediate successor S_{i+1}. Each of these sentences is contained in various relevance intervals; let W_i denote the set of terms with RIs containing S_i, and W_{i+1} denote the set containing S_{i+1}.

For each pair of terms a in W_i and b in W_{i+1}, we compute a connection strength $c(a,b)$, a non-negative real number that reflects how the two terms are related in the context of S_i and S_{i+1}. To include the context, we take into account that some terms in S_i may be closely related, and should support one another in their connections to terms in S_{i+1}, and vice versa, as suggested above. Here, we use PMI values between terms as the basis for connection strength, normalized to a similarity score that ranges between 0 and 1, as follows:

$$s(x,y) = 1 - \frac{1}{\exp\left(PMI(x,y)\right)} \quad (1)$$

The similarity between two terms is set to 0 if this quantity is negative. Also, we assign the maximum value of 1 for self-similarity. We then define connection strength in the following way:

$$c(a,b) = \sum_{x \in W_i} s(x,a)s(x,b) + \sum_{y \in W_{i+1}} s(y,a)s(y,b) \quad (2)$$

That is, the similarity of another term in W_i or W_{i+1} to b or a respectively, will add to the connection strength between a and b, weighted by the similarity of that term to a or b respectively. Note that this formula also includes in the summation the similarity $s(a,b)$ between a and b themselves, when either x or y is set to either a or b.[2] Figure 2 illustrates this procedure. We normalize the connection strength by the total number of pairs in equation (2).

We note in passing that many possible modifications of this formula are easily imagined. One obvious alternative to using the product of two similarity scores is to use the minimum of the two scores. This gives more weight to pair values that are both moderately high, with respect to pairs where one is high and the other low. Apart from this, we could incorporate terms from RIs in sentences beyond these two adjoining sentences, we could weight individual terms in W_i or W_{i+1} according to some independent measure of topical salience, and so on.

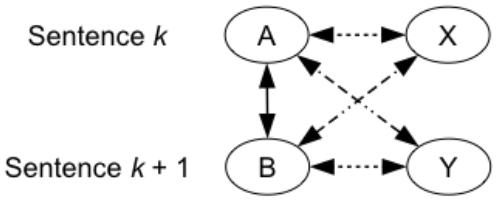

Figure 2. Calculation of connection strength between two nodes

What emerges from this procedure is a weighted graph of connections across slices of a document (sentences, in our experiments). Each node in the graph is labeled with a term and a sentence number, and represents a relevance interval for that term that includes the indicated sentence. The edges of the graph connect nodes associated with adjacent sentences, and are weighted by the connection strength. Because many weak connections are present in this graph, we remove edges that are unlikely to contribute to establishing topical coherence. There are various options for pruning: removing edges with connection strengths below a threshold, retaining only the top n edges, cutting the graph between two sentences where the total connection strength of edges connecting the sentences is small, and using an edge betweenness algorithm (e.g., Girvan and Newman, 2002) to remove edges that have high betweenness (and hence are indicative of a "thin" connection).

[2] In fact, the similarity $s(a_i,b_j)$ will be counted twice, once in each summation in the formula above; we retain this additional weighting of $s(a_i,b_j)$.

