Fine-Grained Hidden Markov Modeling for Broadcast-News Story Segmentation

Warren Greiff, Alex Morgan, Randall Fish, Marc Richards, Amlan Kundu,
MITRE Corporation
202 Burlington Road
Bedford, MA 01730-1420
(greiff, amorgan, fishr, marc, akundu)@mitre.org

ABSTRACT

We present the design and development of a Hidden Markov Model for the division of news broadcasts into story segments. Model topology, and the textual features used, are discussed, together with the non-parametric estimation techniques that were employed for obtaining estimates for both transition and observation probabilities. Visualization methods developed for the analysis of system performance are also presented.

1. INTRODUCTION

Current technology makes the automated capture, storage, indexing, and categorization of broadcast news feasible allowing for the development of computational systems that provide for the intelligent browsing and retrieval of news stories [Maybury, Merlino & Morey '97; Kubula, et al., '00]. To be effective, such systems must be able to partition the undifferentiated input signal into the appropriate sequence of news-story segments.

In this paper we discuss an approach to segmentation based on the use of a fine-grained Hidden Markov Model [Rabiner, `89] to model the generation of the words produced during a news program. We present the model topology, and the textual features used. Critical to this approach is the application of non-parametric estimation techniques, employed to obtain robust estimates for both transition and observation probabilities. Visualization methods developed for the analysis of system performance are also presented.

Typically, approaches to news-story segmentation have been based on extracting features of the input stream that are likely to be different at boundaries between stories from what is observed within the span of individual stories. In [Beeferman, Berger, & Lafferty '99], boundary decisions are based on how well predictions made by a long-range exponential language model compare to those made by a short range trigram model. [Ponte and Croft, '97] utilize Local Context Analysis [Xu, J. and Croft, '96] to enrich each sentence with related words, and then use dynamic programming to find an optimal boundary sequence based on a measure of word-occurrence similarity between pairs of enriched sentences. In [Greiff, Hurwitz & Merlino, `99], a naïve Bayes classifier is used to make a boundary decision at each word of the transcript. In [Yamron, et al., '98], a fully connected Hidden Markov Model is based on automatically induced topic clusters, with one node for each topic. Observation probabilities for each node are estimated using smoothed unigram statistics.

The approach reported in this paper goes further along the lines of find-grained modeling in two respects: 1) differences in feature patterns likely to be observed at different points in the development of a news story are exploited, in contrast to approaches that focus on boudary/no-boundary differences; and 2) a more detailed modeling of the story-length distribution profile, unique to each news source (for example, see the histogram of story lengths for ABC World News Tonight shown in the top graph of Figure 3, below).

2. GENERATIVE MODEL

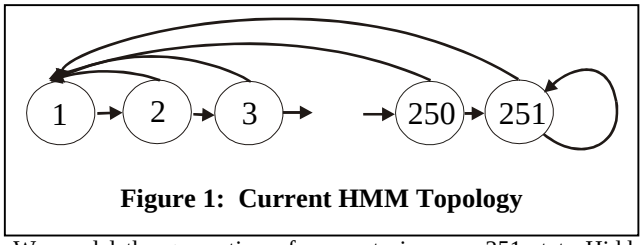

Figure 1: Current HMM Topology

We model the generation of news stories as a 251 state Hidden Markov Model, with the topology shown in Figure 1. States labeled, 1 to 250, correspond to each of the first 250 words of a story. One extra state, labeled 251, is included to model the production of all words at the end of stories exceeding 250 words in length.

Several other models were considered, but this model is particularly suited to the features used, as it allows one to model features that vary with depth into the story (Section 3.1), while simultaneously, by delaying certain features. It also allows one to model features that occur in specific regions the boundaries (Section 3.3). This is possible because all states can feed into the initial state, i.e. all stories end by going into the first word of a new story.

For example, the original model involved a series of beginning and then end states, with a single middle state that could be cycled through (Figure 2). This proved to be a problem because the ends of long stories were being mixed with the ends of short stories which led to problems with our spaced coherence feature (Section 3.1). Another possibility involved splitting the model into two main paths, one to model the shorter stories, and one to model the longer as there is something of a bimodal distribution in story lengths (Figure 4). However, the fine-grained nature of our model would suffer from splitting the data in this manner, and a choice about at which length to fork the model would be somewhat artificial.

