Multimodal News Story Segmentation

Gert-Jan Poulisse[1], Marie-Francine Moens[1]

[1] Katholieke Universiteit Leuven, Department of Computer Science, Celestijnenlaan 200A
Box 2402
B-3001 Heverlee, Belgium
{Gert-Jan.Poulisse, Marie-Francine.Moens}@cs.kuleuven.be

Abstract. In this paper, we describe a multi-modal approach to segmenting news video based on the perceived shift in content. We divide up a video document into logically coherent semantic units known as stories. We investigate the effectiveness of a number of multimedia features which serve as potential indicators of a story boundary. The results show an improvement of performance over current state of the art story segmenters.

Keywords: Video segmentation, feature extraction, story detection.

1 Introduction

The aim of our research is to implement accurate methods for story segmentation in news video. In this context, this means detecting the specific time event at which one news story stops being discussed and a new story starts being reported. In text, a story is a coherent grouping of sentences, discussing related topics and names. The multimedia equivalent, such as found in news video, would be a temporal segment containing imagery accompanied by a spoken description of the single news event.

Three different channels, text, video, and audio are at our disposal for the segmentation task. Our aim is to base the segmentation decision on the detected change in content across the various media. Although considerable work has been done in developing story segmenters that utilize numerous multimodal features we would like to investigate some of the text based features and methods developed in research to date. We wonder whether combining the various approaches into a single, unified segmentation algorithm might not improve performance of segmenting broadcast news video. In order to effectively operate on this multi-modal domain, we also include video and audio features in our investigation. Since our segmentation results form a basis for additional tasks such as summarization and concept detection, we wish to obtain the lowest possible error rate and so we introduce supervision to our segmentation efforts. In order to do this, we train a maximum entropy classifier on various multimedia features.

2 Related Work

Initial efforts at topic segmentation in text determine the lexical cohesion by measuring vocabulary repetition, as expressed by the cosine score of the term vectors representing two adjacent blocks of text. (Hearst 1997) assigns a story break between text blocks whose cosine scores differ greatly. (Choi 2000) computes an inter-sentence ranking from the cosine. Story segments are then identified by maximizing this ranking score while recursively partitioning the text. Other approaches use language models (Beeferman et al, 1999) or lexical chains (Kan et al. 1998, Galley et al. 2003) to compute lexical cohesiveness.

Segmentation of spoken discourse includes work done by (Passonneau et al. 1993, Galley et al., 2003) and makes use of a number of indicators such as cue-words, pause duration, and other forms of speech prosody. Work done for the TRECVID 2004[1] story segmentation task (of news video) is noteworthy as the approaches taken are more grounded in video retrieval. A representative example is IBM (Amir et al. 2004), who combine numerous visual features with specialized commercial and anchor (news reader) detectors, speech prosody, and textual features in order to find story boundaries.

3 Multimedia Features

Our intent is to identify story boundaries, using sentences as the candidate points between which story breaks occur. The following sections describe the features extracted from a multimedia document, and the motivation behind their choice. Ultimately these features will be used to train a maximum entropy classifier which will determine the existence of a story break at a particular sentence.[2]

Segment Likelihood. When considering whether to place a boundary at a candidate point, one can gauge the effect of preserving the segment integrity, versus splitting it up into two new segments, by computing the difference of the likelihoods that words within a segment are generated from the original segment or the two new segments.

$$Score(i) = \frac{\mathcal{L}(original)}{\mathcal{L}(original) + \mathcal{L}(new\ segments)}$$

$L(segment)=\Pi_{words} L(word|segment)= \Pi_{words} \alpha P(word|segment)+(1-\alpha)P(word|wiki)$

$$P(word|\,\text{segment}) = \frac{\#word\ \text{occurences in segment} - 1}{\text{total \# words in segment} - 1}$$

[1] http://www-nlpir.nist.gov/projects/tv2004/tv2004.html

[2] For the sake of brevity, we omit discussions of features related to vocabulary repetition as measured by the cosine metric as implemented by (Hearst 1997) and (Choi 2000), topic similarity as determined by Latent Direchlet Allocation (Blei 2003), and news broadcast program structure, as these features did not contribute to the final solution.

The likelihood function measures term repetition within a segment smoothed by the chance of it occurring naturally, as defined by term frequencies gathered from a large external corpus, which in our case was Wikipedia, which because of its diversity we consider it to be topic neutral. The resultant score is used as a feature.

Story Size. A layout related feature that we consider is the story size of the previous segment. The reasoning behind this is that the highlights section of a news broadcast consists of many, short consecutive passages. Thus the presence of short story segment, corresponding to such a story highlight, might indicate that another short segment might soon follow. This feature is certainly domain driven, but not entirely inconceivable.