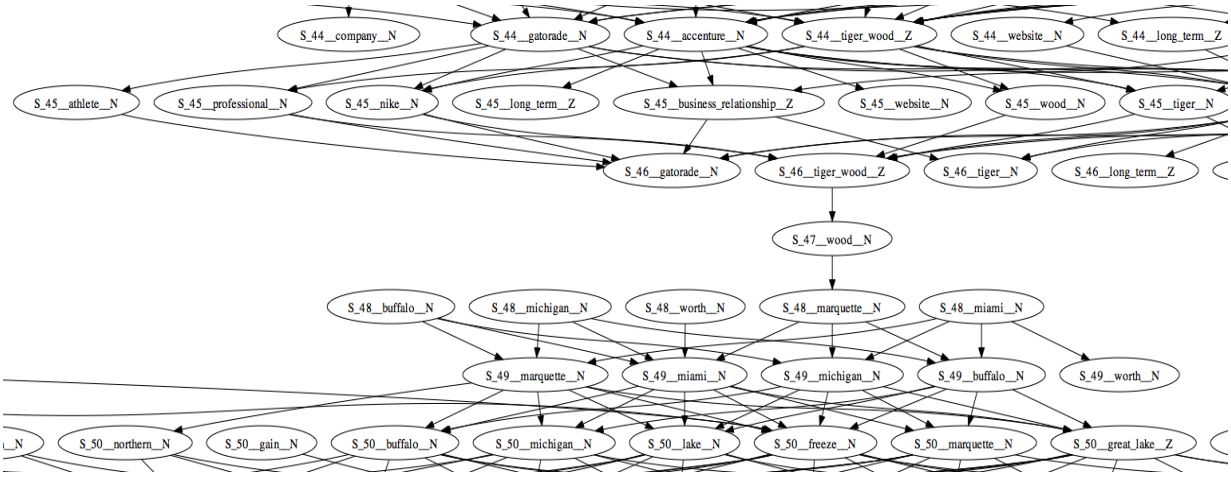

Figure 3. A portion of the graph generated from the first excerpt in Figure 1. Each node is labeled S_*i*__*term_pos*, where *i* indicates the sentence index

We have primarily investigated the first method, removing edges with a connection strength less than 0.5. Two samples of the graphs we produce, corresponding to the excerpts in figure 1, appear in figures 3 and 4.

2.4 Finding segment boundaries in graphs

Segment boundaries in these graphs are hypothesized where there are relatively few, relatively weak connections from a cohort of nodes associated with one sentence to the cohort of nodes associated with the following sentence. If a term has a node in one cohort and in the succeeding cohort (that is, its RI continues across the two corresponding sentences) it counts against a segment boundary at that location, whereas terms with nodes on only one side of the boundary count in favor of a segment. For example, in figure 3, a new set of RIs start in sentence 48, where we see nodes for "Buffalo", "Michigan", "Worth", Marquette", and "Miami", and RIs in preceding sentences for "Tiger Woods", "Gatorade", etc. end. Note that the corresponding excerpt in figure 1 shows a clear topic shift between a story on Tiger Woods ending at sentence 46, and a story about Great Lakes weather beginning at sentence 48.

Similarly, in figure 4, RIs for "Missouri", "city" and "health clinic" include sentences 190. 191, and 192; thus these are evidence against a segment boundary at this location. On the other hand, several other terms, such as "Oregon", "Ron", "Senator", and "bill", have RIs that end at sentence 191, which argues in favor of a boundary there. We present further details of our boundary heuristics in section 4.1.

3 Related Work

The literature on topic segmentation has mostly focused on detecting a set of segments, typically non-hierarchical and non-overlapping, exhaustively composing a document. Evaluation is then relatively simple, employing pseudo-documents constructed by concatenating a set of documents. This is a suitable technique for detecting coarse-grained topic shifts. As Ferret (2007) points out, approaches to the problem vary both in the kinds of knowledge they depend on, and on the kinds of features they employ.

Research on topic segmentation has exploited information internal to the corpus of documents to be segmented and information derived from external resources. If a corpus of documents pertinent to a domain is available, statistical topic models such as those developed by Beeferman et al. (1999) or Blei and Moreno (2001) can be tailored to documents of that type. Lexical cohesion techniques include similarity measures between adjacent blocks of text, as in TextTiling (Hearst, 1994, 1997) and lexical chains based on recurrences of a term or related terms, as in Morris and Hirst (1991), Kozima (1993), and Galley, et al. (2003). In Kan, et al. (1998) recurrences of the same term within a certain number of sentences are used for chains (the number varies with the type of term), and chains are based on entity reference as well as lexical identity. Our method is related to lexical chain techniques, in that the graphs we construct contain chains of nodes that extend the influence of a term beyond the site where it occurs. But we differ in that we do not require a term (or a semantically related term) to recur, in order to build such chains.

