TV News Story Segmentation based on Semantic Coherence and Content Similarity

Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha, and Joemon M. Jose

Dept. of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK
{hemant,hopfgarf,anuj,punitha,jj}@dcs.gla.ac.uk

Abstract. In this paper, we introduce and evaluate two novel approaches, one using video stream and the other using close-caption text stream, for segmenting TV news into stories. The segmentation of the video stream into stories is achieved by detecting anchor person shots and the text stream is segmented into stories using a Latent Dirichlet Allocation (LDA) based approach. The benefit of the proposed LDA based approach is that along with the story segmentation it also provides the topic distribution associated with each segment. We evaluated our techniques on the TRECVid 2003 benchmark database and found that though the individual systems give comparable results, a combination of the outputs of the two systems gives a significant improvement over the performance of the individual systems.

1 Introduction

In most part of the 20th century, consuming news was a solely passive activity. People simply followed news coverage by reading newspapers, listening to radio broadcasts or watching the television news. However, the rise of new technologies has rapidly changed this trend; now-a-days publishers and broadcasters also provide content on the WWW, an increasing percentage of this being video clips [1]. Faced with these developments, processing video clips, television news being one among them, has become an important research area that has attracted a lot of attention. The main focus is to tackle the problems that arise when it is required to retrieve some information from this data. In this context, a basic challenge is to segment videoes into meaningful and manageable *segments* in order to ease the access of the video data. The smallest coherent segment in a video is a shot, a unit that has been constantly filmed using the same camera setting. A simple solution to video segmentation is to divide a video into shots using visual features such as colour, texture and shape. State-of-the-art techniques as evaluated within TRECVid [19] reach a very high performance in detecting shot boundaries. Nevertheless, a more challenging, and also more informative approach, is to segment broadcasts into coherent news stories. Segmenting a news broadcast into such stories is essentially finding the boundaries where one story ends and the other begins.

In this paper, we approach the TV news story segmentation task from lexical content and visual similarity perspectives. Segmenting the teletext stream of a television news into stories is a direct application of text segmentation, an active area of research [12, 20].

We evaluate the performance of our approaches on the TRECVid 2003 data collection [18], a standard benchmark used for the story segmentation task. The corpus consists of over 130 hours of news video in MPEG-1 format that was broadcast in the year 1998. The collection has been split into a test set and a development set. In the current work, we use the test set, which enables us to compare our results with the runs submitted to TRECVid. The test set was split into more than 32000 shots with representative key frames provided for each shot. Moreover, each broadcast was manually split into coherent story segments and the corresponding transcripts were provided. In Section 2, we provide an overview of state-of-the-art story segmentation approaches. In Section 3, we birefly explain our LDA based approach for the task of text segmentation the details of which can be found in [14]. Segmenting the text transcripts of the news broadcast using LDA based approach not only provides the story boundaries but also the topic distribution associated with each story. In Section 4, we introduce our feature-based approach for video segmentation where we extract colour features from each key frame and identify anchor person shots. Neighbouring shots from these anchor persons are merged based on their similarity with respect to shot length difference and visual dissimilarity. Using the resulting time points of the detected boundary key frames, we segment the video broadcast into stories. The performance of both the approaches and their combination is evaluated in Section 5. In Section 6 we draw the main conclusons of this study and outline the future directions.

2 Background

Segmenting TV news broadcasts into story units was one of the main tasks within TRECVid 2003 and 2004 evaluations. The task description of these evaluations defines stories as "segments of a news broadcast with a coherent news focus which contain at least two independent declarative clauses". Various approaches using text, audio and video streams or a combination of them have been studied to segment TV news broadcast into stories.

In text segmentation, some approaches rely on word repetition [12] while the others use cue phrases [16] to identify story boundaries. The later approaches use the information that transcript of a TV news broadcast is typically laced with cues words such as *welcome, bye, good morning, thank you, next to follow* etc., to indicate the beginning or end of a story.

O'Connor et al. [15] performed story segmentation by clustering key frames based on their low-level colour feature. In their approach, two shots that are very similar based on their visual appearance but have been shown at two distant moments during the broadcast will not be placed in the same cluster.

Due to the feature-rich nature of TV news broadcast, it is not premature to assume that a broadcasts' video and text streams may contain complementary information and that their combination can yield a performance that is better than the performance of a system which only uses the information from a single stream. Indeed, analyses [2, 6] have shown that the most successful runs evaluated within TRECVid rely on both text-based and visual-based segmentation approaches to detect story boundaries. Pickering et al. [17], for instance, extracted key entities such as nouns and verbs from the broadcast transcript, computed a term weighting based on their frequency within the text and combined neighbourig shots to accomplish the task.