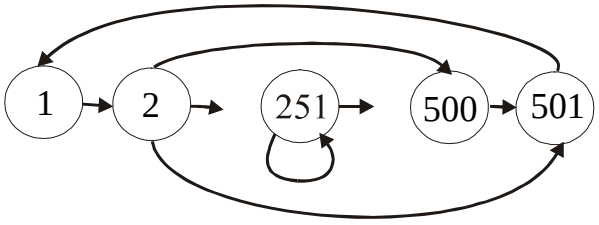

Figure 3: Original Topology

3. FEATURES

Associated with the model is a set of features. For each state, the model assigns a probability distribution over all possible combinations of values the features may take on. The probability assigned to value combinations is assumed to be independent of the state/observation history, conditioned on the state. We further assume that the value of any one feature is independent of all others, once the current state is known. Features have been explicitly designed with this assumption in mind. Three categories of features have been used, which we refer to as *coherence* features, *x-duration* feature, and the *trigger* features.

3.1. Coherence

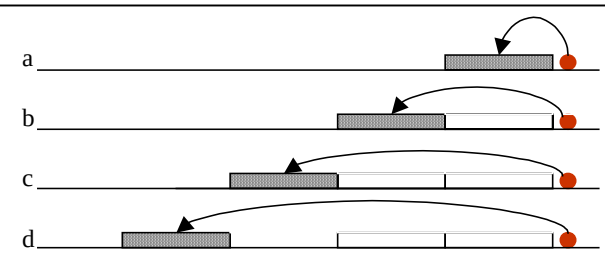

Figure 2: Coherence Features

We have used four coherence features. The COHER-1 feature, shown schematically in Figure 2a, is based on a buffer of 50 words immediately prior to the current word. If the current word does not appear in the buffer, the value of COHER-1 is 0. If it does appear in the buffer, the value is $-\log(s_w/s)$, where s_w is the number of stories in which the word appears, and s is the total number of stories, in the training data. Words that did not appear in the training data, are treated as having appeared once. In this way, rare words get high feature values, and common words get low feature values. Three other features: COHER-2, COHER-3, and COHER-4 (Figures 3b, c & d) correspond to similar features; for these, however, the buffer is separated by 50, 100, and 150 words, respectively, from the current word. Interestingly, the COHER-4 feature actually caused a reduction in performance, and was not used in the final evaluation.

3.2. X-duration

This feature is based on indications given by the speech recognizer that it was unable to transcribe a portion of the audio signal. The existence of an untranscribable section prior to the word gives a non-zero X-DURATION value based on the extent of the section. Empirically this is an excellent predictor of boundaries in that an untranscribable event has uniform likelihood of occurring anywhere in a news story, except prior to the first word of a story, where it is extremely likely to occur.

3.3. Triggers

Trigger features correspond to small regions at the beginning and end of stories, and exploit the fact that some words are far more likely to occur in these positions than in other parts of a news segment. One region, for example, is restricted to the first word of the story. In ABC's World News Tonight, for example, the word "finally" is far more likely to occur in the first word of a story than would be expected by its general rate of occurrence in the training data. For a word, *w*, appearing in the input stream, the value of the feature is an estimate of how likely it is for *w* to appear in the region of interest. The estimate used is given by:

$$\hat{p}(w \in R) = \frac{n_{w \in R} + 1}{n_w + (1/f_R)}$$

where $n_{w \in R}$ is the number of times *w* appeared in *R* in the training data; n_w is the total number of occurrences of *w*; and f_R is the fraction of all tokens of *w* that occurred in the region. This estimate can be viewed as Bayesian estimate with a beta prior. The beta prior is equivalent to a uniform prior and the observation of one occurrence of the word in the region out of $(1/f_R)$ total occurrences. This estimate was chosen so that: 1) the prior probability would not be greatly affected for words observed only a few times in the training data; 2) it would be pushed strongly towards the empirical probability of the word appearing in the region for words that were encountered in *R*; 3) it has a prior probability, f_R, equal to the expectation for a randomly selected word. The regions used for the submission were restricted to the one-word regions for: first word, second word, last word, and next-to-last word. Limited experimentation with multi-state regions, was not fruitful. For example, including the regions, {3,4,…,10} and {-10,-9,…,-3}, where $-i$ is interpreted as i words prior to the end of the story, did not improve segmentation performance.