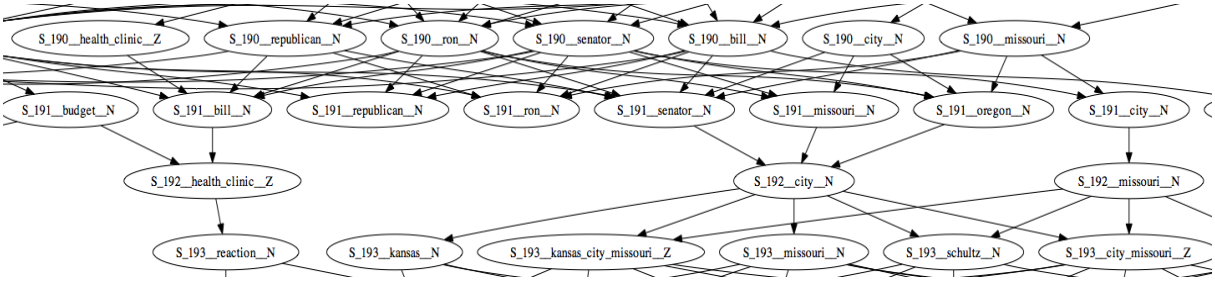

Figure 4. A portion of the graph generated from the second excerpt in Figure 1. Each node is labeled S_*i*__*term_pos*, where *i* indicates the sentence index

In this respect, our approach also resembles that of Matveeva and Levow (2007), who build semantic similarity among terms into their lexical cohesion model through latent semantic analysis. Our techniques differ in that we incorporate semantic relatedness between terms directly into a graph, rather than computing similarities between blocks of text.

In our experiments, we compare our method to C99 (Choi, 2000), an algorithm widely treated as a baseline. Choi's algorithm is based on a measure of local coherence; vocabulary similarity between each pair of sentences in a document is computed and the similarity scores of nearby sentences are ranked, with boundaries hypothesized where similarity across sentences is low.

4 Experiments, results, and evaluation

4.1 Systems compared

As noted above, we tested our system against the C99 segmentation algorithm (Choi, 2000). The implementation of C99 we use comes from the MorphAdorner website (Burns, 2006). We also compared our system to two simpler baseline systems without RIs. One uses graphs that do not represent a term's zone of influence, but contain just a single node for each occurrence of a term. The second represents a term's zone of influence in an extremely simple fashion, as a fixed number of sentences starting from each occurrence of that term. We tried several values ranging from 5 to 20 sentences for this extension. In addition, we varied two parameters to find the best-performing combination of settings: the threshold for pruning low-weight edges, and the threshold for positing a segment boundary. In both the single-node and fixed-extension systems, the connection strength between nodes is calculated in the same way as for our full system. These comparisons aim to demonstrate two things. First, segmentation is greatly improved when we extend the influence of terms beyond the sentences they occur in. Second, the RIs prove more effective than fixed-length extensions in modeling that influence accurately.

Lastly, to establish how much we can gain from using Choi's algorithm to determine termination points for RIs, we also compared two versions of our system: one in which RIs are calculated without information from Choi's algorithm and a second with these boundaries included.

Table 1 lists the systems we compare in the experiments described below.

C99	Implementation of Choi (2000)
SS+C	Our full Segmentation System, incorporating "hard" boundaries determined by modified Choi algorithm
SS	Our system, using RIs without "hard" boundaries determined by modified Choi algorithm
FE	Our system, using fixed extension of a term from its occurrence
SN	Our system, using a single node for each term occurrence (no extension)

Table 1. Systems compared in our experiments

4.2 Data and parameter settings

We tested our method on two sets of data. One set consists of concatenated news stories, following the approach of Choi (2000) and others since; the other consists of closed captions for twelve U.S. commercial television programs. Because the notion of a topic is inherently subjective, we follow many researchers who have reported results on "pseudo-documents"–documents formed by concatenating several randomly selected documents–so that the boundaries of segments are known, sharp, and not dependent on annotator variability (Choi, 2000). However, we also are

interested in our system's performance on more realistic segmentation tasks, as noted in the introduction.

In testing our algorithm, we first generated graphs from the documents in each dataset, as described in section 2. We pruned edges in the graphs with connection strength of less than 0.5. To find segment boundaries, we seek locations where the number of common terms associated with successive sentences is at a minimum. This quantity needs to be normalized by some measure of how many nodes are present on either side of a potential boundary. We tested three normalization factors: the total number of nodes on both sides of the potential segment boundary, the maximum of the numbers of nodes on each side of the boundary, and the minimum of the numbers of nodes on each side of the boundary. The results for all three of these were very similar, so we report only those for the maximum. This measure provides a ranking of all possible boundaries in a document (that is, between each pair of consecutive sentences), with a value of 0 being most indicative of a boundary. After experimenting with a few threshold values, we selected a threshold of 0.6, and posit a boundary at each point where the measure falls below this threshold.