Hsu et al. [10] perform a story boundary segmentation experiment and compare the average precision of different combinations of audio, video and text fusions. They report that a combination of all modalities worked best to identify correct story boundaries. However, as Chang et al. [5] argue, a better understanding of relations between information extracted from the text stream and relations extracted from different audio and visual streams is still needed. Chaisorn et al. [4] approach this problem using a bifid approach. First, they employ a learning based approach to identify story boundaries, and then classify each story into semantic categories by employing heuristic rules.

Different from all these approaches, our LDA based approach not only estimates the segment boundaries, it also categorizes the segments based on their topic distribution. Moreover, the only assmption in feature based approach is that a story always begins with an anchor person shot.

3 Text-based Segmentation

In this section, we briefly describe our recently proposed topic model based approach for story segmentation task [14] which exploits the properties of unsupervised Latent Dirichlet Allocation (LDA) [3, 7] topic model to estimate the coherence of a segment, and in turn the segment boundaries. The details of our approach and its analysis can be found in [14]

LDA is a generative unsupervised approach to model discrete data such as text. The two main assumptions in LDA are: 1) every document is represented by a topic distribution, and 2) every topic has an underlying word distribution.

In this work, we have used Gibbs sampling method, as decribed in [7], to train the LDA model on the well known Reuters collection volume 1 (RCV1). The training consists of estimating the topic distribution in each training document, represented by θ, and word distribution in each topic, represented by ϕ. After the burn-in period of the Gibbs sampling, these two parameters are estimated by the following equations:

$$\theta_{dt} = \frac{K_{dt} + \alpha}{\sum_{k=1}^{T} K_{dk} + T\alpha} \tag{1}$$

$$\phi_{tv} = \frac{J_{tv} + \beta}{\sum_{k=1}^{V} J_{tk} + V\beta} \tag{2}$$

where K_{dt} is the number of times a word in document d has been assigned to topic t, J_{tw} is the number of times word w has been assigned to topic t in the whole training corpus and V is number of unique words in the training corpus (vocabulary size) after removing stop-words; number of topics, T, and Dirichlet priors, α and β, are hyper-parameters, and in our experiments their values were 50, 1 and 0.01, respectively.

During testing, the topic distribution of an unseen document can be estimated by the following iterative equation [8, 13]:

$$\theta_{dt}^{(n+1)} = \frac{1}{l_d} \sum_{v=1}^{V} \frac{C_{dv}\theta_{dt}^{(n)}\phi_{tv}}{\sum_{t'=1}^{T}\theta_{dt'}^{(n)}\phi_{t'v}} \tag{3}$$

where $\theta_{dt}^{(n)}$ is the value of θ_{dt} at nth iteration, C_{dv} is the number of times vocabulary word v has occured in document d, and l_d is the number of words in the document which are present in the training vocabulary. The words in the document which are not in the training vocabulary are dropped, and are not used for estimating the topic distribution.

The likelihood of a document, given its topic distribution, can be estimated as

$$P(C_d|\theta, \phi) = \prod_{v=1}^{V} \left[\sum_{t=1}^{T} \theta_{dt}\phi_{tv} \right]^{C_{dv}} \tag{4}$$

In this paper, the same methodology which is used to compute the likelihood of an unseen document is applied to compute the likelihood of a segment.

For a given text, a coherent segment containing a single story is expected to have only a few active topics (LDA topics as defined in the LDA framework), whereas an incoherent segment, having more than one story in it, may have several active topics. In [13], the authors showed that likelihood of a coherent document is higher as compared to the likelihood of an incoherent document. This observation is the fundamental premise for our LDA based approach: for a given text, the segmentation which provides the highest likelihood is also going to provide the most coherent segments. The task of finding the highest likelihood, and in turn the most coherent segments, is performed in the framework of dynamic programming (DP).

