# Topic Segmentation with Hybrid Document Indexing

**Irina Matveeva**
Department of Computer Science
University of Chicago
Chicago, IL 60637
matveeva@cs.uchicago.edu

**Gina-Anne Levow**
Department of Computer Science
University of Chicago
Chicago, IL 60637
levow@cs.uchicago.edu

## Abstract

We present a domain-independent unsupervised topic segmentation approach based on hybrid document indexing. Lexical chains have been successfully employed to evaluate lexical cohesion of text segments and to predict topic boundaries. Our approach is based in the notion of semantic cohesion. It uses spectral embedding to estimate semantic association between content nouns over a span of multiple text segments. Our method significantly outperforms the baseline on the topic segmentation task and achieves performance comparable to state-of-the-art methods that incorporate domain specific information.

## 1 Introduction

The goal of topic segmentation is to discover story boundaries in the stream of text or audio recordings. Story is broadly defined as segment of text containing topically related sentences. In particular, the task may require segmenting a stream of broadcast news, addressed by the Topic Detection and Tracking (TDT) evaluation project (Wayne, 2000; Allan, 2002). In this case topically related sentences belong to the same news story. While we are considering TDT data sets in this paper, we would like to pose the problem more broadly and consider a domain-independent approach to topic segmentation.

Previous research on topic segmentation has shown that lexical coherence is a reliable indicator of topical relatedness. Therefore, many approaches have concentrated on different ways of estimating lexical coherence of text segments, such as semantic similarity between words (Kozima, 1993), similarity between blocks of text (Hearst, 1994), and adaptive language models (Beeferman et al., 1999). These approaches use word repetitions to evaluate coherence. Since the sentences covering the same story represent a coherent discourse segment, they typically contain the same or related words. Repeated words build lexical chains that are consequently used to estimate lexical coherence. This can be done either by analyzing the number of overlapping lexical chains (Hearst, 1994) or by building a short-range and long-range language model (Beeferman et al., 1999). More recently, topic segmentation with lexical chains has been successfully applied to segmentation of news stories, multi-party conversation and audio recordings (Galley et al., 2003).

When the task is to segment long streams of text containing stories which may continue at a later point in time, for example developing news stories, building of lexical chains becomes intricate. In addition, the word repetitions do not account for synonymy and semantic relatedness between words and therefore may not be able to discover coherence of segments with little word overlap.

Our approach aims at discovering semantic relatedness beyond word repetition. It is based on the notion of semantic cohesion rather than lexical cohesion. We propose to use a similarity metric between segments of text that takes into account semantic associations between words spanning a number of segments. This method approximates lexical chains by averaging the similarity to a number of previous text



*Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 351–359, Prague, June 2007. 

segments and accounts for synonymy by using a hybrid document indexing scheme. Our text segmentation experiments show a significant performance improvement over the baseline.

The rest of the paper is organized as follows. Section 2 discusses hybrid indexing. Section 3 describes our segmentation algorithm. Section 5 reports the experimental results. We conclude in section 6.

## 2 Hybrid Document Indexing

For the topic segmentation task we would like to define a similarity measure that accounts for synonymy and semantic association between words. This similarity measure will be used to evaluate semantic cohesion between text units and the decrease in semantic cohesion will be used as an indicator of a story boundary. First, we develop a document representation which supports this similarity measure.

Capturing semantic relations between words in a document representation is difficult. Different approaches tried to overcome the term independence assumption of the bag-of-words representation (Salton and McGill, 1983) by using distributional term clusters (Slonim and Tishby, 2000) and expanding the document vectors with synonyms, see (Levow et al., 2005). Since content words can be combined into semantic classes there has been a considerable interest in low-dimensional representations. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is one of the best known dimensionality reduction algorithms. In the LSA space documents are indexed with latent semantic concepts. LSA maps all words to low dimensional vectors. However, the notion of semantic relatedness is defined differently for subsets of the vocabulary. In addition, the numerical information, abbreviations and the documents' style may be very good indicators of their topic. However, this information is no longer available after the dimensionality reduction.

We use a hybrid approach to document indexing to address these issues. We keep the notion of latent semantic concepts and also try to preserve the specifics of the document collection. Therefore, we divide the vocabulary into two sets: nouns and the rest of the vocabulary. The set of nouns does not include proper nouns. We use a method of spectral embedding, as described below and compute a low-dimensional representation for documents using only the nouns. We also compute a *tf-idf* representation for documents using the other set of words. Since we can treat each latent semantic concept in the low-dimensional representation as part of the vocabulary, we combine the two vector representations for each document by concatenating them.