4.3 Evaluation metrics

We compute precision, recall, and F-measure based on exact boundary matches between the system and the reference segmentation. As numerous researchers have pointed out, this alone is not a perspicacious way to evaluate a segmentation algorithm, as a system that misses a gold-standard boundary by one sentence would be treated just like one that misses it by ten. We therefore computed two additional, widely used measures, P_k (Beeferman, et al., 1997) and WindowDiff (Pevzner and Hearst, 2002). P_k assesses a penalty against a system for each position of a sliding window across a document in which the system and the gold standard differ on the presence or absence of (at least one) segment boundary. WindowDiff is similar, but where the system differs from the gold standard, the penalty is equal to the difference in the number of boundaries between the two. This penalizes missed boundaries and "near-misses" less than P_k (but see Lamprier, et al., (2007) for further analysis and some criticism of WindowDiff). For both P_k and WindowDiff, we used a window size of half the average reference segment length, as suggested by Beeferman, et al. (1997). P_k and WindowDiff values range between 0 and 1, with lower values indicating better performance in detecting segment boundaries. Note that both P_k and WindowDiff are asymmetrical measures; different values will result if the system's and the gold-standard's boundaries are switched.

4.4 Concatenated *New York Times* articles

The concatenated pseudo-documents consist of *New York Times* articles selected at random from the *New York Times* Annotated Corpus.[3] Each pseudo-document contains twenty articles, with an average of 623.6 sentences. Our test set consists of 185 of these pseudo-documents.[4]

N = 185						
		Prec.	**Rec.**	**F**	**P_k**	**WD**
C99	μ	0.404	0.569	0.467	0.338	0.360
	s.d	0.106	0.121	0.105	0.109	0.135
SS	μ	0.566	0.383	0.448	0.292	0.317
	s.d.	0.176	0.135	0.140	0.070	0.084
SS +C	μ	0.578	0.535	0.537	0.262	0.283
	s.d.	0.148	0.197	0.150	0.081	0.098
FE	μ	0.265	0.140	0.176	0.478	0.536
	s.d.	0.123	0.042	0.055	0.055	0.076
SN	μ	0.096	0.112	0.099	0.570	0.702
	s.d.	0.040	0.024	0.027	0.072	0.164

Table 2. Performance of C99 and SS on segmentation of concatenated *New York Times* articles, without specifying a number of boundaries

Tables 2 and 3 provide summary results on the concatenated news articles. We ran the five systems listed in table 1 on the full dataset without any additional restrictions on the number of article boundaries to be detected. Means and standard deviations for each method on the five metrics are displayed in table 2. C99 typically finds many more boundaries than the 20 that are present (30.65 on average). Our SS system finds fewer than the true number of boundaries (14.52 on average), while the combined system SS+C finds almost precisely the correct number (19.98 on average). We used one-tailed paired t-tests of equal means to determine statistical significance at the 0.01 level. Although only SS+C's performance is significantly better in terms of F-

[3] www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19

[4] Only article text is used, though occasionally some obvious heading material, such as book title, author and publisher at the beginning of a book review, is present also.

measure, both versions of our system outperform C99 according to P_k and WindowDiff.

Using the baseline single node system (SN) yields very poor performance. These results (table 2, row SN) are obtained with the edge-pruning threshold set to a connection strength of 0.9, and the boundary threshold set to 0.2, at which the average number of boundaries found is 26.86. Modeling the influence of terms beyond the sentences they occur in is obviously valuable.

The baseline fixed-length extensions system (FE) does better than SN but significantly worse than RIs. We found that, among the parameter settings yielding between 10 and 30 boundaries per document on average, the best results occur with the extension set to 6 sentences, the edge-pruning threshold set to a connection strength of 0.5, and the boundary threshold set to 0.7. The results for this setting are reported in table 2, row FE (the average number of segments per document is 12.5). Varying these parameters has only minor effects on performance, although the number of boundaries found can of course be tuned. RIs clearly provide a benefit over this type of system, by modeling a term's influence dynamically rather than as a fixed interval.

From here on, we report results only for the two systems: C99 and our best-performing system, SS+C.

For 86 of the documents, in which both C99 and SS+C found more than 20 boundaries, we also calculate the performance on the best-scoring 20 boundaries found by each system. These results are displayed in table 3. Note that when the number of boundaries to be found by each system is fixed at the actual number of boundaries, the values of precision and recall are necessarily identical. Here too our system outperforms C99, and the differences are statistically significant, according to a one-tailed paired t-test of equal means at the 0.01 level.