Since, as described, the current HMM topology does not model end-of-story words (earlier versions of the topology did model these states directly), trigger features for end-of-story regions are delayed. That means that a trigger related to the last word in a story would be delayed by a one word buffer. In this way, it is linked to the first word in the next story. For example, the word "Jennings" (the name of the main anchorperson) is strongly

correlated with the last word in news stories in the ABC World News Tonight corpus. The estimated probability of it being the last word of the story in which it appears is .235 (obtained by the aforementioned method). The trained model associates a high likelihood of seeing the value .235 at state = 1; the intuitive interpretation being, "a word highly likely to appear at the last word of a story, occurred 1-word ago".

4. PARAMETER ESTIMATION

The Hidden Markov Model requires the estimation of transition and conditional observation probabilities. There are 251 transition probabilities to be estimated. Much more of a problem are the observation probabilities, there being 9 features in the model, for each of which a probability distribution over as many as 100 values must be estimated, for each of 251 states. With the goal of developing methods for robust estimation in the context of story segmentation, we have applied non-parametric kernel estimation techniques, using the LOCFIT library [Loader, '99] of the R open-source statistical analysis package, which is based on the S-plus system [Venables & Ripley,

`99; Chambers & Hastie, `92, Becker, Chambers & Wilks, `88]. For the transition probabilities, it is assumed that the underlying probability distribution over story length is smooth, allowing the empirical histogram, shown at the top of Figure 4, to be transformed to the probability density estimate shown at the bottom. From this probability distribution over story lengths, the conditional transition probabilities can be estimated directly.

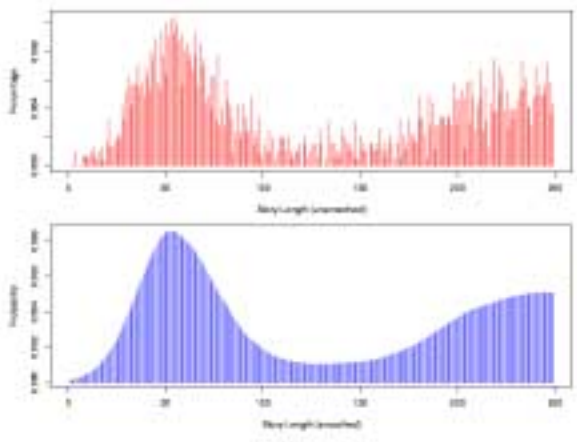

Figure 4: Histograms of story lengths (up to 250 words) -- raw and smoothed --

Conditional observation probabilities are also deduced from an estimate of the joint probability distribution. First, observation values were binned. Binning limits were set in an attempt to 1) be large enough to obtain sufficient counts for the production of robust probability estimates, and yet, 2) be constrained enough so that important distinctions in the probabilities for different feature values will be reflected in the model. For each bin, the observation counts are smoothed by performing a non-parametric regression of the observation counts as a function of state. The smoothed observations counts corresponding to the regression are then normalized so as to sum to the total observation count for the bin. The result is a conditional probability distribution over states for a given binned feature value, *p(State=s|Feature=fv)*. Once this is done for all bin values, each conditional probability is multiplied by the marginal probability, *p(State=s)*, of being in a given state, resulting in a joint distribution, *p(fv,s)*, over the entire space of *(Feature,State)* values. From this joint distribution, the necessary conditional probabilities, *p(Feature=fv|State=s)*, can be deduced directly.

Figure 5 shows the conditional probability estimates, *p(fv | s)*, for the feature value COHER-3=20, across all states, confirming the intuition that, while the probability of seeing a value of 20 is small for all states, the likelihood of seeing it is much higher in latter parts of a story than it is in early-story states.