Lets assume a given text $d = \{w_1 \cdots w_{l_d}\}$ of length l_d. For this text, consider a particular segmentation, S, which is made of m segments, $S = \{S_1 \cdots S_m\}$, where S_i has n_i words in it. Further, let w_i^j be the jth word token in S_i, such that $W_i = \{w_i^1 \cdots w_i^{n_i}\}$. Therefore, $\sum_{i=1}^{m} n_i = l_d$, $d = \{W_1 \cdots W_m\}$ and W_i is dependent only on S_i. The likelihood of segment S can be given by

$$P(S|d) = P(d|S)P(S)/P(d) \tag{5}$$

where $P(d|S)$ is the probability of the document d under segmentation S and P(S), considered as a penalty factor, is a prior over segmentations. $P(d)$ is same for all the possible segmentations of a documents and hence can be dropped.

Therefore

$$P(S|d) \propto \left[\prod_{i=1}^{m} P(W_i|S)\right] P(S) \propto \left[\prod_{i=1}^{m} P(W_i|S_i)\right] P(S) \propto \left[\prod_{i=1}^{m}\prod_{j=1}^{n_i} P(w_i^j|S_i)\right] P(S)$$

The optimal segmentation is the one that maximises this likelihood, that is, $\hat{S} = \underset{S}{\operatorname{argmax}}\ P(S|d)$, and can be obtained by DP which is typically employed to solve the problem of shortest path in many applications. The likelihood of a segment, $P(W_i|S_i)$, is obtained by (4), where the term C_{dv} is replaced by the word frequency occurence in a segment. That is, for each possible segment, (3) is used to compute its θ and subsequently (4) is employed to estimate its likelihood.

A DP algorithm has two passes, a forward-pass followed by a trace-back. In the forward-pass of our DP, for each segment described by a begin word (B) and an end word (E), the likelihood is computed by (4). This likelihood is accummulated and for each E node, the information about the B node which gives the highest score (in orther words, the B node which is the best starting node for this E node) is stored. On reaching the document end, during trace back, the information about the best starting node is used to get segmentation (segment boundaries) which gives the maximum-likelihood path. In our case, $P(S) = (l_d)^{-m*p}$, where $p = 3$ was empirically found to give the best results on another dataset.

4 Feature-based Segmentation

In this section, we focus on exploiting various content features to segment news broadcasts into corresponding story segments. In most new broadcasts, e.g. from CNN, Al Jazeera or BBC, stories are often introduced by an anchor person. This also applies to the TRECVid 2003 corpus. The first step in our feature-based story segmentation approach is therefore to identify the first anchor person shot in the video. An analysis revealed that the first anchor person shot usually appears within the initial 25–55 seconds of each broadcast. Since anchor persons are usually filmed in a studio setting with similar visual appearance in each broadcast, identifying these shots is a pattern matching task. Utilising this observation, we first identify the first possible anchor person key frame. In the beginning, we consider each shot in the first 25–55 seconds of video to be the possible anchor person shot candidate. We hence need to identify the key frames which appear more often than any other key frame in the broadcast. We start by computing the visual distance of the MPEG-7 colour structure feature between every candidate within this range and the remaining key frames of the broadcast. Since some shots might be re-appearing shots belonging to the same story, we skip a few shots Δk which may be repeated shots in the neighbourhood of the anchor person shot. The candidate frame with the lowest average visual similarity is considered to be the first anchor person key frame.

The next task is to identify other anchor person shots within the video. In order to classify a shot as an anchor person shot, we also take the neighbouring three shots on both sides, a region of support, into account. This region of support is used to determine whether the anchor person introduces a new story or not. Accordingly, a shot will be treated as story boundary candidate only if the neighbouring shots differ significantly from each other. Unfortunately, no ground truth data exist which can be used to evaluate our anchor person detection approach. Therefore, our evaluation is focused on the actual story boundary detection task, which we treat as a classification task. Twenty sample videos from both CNN and ABC videos of the TRECVid 2003 corpus are used to train an SVM for each collection. Ground truth provided within TRECVid is used to identify true story boundaries in training samples. The following features are used to train an SVM to identify anchor person shots:

- **Distance from Anchor Person Template:** We compute the visual distance between the previously identified template and the current key frame using the MPEG-7 Colour Feature.
- **Semantic Text Similarity:** Following Kolb [11], we compute the semantic similarity between the transcript of the left region of support and the right region of support. We assume that the transcript is similar on both sides if both transcripts form part of the same story.
- **Shot Length Distance:** We compute the absolute difference between the numbers of key frames in the left and the right region of the support. Action-loaded news like sports reports are expected to have more key frames than calmer news, e.g. reports about political party agendas. Therefore, it is an effective feature to distinguish between stories.
- **Average Visual Dissimilarity:** We determine the average difference of the MPEG-7 colour structure feature between the shots from the left and the right region of support. This value can identify the shots which are visually similar to the neighbourhood and so are very less probable to start a new story.
- **Minimum Visual Dissimilarity:** We compute the minimum difference between the shots from the left and right region of support using the colour structure feature. This value is useful to detect when a shot is repeated in a news story, as the minimum distance will be very low in this case.