N = 86				
		Prec.=Rec.=F	**P_k**	**WD**
C99	μ	0.530	0.222	0.231
	s.d	0.105	0.070	0.074
SS + C	μ	0.643	0.192	0.201
	s.d.	0.130	0.076	0.085

Table 3. Performance of C99 and SS on segmentation of concatenated *New York Times* articles, selecting the 20 best-scoring boundaries

4.5 Human-annotated television program closed-captions

We selected twelve television programs for which we have closed-captions; they are a mix of headline news (3 shows), news commentary (4 shows), documentary/lifestyle (3 shows), one comedy/drama episode, and one talk show. Only the closed captions are used, no speaker intonation, video analysis, or metadata is employed. The closed captions are of variable quality, with numerous spelling errors.

Five annotators were instructed to indicate topic boundaries in the closed-caption text files. Their instructions were open-ended in the sense that they were not given any definition of what a topic or a topic shift should be, beyond two short examples, were not told to find a specific number of boundaries, but were allowed to indicate how important a topic was on a five-point scale, encouraging them to indicate minor segments or subtopics within major topics if they chose to do so. For some television programs, particularly the news shows, major boundaries between stories on disparate topics are likely be broadly agreed on, whereas in much of the remaining material the shifts may be more fine-grained and judgments varied. In addition, the scripted nature of television speech results in many carefully staged transitions and teasers for upcoming material, making boundaries more diffuse or confounded than in some other genres.

We combined the five annotators' segmentations, to produce a single set of boundaries as a reference. We used a three-sentence sliding window, and if three or more of the five annotators place a boundary in that window, we assign a boundary where the majority of them place it (in case of a tie, we choose one location arbitrarily). Although the annotators are rather inconsistent in their use of this rating system, a given annotator tends to be consistent in the granularity of segmentation employed across all documents. This observation is consistent with the remarks of Malioutov and Barzilay (2006) regarding varying topic granularity across human annotators on spoken material. We thus computed two versions of the combined boundaries, one in which all boundaries are used, and another in which we ignore minor boundaries—those the annotator assigned a score of 1 or 2. We ran our experiments with both versions of the combined boundaries as the reference segmentation.

We use P_k to assess inter-annotator agreement among our five annotators. Table 4 presents two

P_k values for each pair of annotators; one set of values is for all boundaries, while the other is for "major" boundaries, assigned an importance of 3 or greater on the five-point scale. The P_k value for each annotator with respect to the two reference segmentations is also provided.

	A	**B**	**C**	**D**	**E**	***Ref***
A		0.36 *0.48*	0.30 *0.45*	0.27 *0.44*	0.42 *0.67*	0.20 *0.38*
B	0.29 *0.40*		0.29 *0.32*	0.27 *0.33*	0.33 *0.55*	0.20 *0.25*
C	0.57 *0.48*	0.60 *0.44*		0.41 *0.20*	0.67 *0.61*	0.40 *0.18*
D	0.36 *0.46*	0.41 *0.46*	0.27 *0.20*		0.53 *0.63*	0.22 *0.26*
E	0.33 *0.35*	0.31 *0.34*	0.33 *0.30*	0.32 *0.31*		0.25 *0.27*
Ref	0.25 *0.39*	0.32 *0.35*	0.24 *0.17*	0.21 *0.22*	0.42 *0.58*	

Table 4. P_k values for the segmentations produced by each pair of annotators (A-E) and for the combined annotation described in section 4.5; upper values are for all boundaries and *lower values* are for boundaries of segments scored 3 or higher

These numbers are rather high, but comparable to those obtained by Malioutov and Barzilay (2006) in a somewhat similar task of segmenting video recordings of physics lectures. The P_k values are lower for the reference boundary set, which we therefore feel some confidence in using as a reference segmentation.

		Prec.	**Rec.**	**F**	**P_k**	**WD**
All topic boundaries						
C99	μ	0.197	0.186	0.184	0.476	0.507
	s.d	0.070	0.072	0.059	0.078	0.102
SS +C	μ	0.315	0.208	0.240	0.421	0.462
	s.d.	0.089	0.073	0.064	0.072	0.084
Major topic boundaries only						
C99	μ	0.170	0.296	0.201	0.637	0.812
	s.d.	0.063	0.134	0.060	0.180	0.405
SS +C	μ	0.271	0.316	0.271	0.463	0.621
	s.d.	0.102	0.138	0.077	0.162	0.445

Table 5. Performance of C99 and SS+C on segmentation of closed-captions for twelve television programs, with the two reference segmentations using "all topic boundaries" and "major topic boundaries only"

As the television closed-captions are noisy with respect to data quality and inter-annotator disagreement, the performance of both systems is worse than on the concatenated news articles, as expected. We present the summary performance of C99 and SS+C in table 5, again using two versions of the reference. Because of the small test set size, we cannot claim statistical significance for any of these results, but we note that on average SS+C outperforms C99 on all measures.