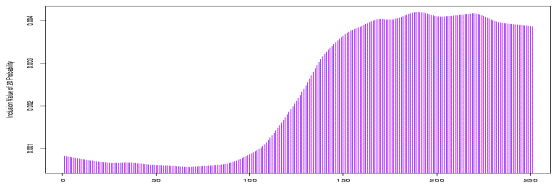

Figure 5: Likelihood of COHER-3=2 over all states

5. SEGMENTATION

Once parameters for the HMM have been determined, segmentation is straightforward. The Viterbi algorithm [Rabiner, `89], is employed to determine the sequence of states most likely to have produced the observation sequence associated with the broadcast. A boundary is then associated with each word produced from State 1 for the maximum likelihood state sequence.

The version of the Viterbi algorithm we have implemented provides for the specification of "state-penalty" parameters, which we have used for the "boundary state", state 1. In effect, the probability for each path in consideration is multiplied by the value of this parameter (which can be less than, equal to, or greater than, 1) for each time the path passes through the boundary state. Variation of the parameter effectively controls the "aggressiveness" of segmentation, allowing for tuning system behavior in the context of the evaluation metric.

6. RESULTS

Preliminary test results of this approach are encouraging. After training on all but 15 of the ABC World News Tonight programs from the TDT-2 corpus [Nist, '00], a test on the remaining 15 produced a false-alarm (boundary predicted incorrectly) probability of .11, with a corresponding miss (true boundary not predicted) probability of .14, equal to the best performance reported to date, for this news source.

A more intuitive appreciation for the quality of performance can be garnered from the graphs in Figure 6, which contrast the segmentation produced by the system (middle) with ground truth (the top graph), for a typical member of the ABC test set. The x-axis corresponds to time (in units of word tokens); i.e., the index of the word produced by the speech recognizer, and the y-axis

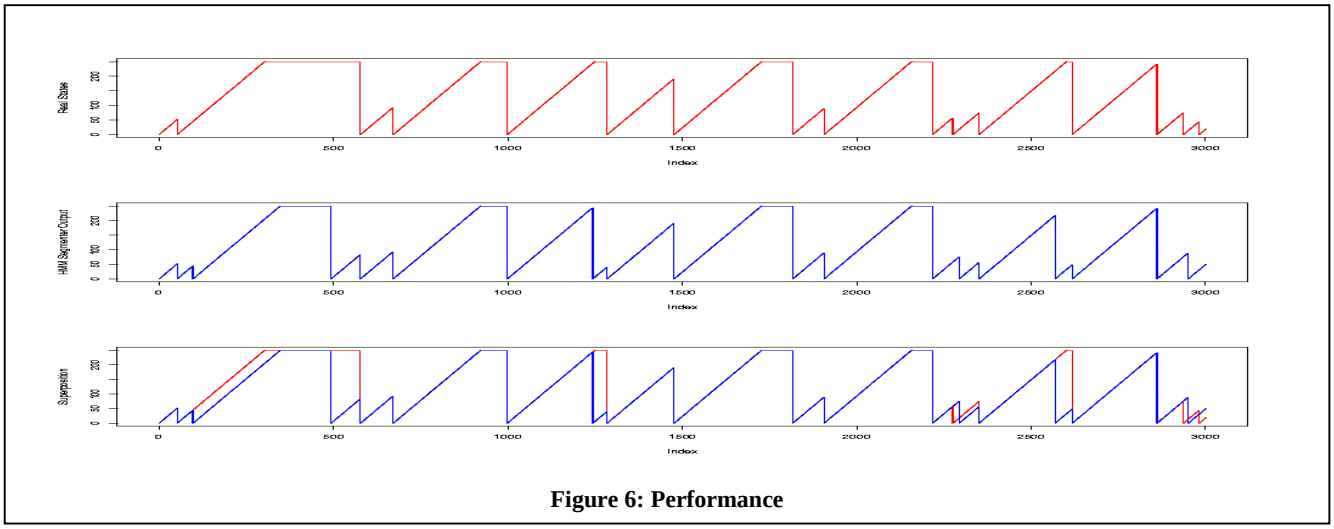

Figure 6: Performance

corresponds to the state of the HMM model. A path passing through the point (301, 65), for example, corresponds to a path through the network that produced the 65th word from state 301. Returns to state=1 correspond to boundaries between stories. The bottom graph shows the superposition of the two to help illustrate the agreement between the path chosen by the system and the path corresponding to perfect segmentation..