Despite the assumption that any story starts with the anchor person, it is not always true that a story ends with the appearance of the next anchor person. Hsu et al. [9] argue that within an anchor person shot, there can be a possible presence of a story boundary, as the anchor person continues with the previous story and changes to the new story only towards the mid of the shot. It could also happen that an anchor person introduces stories without any supporting video clips. This gives rise to possible, intra-shot story boundaries. Hence, it is required to split and merge anchor person shots accordingly.

In order to detect such boundaries, we first extract two frames per second of all anchor person shots. As shown in Figure 1, we first split each frame into four regions, with R_1 and R_2 being the first and second quadrant, respectively. We assume that in these two quadrants, anchor person shots will contain the face

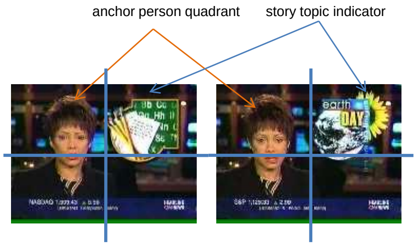

Fig. 1. Example of an intra-shot story boundary

of the anchor person and a graphic or video indicating the topic of the actual story. Consequently, the visual appearance of the anchor person quadrant will be similar over all frames of the anchor person shot, while the visual appearance of the other quadrant will change whenever a new story begins. Therefore, we determine the eigen difference E_1 and E_2 for both quadrants. If either E_1 or E_2 is under a predefined threshold while the other value is above a threshold, we define this frame as a story boundary.

5 Results and analysis

5.1 Boundary Detection Task

Following the TRECVid guidelines, we evaluate the segmentation performance of both approaches using the precision P_{seg} and recall R_{seg} metrics as defined by (6) and (7). Moreover, we compute the F_1 values using both metrics.

$$P_{seg} = \frac{|\text{determined boundaries}| - |\text{wrong boundaries}|}{|\text{determined boundaries}|} \tag{6}$$

$$R_{seg} = \frac{|\text{detected reference boundaries}|}{|\text{reference boundaries}|} \tag{7}$$

As outlined by Hsu et al. [9], boundaries are correctly detected when a determined boundary lies within five seconds of an actual reference story boundary. Otherwise, the boundary is considered to be wrong. Table 1 shows the independent metrics for both ABC and CNN videos as well as for the combination of both datasets. In the remainder of this section, we will denote these metrics as “baseline” results. As can be seen, the overall performance of both approaches for both datasets is similar.

The main weakness of the feature-based approach seems to be the actual detection of anchor person shots. Whenever an anchor person shot has been missed, a potential story boundary will be ignored, hence resulting in a drop in precision and recall. Moreover, stories that do not start with an anchor person shot will be missed as well, which is a drawback of our feature based approach.

Table 1. Precision, Recall and F_1 measures for both approaches

	CNN			ABC			ABC & CNN		
	R_{seg}	P_{seg}	F_1	R_{seg}	P_{seg}	F_1	R_{seg}	P_{seg}	F_1
Feature-based	0.33	0.69	0.44	0.27	0.69	0.38	0.30	0.70	0.41
LDA	0.30	0.70	0.42	0.32	0.52	0.40	0.31	0.62	0.41
LDA (adapted)	0.31	0.71	0.43	0.38	0.58	0.45	0.34	0.65	0.44

The potential drawback of the LDA approach is that if the test data is from a different domain and as a consequence there is a vocabulary mismatch, those words which did not appear during training will be dropped from the estimations. Therefore, a percentage of content words is lost. To alleviate this problem of vocabulary mismatch between Reuters data used for LDA training and the TRECVid transcripts used for evaluation, we propose to train the LDA model with combined Reuters and TRECVid development data. The results of this LDA adaptation is shown in the last row of Table 1. As can be seen, the performance of the LDA method improves, suggesting that a bigger in-domain adaptation data may have improved the performance even further. A quick analysis of the segmented output reveals that on most occasions the boundaries are estimated correctly or missed by a sentence or two. It is observed that the short segments are typically missed and it is because LDA requires some minimum amount of data for reliable estimation. For two example broadcasts, Figures 2 and 3 show the boundaries, in terms of word number in the transcript, identified by both approaches, as well as the actual boundaries. Figure 2 reveals that most of the time, the boundaries in the ABC broadcast which are identified by both approaches are correct. In the CNN broadcast shown in Figure 3, however, various boundaries have been missed. The reason for this miss is that CNN stories are rather short which is a problem for our text based approach. Both figures illustrate that the two approaches do not identify the same boundaries all the time. This complementarity can be exploited by combining the results of both approaches. It supports the general assumption [6] that a combination of differ-