5 Conclusions and future work

We have presented an approach to text segmentation that relies on a novel graph based representation of document structure and semantics. It successfully models topical coherence using long-range influence of terms and a contextually determined measure of semantic relatedness. Relevance intervals, calculated using PMI and other criteria, furnish an effective model of a term's extent of influence for this purpose. Our measure of semantic relatedness reinforces global co-occurrence statistics with local contextual information, leading to an improved representation of topical coherence. We have demonstrated significantly improved segmentation resulting from this combination, not only on artificially constructed pseudo-documents, but also on noisy data with more diffuse boundaries, where inter-annotator agreement is fairly low.

Although the system we have described here is not trained in any way, it provides an extensive set of parameters that could be tuned to improve its performance. These include various techniques for calculating the similarity between terms and combining those similarities in connection strengths, heuristics for scoring potential boundaries, and thresholds for selecting those boundaries. Moreover, the graph representation lends itself to techniques for finding community structure and centrality, which may also prove useful in modeling topics and topic shifts.

We have also begun to explore segment labeling, identifying the most "central" terms in a graph according to their connection strengths. Those terms whose nodes are strongly connected to others within a segment appear to be good candidates for segment labels.

Finally, although we have so far applied this method only to linear segmentation, we plan to explore its application to hierarchical or overlapping topical structures. We surmise that strongly connected subgraphs may correspond to these more fine-grained aspects of discourse structure.

Acknowledgements

We thank our colleagues David Houghton, Olivier Jojic, and Robert Rubinoff, as well as the anonymous referees, for their comments and suggestions.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1997. Text Segmentation Using Exponential Models. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, 35-46.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

David M. Blei and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. *Proceedings of the 24th Annual Meeting of ACM SIGIR*, 343–348.

Burns, Philip R. 2006. MorphAdorner: Morphological Adorner for English Text. http://morphadorner.northwestern.edu/morphadorner/textsegmenter/.

Freddy Y.Y. Choi. 2000. Advances in domain independent linear text segmentation. *Proceedings of NAACL 2000*, 109-117.

Anthony Davis, Phil Rennert, Robert Rubinoff, Tim Sibley, and Evelyne Tzoukermann. 2004. Retrieving what's relevant in audio and video: statistics and linguistics in combination. *Proceedings of RIAO 2004*, 860-873.

Olivier Ferret. 2007. Finding document topics for improving topic segmentation. *Proceedings of the 45th Annual Meeting of the ACL*, 480–487.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse Segmentation of Multi-Party Conversation. *Proceedings of the 41st Annual Meeting of the ACL*, 562-569.

Michelle Girvan and M.E.J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:12, 7821-7826.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. *Proceedings of the 32nd Annual Meeting of the ACL*, 9-16.

Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23:1, 33-64.

Min-Yen Kan, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear Segmentation and Segment Significance. *Proceedings of the 6th International Workshop on Very Large Corpora*, 197-205.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. *Proceedings of the 16th International Conference on Computational Linguistics*, 113-118.

Hideki Kozima. 1993. Text segmentation based on similarity between words. *Proceedings of the 31st Annual Meeting of the ACL (Student Session)*, 286-288.

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat and Frederic Saubion. 2007. On Evaluation Methodologies for Text Segmentation Algorithms. *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, 19-26.

Igor Malioutov and Regina Barzilay. 2006. Minimum Cut Model for Spoken Lecture Segmentation. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 25–32.

Irina Matveeva and Gina-Anne Levow. 2007. Topic Segmentation with Hybrid Document Indexing. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 351–359,

David Milne and Ian H. Witten. 2009. An Open-Source Toolkit for Mining Wikipedia. http://www.cs.waikato.ac.nz/~dnk2/publications/AnOpenSourceToolkitForMiningWikipedia.pdf.

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

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28:1, 19–36.