Speech Pauses. Work done by among others (Passonneau 1993) has shown that speech prosody can contribute to the detection of story breaks, with speaker pause duration often being the most important feature. We assume a larger pause between stories than between the sentences of a story. We therefore used a voice-activity detector (Dekens et al. 2007) to extract the duration of all the silences in the audio channel. We identified the longest silence fragment immediately preceding a sentence, and used this duration as a feature.

Shot Cuts. The rapid change in visual content, usually due to some camera motion or change in scenery is referred to as a shot cut. One supposes that a visual change would be correlated with a change in semantic content. Like (Amir et al. 2004), we identify visual transitions using a shot cut detector provided by (Osian and van Gool 2004). Unfortunately, shot cuts do not necessarily indicate a story boundary. There are generally many shot cuts within a single story unit. Often shot cuts do not always precisely align with sentence boundaries. As a result we computed the shot cut feature by examining each sentence in the document and assigning a binary feature, which indicated the presence or absence of a shot cut during the time period in which a sentence was uttered. This time window was slightly offset, in order to catch shot cuts that occurred immediately prior to, or after, the sentence was read.

Cue Words: Chi-Square. In many works, such as (Beeferman 1999) and (Galley et al., 2003), cue words and phrases are used as a feature for the detection of topic breaks when segmenting text. Phrases such as “good morning,” or “this is Alastair Yates,” are commonly said by news anchors or reporters to begin or end a particular news story, and their detection might help in recognizing story breaks. We examined the phrases immediately preceding or following a story break in a training set, and compared this with how often they occurred in the rest of the corpus. We performed χ^2 tests at a significance level of 0.01 in order to determine the phrases indicative of a story transition. We discovered such phrases as, “Hello and welcome to BBC news”, “Good evening”, “Stay with us”, and “These are the headlines.” When performing feature extraction, each sentence receives a binary score indicating the presence or absence of cue phrases.

Implicit Cue Words. We implicitly learned cue phrases by training a maximum entropy classifier on sentences in our training set which were on, or away from, story boundaries. The probability score returned by this classifier forms one of our features.

Named Entity Chains. (Hearst 1997, Kan et al. 1998, Galley et al. 2003) use lexical chains to measure entity repetition and thus infer the cohesiveness of a prospective story segment. We observed that news stories often have disjoint sets of named entities, i.e. the proper names of people, places, or organizations. Applying a limited form of lexical chaining of named entities, the density of these chains over a segment closely mirrors the actual story segment; boundaries occur where there are few chains.

It should be noted that longer interviews do not as closely follow this pattern as there generally are less named entities due to dialogue etc. Also, very short story segments, such as the highlights section of a news broadcast, cannot be distinguished using this method as they generally have a similar amount of named entities.

We used the Stanford Named Entity Recognizer (Finkel et al, 2005). We calculated the number of named entity chains spanning every sentence as a feature.

Lexical Chains: Galley. We include the score feature developed by (Galley et al. 2003), which measures based on lexical chains. He combines the frequency of term repetition with chain compactness to arrive at a descriptor for lexical cohesiveness. This is then used to compute the rate of change in lexical cohesiveness, where a low score indicates a story boundary.

4 Story Boundary Selection

The number of story segments to find in a document can either be specified or be estimated from the training data. We adopt the following heuristic, we determine the ratio of sentences to story segments from the training data, and apply this figure to determine the number of segments in unseen documents in our test set.

After a random initialization, where we place that amount of boundaries in the document, we use an iterative method to reassign story boundaries based on a fitness criterion-a maximum entropy classifier trained using features from section 3. Purported story segments with a low fitness score will be removed and new story boundaries will be place at random position elsewhere. Over many iterations, certain candidate boundary positions (sentences) will more often have a story boundary occur on them. The candidate boundary positions for which this occurs most frequently are returned by the algorithm.

5 Evaluation

We have collected and annotated 14 news broadcasts from the BBC, which have a combined duration of around 7 hours. In addition, we have the corresponding transcripts for each broadcast, which consist of over 3000 sentences in total. The

transcripts were sometimes noisy due to transcription errors, and an effort was made to correct the worst distortions by hand.

In text based segmentation, quite a few evaluation metrics have been proposed. Early papers such as (Hearst 1994) used the precision/recall metrics, although this metric is too strict in that it penalizes boundaries that have been placed very close to, but not on the ground-truth boundary. As a result, degenerate algorithms which place a boundary after every possible sentence can actually achieve a higher precision and recall score. (Beeferman et al. 1999) proposed a metric, P_k, which penalizes degenerate algorithms yet also gives partial credit for boundaries which are close to the actual boundary. An improvement on the P_k metric, called WindowDiff (WD), was introduced by (Pevzner and Hearst 2001).

We performed leave-one-out (leaving out 1 broadcast every time) cross-validation to train and evaluate our segmentation algorithm. The average P_k and WD scores were computed from the held-out test sets.