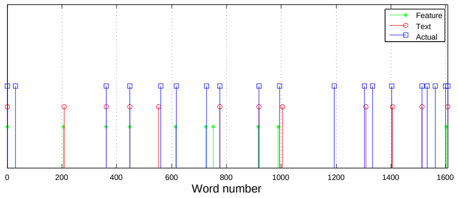

Fig. 2. Performance of both approaches in detecting story boundaries (ABC footage)

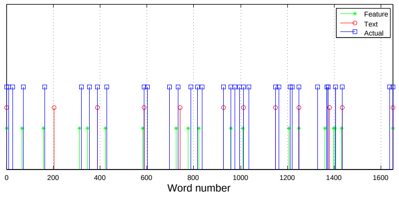

Fig. 3. Performance of both approaches in detecting story boundaries (CNN footage)

ent modalities, text and visual features in our case, can improve the accuracy of story segmentation approaches. Therefore, we fuse detected boundaries from both approaches using the “or” operator. Boundaries from both approaches that are within a one second time window distance from each other are merged to form one single boundary. This buffer will reduce the number of false positives. As Table 2 reveals, this fusion results in a huge improvement in both recall and F_1 measures in comparison to the baseline results shown in Table 1. Precision goes down slightly, indicating that the relative number of wrong boundaries has marginally increased.

Table 2. Precision, Recall and F_1 measures

	CNN			ABC			ABC & CNN		
Combination	R_{seg}	P_{seg}	F_1	R_{seg}	P_{seg}	F_1	R_{seg}	P_{seg}	F_1
Feature + LDA	0.51	0.67	0.58	0.52	0.57	0.54	0.52	0.62	0.56
Feature + LDA (adapted)	0.52	0.67	0.58	0.56	0.60	0.58	0.54	0.64	0.58

In comparison with state-of-the-art approaches evaluated within TRECVid 2003, our simple approach ranks in the upper field of all submissions. In addition to the other approaches, however, our LDA based method can also be exploited to categorise detected stories. This categorisation is shown in the following section.

5.2 Story Categorisation Task

All the results previously published in the literature typically concentrated on the segmentation performance (either some error metric or time complexity). Though estimating the segment boundaries is important, if the segments can be identified by a topic (or topic distribution), this information can have profound impact in several other applications such as discourse analysis and information retrieval. LDA being a topic model is in a position to output this information along with the segment boundaries. In this section, we show an example output of the text segmentation phase. To save space, long sentences were terminated by “...” to show continuation beyond the printed words.

“*the holy grail of hiv research is to develop a safe and ... when you have an epidemic like this the way to put an ... a few potential vaccines are in human trials but final results are ... this year s conference represents a major change in emphasis and mood it is somber because hiv continues to be such an elusive foe george strait abc news geneva* **ESTIMATED BOUNDARY is CORRECT: TOPIC 43 has highest probability (0.39)** *now for news back home there is a new face in the ... she s a friend of the lewinsky family and she is telling ... she has testified before kenneth starr s grand jury she has also given an interview to newsweek magazine here is abc s karla davis abc news has confirmed that dale young a forty seven year old ... it is just the most unfortunate sense of timing* *tomorrow in another washington courtroom team clinton will argue presidential adviser bruce ... the one person not scheduled to be in court is monica lewinsky she and her new legal team still have not reached a deal ... karla davis abc news washington*" **ESTIMATED BOUNDARY is CORRECT: TOPIC 28 has highest probability (0.47)**

The top 10 words of TOPIC 43 and TOPIC 28, obtained after LDA training, are printed below for reference:
TOPIC 43: *'health' 'medical' 'mother' 'hospital' 'people' 'church' 'drug' 'heart' 'doctors' 'disease'*
TOPIC 28: *'pay' 'lead' 'type' 'sep' 'today' 'investigation' 'evidence' 'trial' 'case' 'lewinsky'*

From this example, we notice that the top topic associated with each segment is mostly relevant to the words present in that segment.