### 2.1 Spectral Embedding

A vector space representation for documents and sentences is convenient and makes the similarity metrics such as cosine and distance readily available. However, those metrics will not work if they don't have a meaningful linguistic interpretation.

Spectral methods comprise a family of algorithms that embed terms and documents in a low-dimensional vector space. These methods use pair-wise relations between the data points encoded in a similarity matrix. The main step is to find an embedding for the data that preserves the original similarities.

**GLSA** We use Generalized Latent Semantic Analysis (GLSA) (Matveeva et al., 2005) to compute spectral embedding for nouns. GLSA computes term vectors and since we would like to use spectral embedding for nouns, it is well-suited for our approach. GLSA extends the ideas of LSA by defining different ways to obtain the similarities matrix and has been shown to outperform LSA on a number of applications (Matveeva and Levow, 2006).

GLSA begins with a matrix of pair-wise term similarities $S$, computes its eigenvectors $U$ and uses the first $k$ of them to represent terms and documents, for details see (Matveeva et al., 2005). The justification for this approach is the theorem by Eckart and Young (Golub and Reinsch, 1971) stating that inner product similarities between the term vectors based on the eigenvectors of $S$ represent the best element-wise approximation to the entries in $S$. In other words, the inner product similarity in the GLSA space preserves the semantic similarities in $S$.

Since our representation will try to preserve semantic similarities in $S$ it is important to have a matrix of similarities which is linguistically motivated.

| Word | Nearest Neighbors in GLSA Space | | | | | |
|---|---|---|---|---|---|---|
| witness | testify | prosecutor | trial | testimony | juror | eyewitness |
| finance | fund | bank | investment | economy | crisis | category |
| broadcast | television | TV | satellite | ABC | CBS | radio |
| hearing | hearing | judge | voice | chatter | sound | appeal |
| surprise | announcement | disappointment | stunning | shock | reaction | astonishment |
| rest | stay | remain | keep | leave | portion | economy |

Table 1: Words' nearest neighbors in the GLSA semantic space.

### 2.2 Distributional Term Similarity

**PMI** Following (Turney, 2001; Matveeva et al., 2005), we use point-wise mutual information (PMI) to compute the matrix $S$. PMI between random variables representing the words $w_i$ and $w_j$ is computed as

$$PMI(w_i, w_j) = \log \frac{P(W_i = 1, W_j = 1)}{P(W_i = 1)P(W_j = 1)}. \quad (1)$$

Thus, for GLSA, $S(w_i, w_j) = PMI(w_i, w_j)$.

**Co-occurrence Proximity** An advantage of PMI is the notion of proximity. The co-occurrence statistics for PMI are typically computed using a sliding window. Thus, PMI will be large only for words that co-occur within a small context of fixed size.

**Semantic Association vs. Synonymy** Although GLSA was successfully applied to synonymy induction (Matveeva et al., 2005), we would like to point out that the GLSA discovers semantic association in a broad sense. Table 1 shows a few words from the TDT2 corpus and their nearest neighbors in the GLSA space. We can see that for "witness", "finance" and "broadcast" words are grouped into corresponding semantic classes. The nearest neighbors for "hearing" and "stay" represent their different senses. Interestingly, even for the abstract noun "surprise" the nearest neighbors are meaningful.

### 2.3 Document Indexing

We have two sets of the vocabulary terms: a set of nouns, $N$, and the other words, $T$. We compute *tf-idf* document vectors indexed with the words in $T$:

$$\vec{d_i} = (\alpha_i(w_1), \alpha_i(w_2), ..., \alpha_i(w_{|T|})), \quad (2)$$

where $\alpha_i(w_t) = \mathrm{tf}(w_t, d_i) * \mathrm{idf}(w_t)$.

We also compute a $k$-dimensional representation with latent concepts $c_i$ as a weighted linear combination of GLSA term vectors $\vec{w_t}$:

$$\vec{d_i} = (c_1, ..., c_k) = \sum_{t=1:|N|} \alpha_i(w_t) * \vec{w_t}, \quad (3)$$

We concatenate these two representations to generate a hybrid indexing of documents:

$$\vec{d_i} = (\alpha_i(w_1), ..., \alpha_i(w_{|T|}), c_1, ...c_k) \quad (4)$$

In our experiments, we compute document and sentence representation using three indexing schemes: the *tf-idf* baseline, the GLSA representation and the hybrid indexing. The GLSA indexing computes term vectors for all vocabulary words; document and sentence vectors are generated as linear combinations of term vectors, as shown above.