7. VISUALIZATION

The evolution of the segmentation algorithm was driven by analysis of the behavior of the system, which was supported by visualization routines developed using the graphing capability of the R package. Figure 7 gives an example of the kind of graphical displays that were used for analysis of the segmentation of a specific broadcast news program; in this case, analysis of the role of the X-DURATION feature. This graphical display allows for the comparison of the maximum likelihood path produced by the HMM to the path through the HMM that would be produced by a perfect system – one privy to ground-truth.

The top graph corresponds to the bottom graph of Figure 6, showing the states traversed by the two systems. The second graph shows the value of the X-DURATION feature corresponding to each word of the broadcast. So, the plotting of a point at (301, 3) corresponds to an X-DURATION value of 3 having been observed at time, 301. One thing that can be seen from this graph is that being at a story boundary (low-points on the thicker-darker line of the top graph) is more frequent when higher values of the X-DURATION cue are observed, than when lower values are observed, as could be expected.

The third graph shows, on a log scale, how many times more likely it is that the observed X-DURATION value would be generated from the true state than from the state predicted by the system. Most points are close to 0, indicating that the X-DURATION value observed was as likely to have come from the true state as it is to have come from the state predicted by the Viterbi algorithm. Of course, this is the case wherever the true state has been correctly predicted. Negative points indicate that the X-DURATION value observed is less likely to be produced from the true state than from the predicted state. Strongly negative points are a major component of the probability calculation that resulted in the system preferring the path it chose over the true path. These points suggest potential deficiencies in the modeling. Their identification directs the focus of analysis so that system performance can be improved by correcting weaknesses of the existing model.

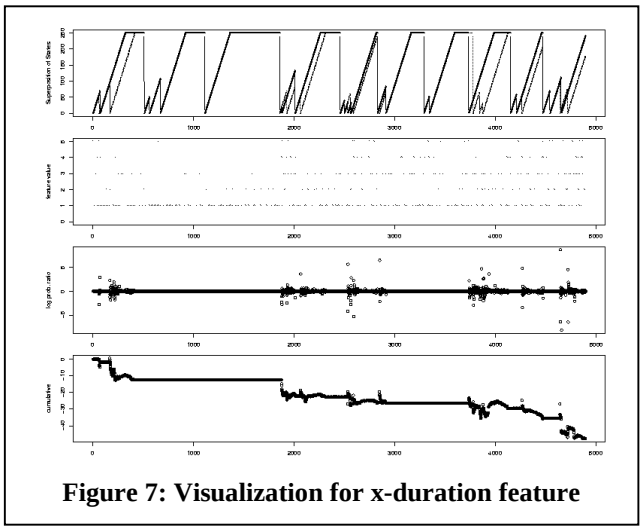

Figure 7: Visualization for x-duration feature

The final graph shows the cumulative sum of the values from the graph above it. (Note that the sum of the logs of the probabilities is equivalent to the cumulative product of probabilities on a log scale.) The graphing of the cumulative sum can be very useful when the system is performing poorly due to a small but consistent preference for the observations having been produced by the state sequence chosen by the system. This phenomenon is made evident by a steady downward trend in the graph of the cumulative sum. This is in contrast to an overall level trend with occasional downward dips. Note, that a similar graph for the total probability (equal to the product of all the individual feature value probabilities) will always have an overall downward trend, since the maximum likelihood path will always have a likelihood

greater than the likelihood of any other path.

Aside from supporting the detailed analysis of specific features, the productions of these graphs for each of the features, together with the corresponding graph for the total observation probability, allowed us to quickly asses which of the features was most problematic at any given stage of model development.