6 Conclusions

In this paper, we investigated and compared two approaches for segmenting TV news broadcast into stories: an LDA based method for text segmentation and a low-level feature-based approach for video segmentation. LDA has been previously demonstrated as an approach comprabale to the state-of-the-art approaches for the task of text segmentation [14]; it also outputs the topic distribution of segments. An analysis of the identified story boundaries revealed the complementarity of both the approaches, suggesting that they can be combined to form a more precise segmentation. Indeed, a simple fusion using an "or" operator already leads to significant improvement in performance. With respect to precision and recall, these results are above average in comparison with systems evaluated within TRECVid, outperformed by a few approaches only. While these best performing approaches are tailored to pre-defined rules, e.g., the appearance of cue phrases in the transcript, we base our approach on one assumption only, that is stories always start with an anchor person shot. Our approach is therefore a more general solution to tackle the television news segmentation task. Unlike other approaches, the proposed method computes topic distributions jointly with segmentation, thus allowing to collect information about the thematic content of each segment. This information can be used to keep track of recurring topics. In future work, we aim at including other multimedia domains, such as the audio layer of the news broadcast since TRECVid results support the effectiveness of considering this domain in a segmentation task.

Acknowledgments

This research was supported by the EC under contract FP6-027122-SALERO.

References

1. D. Ahlers. News Consumption and the New Electronic Media. *The Harvard International Journal of Press/Politics*, 11(1):29–52, 2006.
2. J. Arlandis, P. Over, and W. Kraaij. Boundary error analysis and categorization in the TRECVID news story segmentation task. In *CIVR'05*, pages 103–112. 2005.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 14, pages 601–608, Cambridge, MA, 2002. MIT Press.
4. L. Chaisorn, T.-S. Chua, C.-H. Lee, and Q. Tian. A hierarchical approach to story segmentation of large broadcast news video corpus. In *ICME'04*, pages 1095–1098. 2004.
5. S.-F. Chang, R. Manmatha, and T.-S. Chua. Combining Text and Audio-Visual Features in video Indexing. In *ICASSP'05 – Proceedings of Acoustics, Speech, and Signal Processing Conference*, pages 1005–1008, 03 2005.
6. T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu. Story boundary detection in large broadcast news video archives: techniques, experience and trends. In *MM'04*, pages 656–659. ACM, 2004.
7. T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (supl 1):5228–5235, 2004.
8. A. Heidel, H. an Chang, and L. shan Lee. Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm. In *Proceedings of EuroSpeech*, Antwerp, Belgium, 2007.
9. S.-F. Hsu, Winston H. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar. Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation. In *IS&T/SPIE Electronic Imaging, San Jose, CA*, 2004.
10. W. H. Hsu, L. S. Kennedy, S.-F. Chang, M. Franz, and J. R. Smith. Columbia-IBM News Video Story Segmentation in TRECVID 2004. In *TREC*, 2004.
11. P. Kolb. DISCO: A Multilingual Database of Distributionally Similar Words. In *KONVENS 2008*, 2008.
12. H. Kozima. Text segmentation based on similarity between words. In *Meeting of the Association for Computational Linguistics*, pages 286–288, Ohio, U.S.A., 1993.
13. H. Misra, O. Cappé, and F. Yvon. Using LDA to detect semantically incoherent documents. In *Proceedings of CoNLL*, pages 41–48, Manchester, U.K., 2008.
14. H. Misra, F. Yvon, J. M. Jose, and O. Cappé. Text segmentation via topic modeling: An analytical study. In *Proceedings of CIKM*, Hong Kong, China, 2009.
15. N. O'Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy, and A. Smeaton. News Story Segmentation in the Físchlár Video Indexing System. In *ICIP'01*, 2001.
16. R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Comput. Linguist.*, 23(1):103–139, 1997.
17. M. J. Pickering, L. Wong, and S. Rüger. ANSES: Summarisation of news video. *Image and Video Retrieval*, 2788:481–486, 2003.
18. A. F. Smeaton, W. Kraaij, and P. Over. TRECVID 2003 – An Overview. In *TRECVid 2003*, 2003.
19. A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06*, pages 321–330, New York, NY, USA, 2006. ACM Press.
20. M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Meeting of the Association for Computational Linguistics*, pages 491–498, Bergen, Norway, 2001.