### 2.4 Document similarity

One can define document similarity at different levels of semantic content. Documents can be similar because they discuss the same people or events and because they discuss related subjects and contain semantically related words. Hybrid Indexing allows us to combine both definitions of similarity. Each representation supports a different similarity measure. *tf-idf* uses term-matching, the GLSA representation uses semantic association in the latent semantic space computed for all words, and hybrid indexing uses a combination of both: term-matching for named entities and content words other than nouns combined with semantic association for nouns.

In the GLSA space, the inner product between document vectors contains all pair-wise inner product between their words, which allows one to detect semantic similarity beyond term matching:

$$\langle \vec{d_i}, \vec{d_j} \rangle = \sum_{w \in d_i} \sum_{v \in d_j} \alpha_i(w)\alpha_j(v)\langle \vec{w}, \vec{v} \rangle \quad (5)$$

If documents contain words which are different but semantically related, the inner product between the term vectors will contribute to the document similarity, as illustrated with an example in section 5.

When we compare two documents indexed with the hybrid indexing scheme, we compute a combination of similarity measures:

$$\langle \vec{d_i}, \vec{d_j} \rangle = \sum_{n_k \in d_i} \sum_{n_m \in d_j} \alpha_i(n_k)\alpha_j(n_m)\langle \vec{n_k}, \vec{n_m} \rangle + \sum_{t \in T} \alpha_i(t) * \alpha_j(t). \quad (6)$$

Document similarity contains semantic association between all pairs of nouns and uses term-matching for the rest of the vocabulary.

## 3 Topic Segmentation with Semantic Cohesion

Our approach to topic segmentation is based on semantic cohesion supported by the hybrid indexing. Topic segmentation approaches use either sentences (Galley et al., 2003) or blocks of words as text units (Hearst, 1994). We used both variants in our experiments. When using blocks, we computed blocks of a fixed size (typically 20 words) sliding over the documents in a fixed step size (10 or 5 words). The algorithm predicts a story boundary when the semantic cohesion between two consecutive units drops. Blocks can cross story boundaries, thus many predicted boundaries will be displaced with respect to the actual boundary.

**Averaged similarity** In our preliminary experiments we used the largest difference in score to predict story boundary, following the TextTiling approach (Hearst, 1994). We found, however, that in our document collection the word overlap between sentences was often not large and pair-wise similarity could drop to zero even for sentences within the same story, as will be illustrated below. We could not obtain satisfactory results with this approach.

Therefore, we used the average similarity by using a history of fixed size $n$. The semantic cohesion score was computed for the position between two text units, $t_i$ and $t_j$ as follows:

$$\mathrm{score}(t_i, t_j) = \frac{1}{n} \sum_{k=0}^{n-1} \langle t_{i-k}, t_j \rangle \quad (7)$$

Our approach predicts story boundaries at the minima of the semantic cohesion score.

**Approximating Lexical Chains** One of the motivations for our cohesion score is that it approximates lexical chains, as for example in (Galley et al., 2003). Galley et al. (Galley et al., 2003) define lexical chains $R_1, .., R_N$ by considering repetitions of terms $t_1, .., t_N$ and assigning larger weights to short and compact chains. Then the lexical cohesion score between two text units $t_i$ and $t_j$ is based on the number of chains that overlap both of them:

$$\mathrm{score}(t_i, t_j) = \sum_{k=1}^{N} w_k(t_i) w_k(t_j), \quad (8)$$

where $w_k(t_i) = \mathrm{score}(R_j)$ if the chain $R_j$ overlaps $t_i$ and zero otherwise. Our cohesion score takes into account only the chains for words that occur in $t_j$ and have another occurrence within $n$ previous sentences. Due to this simplification, we compute the score based on inner products. Once we make the transition to inner products, we can use hybrid indexing and compute semantic cohesion score beyond term repetition.

## 4 Related Approaches

We compare our approach to the LCseg algorithm which uses lexical chains to estimate topic boundaries (Galley et al., 2003). Hybrid indexing allows us to compute semantic cohesion score rather than the lexical cohesion score based on word repetitions.