8. FURTHER WORK

It should be kept in mind that experimentation with this approach has been based on relatively primitive features – our focus, to this point, having been on the development of the core segmentation mechanism. Features based on more sophisticated extraction techniques, which have been reported in the literature – for example, the use of exponential models for determining trigger cues used in [Beeferman, Berger, & Lafferty '99] – can easily be incorporated into this general framework. Integration of such techniques can be expected to result in significant further improvement in segmentation quality.

To date, the binning method described has given much better results than two dimensional kernel density estimation techniques which we also attempted to employ. One of the main difficulties with using traditional kernel density estimation techniques is that they tend to inaccurately estimate the density at areas of discontinuity, such as state=1 in our model and our trigger features. Preliminary work with boundary kernels [Scott, '92] is very promising. It is certainly an area worthy of more in-depth investigation.

Work done by another group [Liu, '00] to segment documentaries based on video cues alone has been moderately successful in the past. We engineered a neural network in an attempt to identify video frames containing an anchorperson, a logo, and blank frames, with a belief that these are all features that would contain information about story boundaries. Preliminary work was also done to extract features directly from the audio signal, such as trying to identify speaker change. Initial work with the audio and video has been unable to aid in segmentation, but we feel this is also an area worth continuing to pursue.

9. REFERENCES

1. [Becker, Chambers & Wilks, `88] Becker, Richard A., Chambers, John M., and Wilks, Allan R. *The New S Language*. Wadsworth & Brooks/Cole, Pacific Grove, Cal.
2. [Beeferman, Berger, & Lafferty '99] D. Beeferman, D., A. Berger, A. and Lafferty, J. Statistical models for text segmentation. *Machine Learning*, vol. 34, pp. 1-34, 1999.
3. [Chambers & Hastie, `88] Chambers, John M. and Hastie, Trevor, J. *Statistical Models in S.* Wadsworth & Brooks/Cole, Pacific Grove, Cal., 1988.
4. [Greiff, Hurwitz & Merlino, `99] Greiff, Warren, Hurwitz, Laurie, and Merlino, Andrew. MITRE TDT-3 segmentation system. *TDT-3 Topic Detection and Tracking Conference*, Gathersburg, Md, February, 2000.
5. [Kubula, et al., '00] Kubula, F., Colbath, S., Liu, D., Srivastava, A. and Makhoul, J. Integrated technologies for indexing spoken language, *Communication of the ACM*, vol. 43, no. 2, Feb., 2000.
6. [Liu, '00] Liu, Tiecheng and Kender, John R. A hidden Markov model approach to the structure of documentaries. *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries, 2000.*
7. [Loader, `99] Loader, C. *Local Regression and Likelihood.* Springer, Murray Hill, N.J., 1999.
8. [Maybury, Merlino & Morey '97] Maybury, M., Merlino, A. Morey, D. Broadcast news navigation using story segments. *Proceedings of the ACM International Multimedia Conference*, Seattle, WA, Nov., 1997.
9. [Nist, '00] Topic Detection and Tracking (TDT-3) Evaluation Project. http://www.nist.gov/speech/tests/tdt/tdt99/.
10. [Ponte and Croft, '97] Ponte, J.M. and Croft, W.B. Text segmentation by topic, *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pp. 120--129, 1997.
11. [Rabiner, `89] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, vol. 37, no. 2, pp. 257-86, February, 1989.
12. [Scott, '92] David W. Scorr, Boundary kernels, *Multivariate Density Estimation: Theory and Practice*, pp 146-149, 1992.
13. [Venables & Ripley, `99] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S-PLUS.* Springer, Murray Hill, N.J., 1999.
14. [Xu, J. and Croft, '96] Xu, J. and Croft, W.B., Query expansion using local and global document analysis, *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4--11, 1996
15. [Yamron, et al., '98] Yamron, J. P., Carp, I., Gillick, L., Lowe, S. and van Mulbregt, P. A Hidden Markov Model approach to text segmentation and event tracking. *Proceedings ICASSP-98*, Seattle, WA. May, 1998.