We established a strong baseline, by using our likelihood function for lexical cohesion, story segment size, and automatically learned cue phrases. We then incrementally added in the remaining features into our maximum entropy classifier. We then considered the effects of a late fusion of our best performing maximum entropy classifier with the remaining unused features. Late fusion was performed as follows:

$$P(story\ break|Classifier) = \lambda_i P(story\ break|MaxEnt_{best}) + \lambda_{i+1} P(story\ break|feature) + \cdots$$

A new classifier was trained on our training set by interpolating the results of our best Maximum Entropy classifiers with our unused features. Interpolation weights were determined with the use of the Expectation Maximization (EM) algorithm. Only cue words and the lexical chain (GAL) feature gave an improvement.
We list the runs which showed a consistent performance increase in Table 1.

Table 1. Best performing feature combinations.

	BASE	BASE+ NE+ GAL	BASE+ PAUSE+ CUE+ GAL	BASE+NE+GAL+ Late Fusion(CUE)	BASE+PAUSE+ CUE+GAL+ Late Fusion(CUE)	BASE+PAUSE+ CUE+GAL+Late Fusion (CUE+GAL)
WD	0.212	0.209	0.210	**0.194**	0.198	0.197
P_K	0.135	0.142	0.141	**0.119**	0.122	0.121
BASE is the combination of segment likelihood, segment size, and implicit cue words						
PAUSE is the pause duration feature				Cue is the χ2 Cue Words		
NE is Named Entity chain feature				GAL is the score defined by Galley		

We then evaluated our resultant system against Galley's segmenter (Galley et al., 2003), LCSeg, and Choi's segmenter (Choi 2000), C99. These are two state of the art systems which segment text in an unsupervised fashion based on text-only features. Our system differs from theirs in that we use multiple features, including those specific to multimodal data. Both C99 and LCSeg, like ours, are capable of automatically determining the number of segments in a document. We provide results for both modes of operation, referred to as *known* (number of segments is known) and *unknown* respectively in Table 2. Both C99 and LCSeg ran using their default parameters.

Table 2. Comparison of the results from C99 and LCSeg.

	BASE+NE+GAL+ Late Fusion $(CUE)_{known}$	BASE+NE+GAL+ Late Fusion $(CUE)_{unknown}$	$C99_{known}$	$C99_{unknown}$	$LCSseg_{known}$	$LCSeg_{unknown}$
WD	**0.194**	0.220	0.363	0.307	0.276	0.243
P_k	**0.119**	0.149	0.323	0.268	0.218	0.191

6 Conclusion

Our initial belief that a unification of several features and methods from the textual modality would result in improved segmentation performance was validated by our final result. The inclusion of multimedia features, such as shot cuts and pause duration, only resulted in a performance increase with the pause feature. Data analysis revealed that too many shot cuts occurred during a single story segment, both on actual boundaries and on intra-segment sentences, meaning this feature was insufficiently discriminative. Ironically our best segmentation of our news video dataset used only features from the textual modality.

We acknowledge that many previous segmentation efforts in research have focused on unsupervised methods, yet we also feel we have more than adequately demonstrated the improved performance made possible by the use of a supervised training method. In light of the small development set requirement (13 broadcasts), and future mission critical applications (document summarization and concept detection), this seems a justified choice, as the performance of our segmenter exceeds that of other methods.

References

A. Amir, J. Argillander, M. Berg, S-F. Chang, et al. "IBM Research TRECVID-2004 Video Retrieval System." Proc. TRECVID 2004.

D. Beeferman, A. Berger, J. Lafferty. "Statistical models for text segmentation." Machine Learning 34, 1999: (1–3)177–210.

D. M. Blei, A. Y. Ng, M. I. Jordan. "Latent Dirichlet Allocation." Journal of Machine Learning Research, 2003: (3) 993-1022.

F. Choi, "Advances in domain independent linear text segmentation." Proc NAACL. 2000.

T. Dekens, M. Demol, W. Verhelst, F. Beaugendre. "Voice Activity Detection based on Inverse Normalized Noise Likelihood Estimation." CIE 2007, Santa Clara, Cuba, June 18-22, 2007

J. Finkel, T. Grenager, C. Manning. " Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling." ACL. 2005.

M. Galley, K. McKeown, E. Fosler-Lussier, H. Jing. "Discourse Segmentation of Multi-party Conversation." Proc. ACL. Sapporo, Japan, 2003. 562-569.

M. A. Hearst, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages." Computational Linguistics, 1997: 23(1)33-64

M-Y. Kan, J. L. Klavans, K. R. McKeown. "Linear Segmentation and Segment Significance." Proc. 6th Workshop on Very Large Corpora. 1998.

M. Osian, L. Van Gool, “Video shot characterization.” Machine Vision and Applications, vol. 15, no. 3, p. 172-177, 2004.

R. J. Passonneau, D. J. Litman. "Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues." Proc. ACL. 1993. 148-155.