Choi at al. used LSA for segmentation (Choi et al., 2001). LSA (Deerwester et al., 1990) is a special case of spectral embedding and Choi at al. (Choi et al., 2001) used all vocabulary words to compute low-dimensional document vectors. We use GLSA (Matveeva et al., 2005) because it computes term vectors as opposed to the dual document-term representation with LSA and uses a different matrix of pair-wise similarities. Furthermore, Choi at al. (Choi et al., 2001) used clustering to predict boundaries whereas we used the average similarity scores.

| | |
|---|---|
| s1: The **Cuban** news agency Prensa Latina called Clinton 's announcement Friday that Cubans picked up at sea will be taken to Guantanamo Bay naval base a " new and dangerous element " in U S immigration policy. | |
| s2: The **Cuban** government has not yet publicly reacted to Clinton 's announcement that Cuban rafters will be turned away from the United States and taken to the U S base on the southeast tip of Cuba. | |
| s5: The arrival of **Cuban** emigrants could be an " extraordinary aggravation " to the situation , Prensa Latina said. | |
| s6: It noted that **Cuba** had already denounced the use of the base as a camp for Haitian refugees. whom it had for many years encouraged to come to the United States. | |
| s8: **Cuba** considers the land at the naval base , leased to the United States at the turn of the century, to be illegally occupied. | |
| s10: **General Motors Corp** said Friday it was recalling 5,600 1993-94 model Chevrolet Lumina, Pontiac Trans Sport and Oldsmobile Silhouette minivans equipped with a power sliding door and built-in child seats. | |
| s14: If this occurs , the shoulder belt may not properly retract , the *carmaker* said. | |
| s15: **GM** is the only company to offer the power-sliding door. | |
| s16: The *company* said it was not aware of any accidents or injuries related to the defect. | |
| s17: To correct the problem , **GM** said dealers will install a modified interior trim piece that will reroute the seat belt. | |

Table 2: TDT. The first 17 sentences in the first file.

Existing approaches to hybrid indexing used different weights for proper nouns, nouns phrase heads and use WordNet synonyms to expand the documents, for example (Hatzivassiloglou et al., 2000; Hatzivassiloglou et al., 2001). Our approach does not require linguistic resources and learning the weights. The semantic associations between nouns are estimated using spectral embedding.

## 5 Experiments

### 5.1 Data

The first TDT collection is part of the LCseg toolkit[1] (Galley et al., 2003) and we used it to compare our approach to LCseg. We used the part of this collection with 50 files with 22 documents each.

We also used the TDT2 collection[2] of news articles from six news agencies in 1998. We used only 9,738 documents that are assigned to one topic and have length more than 50 words. We used the Lemur toolkit[3] with stemming and stop words list for the *tf-idf* indexing; we used Bikel's parser[4] to obtain the POS-tags and select nouns; we used the PLAPACK package (Bientinesi et al., 2003) to compute the eigenvalue decomposition.

[1] http://www1.cs.columbia.edu/ galley/tools.html

[2] http://nist.gov/speech/tests/tdt/tdt98/

[3] http://www.lemurproject.org/

[4] http://www.cis.upenn.edu/ dbikel/software.html

**Evaluation** For the TDT data we use the error metric $p_k$ (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002) which are implemented in the LCseg toolkit. We also used the TDT cost metric Cseg[5], with the default parameters P(seg)=0.3, Cmiss=1, Cfa=0.3 and distance of 50 words. All these measures look at two units (words or sentences) $N$ units apart and evaluate how well the algorithm can predict whether there is a boundary between them or not. Lower values mean better performance for all measures.

**Global vs. Local GLSA Similarity** To obtain the PMI values we used the TDT2 collection, denoted as $GLSA_{local}$. Since co-occurrence statistics based on larger collections give a better approximation to linguistic similarities, we also used 700,000 documents from the English GigaWord collection, denoted as GLSA. We used a window of size 8.

### 5.2 Topic Segmentation

The first set of experiments was designed to evaluate the advantage of the GLSA representation over the baseline. We compare our approach to the LCseg algorithm (Galley et al., 2003) and use sentences as segmentation unit. To avoid the issue of parameters setting when the number of boundaries is not known, we provide each algorithm with the actual numbers

[5] www.nist.gov/speech/tests/tdt/tdt98/doc/tdt2.eval.plan.98.v3.7.ps

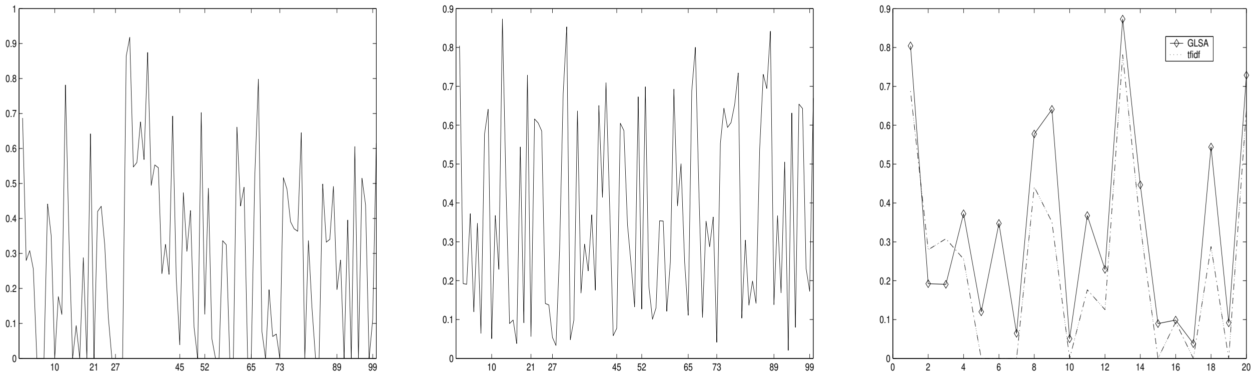


Figure 1: TDT. Pair-wise sentence similarities for *tf-idf* (left), GLSA (middle); x-axis shows story boundaries. Details for the first 20 sentences, table 2 (right).

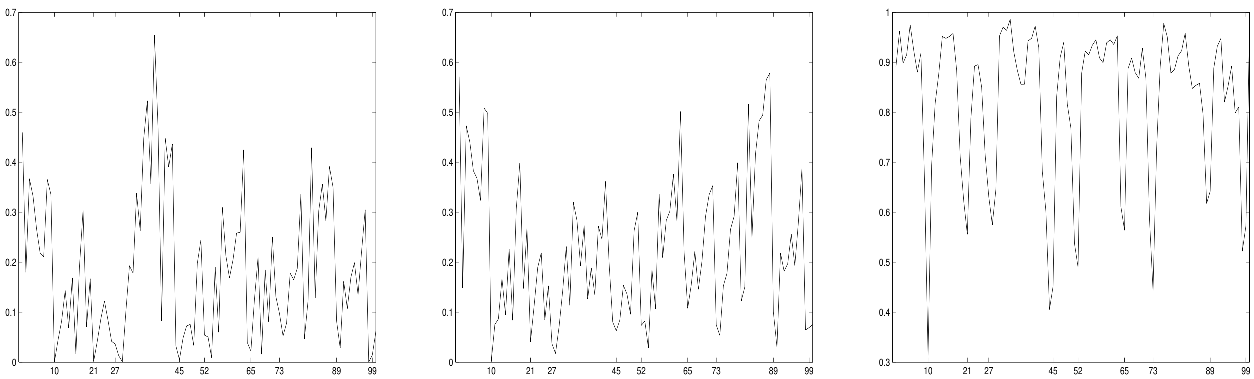


Figure 2: TDT. Pair-wise sentence similarities for *tf-idf* (left), GLSA (middle) averaged over 10 preceeding sentences; LCseg lexical cohesion scores (right). X-axis shows story boundaries.

of boundaries.

**TDT** We use the LCseg approach and our approach with the baseline *tf-idf* representation and the GLSA representation to segment this corpus. Table 2 shows a few sentences. Many content words are repeated, so the lexical chains is definitely a sound approach. As shown in Table 2, in the first story the word "Cuba" or "Cuban" is repeated in every sentence thus generating a lexical chain. On the topic boundary, the word overlap between sentences is very small. At the same time, the repetition of words may also be interrupted within a story: sentence 5, 6 and sentences 14, 15, 16 have little word overlap. LCseg deals with this by defining several parameters to control chain length and gaps. This simple example illustrates the potential benefit of semantic cohesion. Table 2 shows that "General Motors" or "GM" are not repeated in every sentence of the second story. However, "GM", "carmaker" and "company" are semantically related. Making this information available to the segmentation algorithm allows it to establish a connection between each sentence of the second story.

We computed pair-wise sentence similarities between pairs of consecutive sentences in the *tf-idf* and GLSA representations. Figure 1 shows the similarity values plotted for each sentence break. The pairwise similarities based on term-matching are very spiky and there are many zeros within the story. The GLSA-based similarity makes the dips in the similarities at the boundaries more prominent. The last plot gives the details for the sentences in table 2. In the *tf-idf* representation sentences without word overlap receive zero similarity but the GLSA representation is able to use the semantic association between between "emigrants" and "refugees" for sentences 5 and 6, and also the semantic association between "carmaker" and "company" for sentences 14

| Measure | *tf-idf* | GLSA | LCseg |
|---|---|---|---|
| Pmiss | 0.29 | 0.19 | N/A |
| Pfa | 0.14 | 0.09 | N/A |
| Cseg | 0.18 | 0.08 | N/A |
| $p_k$ | 0.24 | 0.17 | 0.07 |
| wd | 0.27 | 0.21 | 0.10 |

Table 3: TDT segmentation results.

and 15.

This effect increases as we use the semantic cohesion score as in equation 7. Figure 2 shows the similarity values for *tf-idf* and GLSA and also the lexical cohesion scores computed by LCseg. The GLSA-based similarities are not quite as smooth as the LCseg scores, but they correctly discover the boundaries. LCseg parameters are fine-tuned for this document collection. We used a general TDT2 GLSA representation for this collection, and the only segmentation parameter we used is to avoid placing next boundary within $n$=3 sentences of the previous one. For this reason the predicted boundary may be one sentence off the actual boundary. These results are summarized in Table 3. The GLSA representation performs significantly better than the *tf-idf* baseline. Its $p_k$ and WindowDiff scores with default parameters for LCseg are worse than for LCseg. We attribute it to the fact that we did not fine-tuned our method to this collection and that boundaries are often placed one position off the actual boundary.

**TDT2** For this collection we used three different indexing schemes: the *tf-idf* baseline, the GLSA representation and the hybrid indexing. Each representation supports a different similarity measure. Our TDT experiments showed that the semantic cohesion score based on the GLSA representation improves the segmentation results. The variant of the TDT corpus we used is rather small and well-balanced, see (Galley et al., 2003) for details. In the second phase of experiments we evaluate our approach on the larger TDT2 corpus. The experiments were designed to address the following issues:

- performance comparison between GLSA and Hybrid indexing representations. As mentioned before, GLSA embeds all words in a low-dimensional space. Whereas semantic classes for nouns have theoretical linguistic justification, it is harder to motivate a latent space representation for example for proper nouns. Therefore, we want to evaluate the advantage of using spectral embedding only for nouns.

- collection dependence of similarities. The similarity matrix $S$ is computed using the TDT2 corpus ($GLSA_{local}$) and using the larger GigaWord corpus. The larger corpus provides more reliable co-occurrence statistics. On the other hand, word distribution is different from that in the TDT2 corpus. We wanted to evaluate whether semantic similarities are collection independent.

| #b known | | | |
|---|---|---|---|
| Method | Pmiss | Pfa | Cseg |
| *tf-idf* | 0.52 | 0.14 | 0.19 |
| GLSA | 0.4 | 0.1 | 0.14 |
| GLSA_$local$ | 0.44 | 0.12 | 0.16 |
| Hybrid | 0.34 | 0.10 | 0.12 |
| Hybrid_$local$ | 0.38 | 0.09 | 0.13 |
| LCseg | 0.80 | 0.19 | 0.28 |
| #b unknown | | | |
| Method | Pmiss | Pfa | Cseg |
| *tf-idf* | 0.42 | 0.2 | 0.17 |
| GLSA | 0.37 | 0.13 | 0.14 |
| GLSA_$local$ | 0.35 | 0.19 | 0.14 |
| Hybrid | 0.26 | 0.16 | 0.11 |
| Hybrid_$local$ | 0.27 | 0.18 | 0.12 |

Table 4: TDT2 segmentation results. Sliding blocks with size 20 and stepsize 10; similarity averaged over 10 preceeding blocks.

Table 4 shows the performance evaluation. We show the results computed using blocks containing 20 words (after preprocessing) with step size 10. We tried other parameter values but did not achieve better performance, which is consistent with other research (Hearst, 1994; Galley et al., 2003). We show the results for two settings: predict a known number of boundaries, and predict boundaries using a threshold. In our experiments we used the average of the smallest $N$ scores as threshold, $N = 4000$ showing best results.

The spectral embedding based representations (GLSA, Hybrid) significantly outperform the baseline. This confirms the advantage of the semantic cohesion score vs. term-matching. Hybrid indexing outperforms the GLSA representation supporting our intuition that semantic association is best defined for nouns.

We used the GigaWord corpus to obtain the pair-wise word associations for the GLSA and Hybrid representations. We also computed $\text{GLSA}_{local}$ and $\text{Hybrid}_{local}$ using the TDT2 corpus to obtain the pair-wise word associations. The co-occurrence statistics based on the GigaWord corpus provide more reliable estimations of semantic association despite the difference in term distribution. The difference is larger for the GLSA case when we compute the embedding for all words, GLSA performs better than $\text{GLSA}_{local}$. $\text{Hybrid}_{local}$ performs only slightly worse than Hybrid. This seems to support the claim that semantic associations between nouns are largely collection independent. On the other hand, semantic associations for proper names are collection dependent at least because the collections are static but the semantic relations of proper names may change over time. The semantic space for a name of a president, for example, is different for the period of time of his presidency and for the time before and after that.

Disappointingly, we could not achieve good results with LCseg. It tends to split stories into short paragraphs. Hybrid indexing could achieve results comparable to state-of-the art approaches, see (Fiscus et al., 1998) for an overview.

## 6 Conclusion and Future Work

We presented a topic segmentation approach based on semantic cohesion scores. Our approach is domain independent, does not require training or use of lexical resources. The scores are computed based on the hybrid document indexing which uses spectral embedding in the space of latent concepts for nouns and keeps proper nouns and other specifics of the documents collections unchanged. We approximate the lexical chains approach by simplifying the definition of a chain which allows us to use inner products as basis for the similarity score. The similarity score takes into account semantic relations between nouns beyond term matching. This semantic cohesion approach showed good results on the topic segmentation task.

We intend to extend the hybrid indexing approach by considering more vocabulary subsets. Syntactic similarity is more appropriate for verbs, for example, than co-occurrence. As a next step, we intend to embed verbs using syntactic similarity. It would also be interesting to use lexical chains for proper names and learn the weights for different similarity scores.

## References


J. Allan, editor. 2002. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers.

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Paolo Bientinesi, Inderjit S. Dhilon, and Robert A. van de Geijn. 2003. A parallel eigensolver for dense symmetric matrices based on multiple relatively robust representations. *UT CS Technical Report TR-03-26*.

Freddy Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for text segmentation. In *Proceedings of EMNLP*, pages 109–117.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

J. G. Fiscus, George Doddington, John S. Garofolo, and Alvin Martin. 1998. NIST's 1998 topic detection and tracking evaluation (tdt2). In *Proceedings of NIST's 1998 Topic Detection and Tracking Evaluation*.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*.

G. Golub and C. Reinsch. 1971. *Handbook for Matrix Computation II, Linear Algebra*. Springer-Verlag, New York.

V. Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of SIGIR*, pages 224–231.

V. Hatzivassiloglou, Regina Barzilay Min-Yen Kan Judith L. Klavans, Melissa L. Holcombe, and Kathleen R. McKeown. 2001. Simfinder: A flexible clustering tool for summarization. In *Proceedings of NAACL*, pages 41–49.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of ACL*, pages 9–16.

Hideki Kozima. 1993. Text segmentation based on similarity between words. In *Proceedings of ACL*, pages 286–288.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: Special Issue on Cross-language Information Retrieval*.

Irina Matveeva and Gina-Anne Levow. 2006. Graph-based Generalized Latent Semantic Analysis for document representation. In *Proc. of the TextGraphs Workshop at HLT/NAACL*.

Irina Matveeva, Gina-Anne Levow, Ayman Farahat, and Christian Royer. 2005. Generalized Latent Semantic Analysis for term representation. In *Proc. of RANLP*.

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

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Noam Slonim and Naftali Tishby. 2000. Document clustering using word clusters via the information bottleneck method. In *Research and Development in Information Retrieval*, pages 208–215.

Peter D. Turney. 2001. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167:491–502.

C. Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 1487–1494.