MITRE TDT-2000 SEGMENTATION SYSTEM

Warren Greiff, Alex Morgan, Randall Fish, Marc Richards, Amlan Kundu,
(greiff, morgan, fishr, marc, akundu)@mitre.org

MITRE Corporation
202 Burlington Road
Bedford, MA 01730-1420

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

For last year's, TDT-3, evaluation, the MITRE system utilized a Naïve Bayes classifier [Greiff, Hurwitz & Merlino, `99]. For the TDT-2000 evaluation, we have explored the use of a fine-grained, multi-state Hidden Markov Model (HMM). After experimentation and performance analysis with an initial design, the model topology, as well as the feature definitions used were found to be unsatisfactory in a number of respects. In the sections that follow, we describe the topology and features that have evolved from this experience, and are used in the current version of the segmentation system. Also described is our approach to the determination of model parameters using non-parametric kernel estimation methods. After a brief description of how the resulting model is used for actual segmentation, we discuss visualization aids developed to support analysis of the system, and conclude with a summary of our TDT-2000 submission results, and plans for future work.

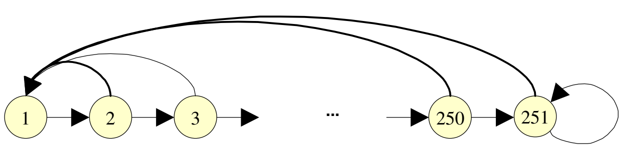

Figure 1: HMM Topology

2. APPROACH

The model we have used is a 251 state Hidden Markov Model, with the topology shown in Figure 1. States labeled, 1 to 250, correspond to 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. There are two principal motivations for this approach. First, that with a model of this nature we will be able to exploit the distribution of story-lengths that is unique to each source. For example, in the top graph of Figure 3, below, we see a histogram of story lengths (up to 250 words) for the ABC_WNT news source. The histogram displays a bimodal distribution with a clear preference for stories in the range of 30 to 80 words. This is evidence that can be used to improve segmentation. The second motivating factor was to go beyond the approach we employed in TDT-3, which used a Naïve Bayes classifier to classify each word as to whether or not it is at a boundary. Classification of each word was independent of all others. The goal of the fine-grained HMM approach is to enable the exploitation of features that give strong indications of which portion of a story we are in; indications which go beyond the binary boundary/no-boundary distinction. The HMM approach also supports the production of a single coherent segmentation sequence in place of a series of independent boundary/no-boundary decisions.

3. FEATURES

The Hidden Markov Model is defined in terms of a set of features. The model assigns, for each state, 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. The MITRE system further assumes 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, *trigger* features, and the *x-duration* feature.

3.1. Coherence

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 (to which the feature value corresponds). If the current

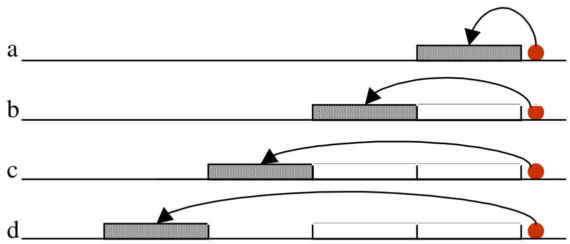

Figure 2: Coherence features

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. In this way, rare words get high feature values, and common words get low feature values. The idea behind this feature is that high values will be very unlikely to be observed at early points in the story, where the majority of the buffer corresponds to words of the previous story. Three other features: COHER-2, COHER-3, and COHER-4 (Figures 2b, c & d) correspond to similar features; for these, however, the buffer is separated by 50, 100, and 150 words, respectively, from the current word. Presumably, the probability of observing a high value of COHER-3, for example, would be very low in states 1 to 100, where the buffer is entirely over words of the previous story; much higher in states 150 - 250, where the buffer is entirely over words of the same story; and grow monotonically between states 100 and 150.

3.2. X-duration

This feature is based on the untranscribable portions of the audio signal, as indicated by X entries in the ASR file. The absence of an X entry prior to a word corresponds to an X-DURATION feature value of 0. The existence of an X entry prior to the word gives a non-zero value, with large X entry durations corresponding to high values, and low X entry durations corresponding to low values.

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. For ABC_WNT, 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 a Bayesian estimate with a beta prior. The beta prior is equivalent to a uniform prior together with 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 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. Suppose the delay were 10 words, sufficient to cover all end-of-story regions, and the region of interest were {-4,-3,-2}. High values for this trigger feature could be interpreted as "a word showing marked preference for occurring 2 to 4 words prior to the end of a story, occurred 10 words ago", and could be expected to be observed most often in states 6, 7 and 8.

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, 1999) of the R open-source statistical analysis package, which is based on the S-plus system [Venables & Ripley,

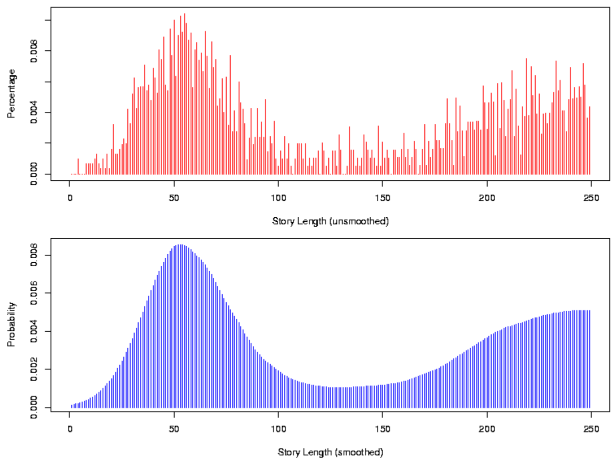

Figure 3: Histograms of story lengths (raw and smoothed)

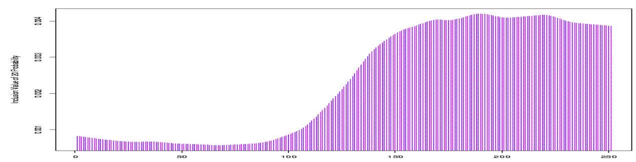

Figure 4: Likelihood of COHER-3=20 over all states.

`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 3, 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.

A similar approach is taken with regard to observations, in that conditional probabilities are deduced from an estimate of the joint probability distribution. The following procedure was used. 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. At this time, the limits have been set manually, as the result of studying the data distributions, and comparing the results of different binning strategies. 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 4 gives a comparison of 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 later parts of a story than it is in early-story states.

The preceding description assumes that probabilities are smooth across all states. This is not always the case. For example, exploratory analysis of the data reveals that state 1 is notably different from other states with respect to the X-DURATION feature. Because of this, a smoothing over all states is not appropriate. Fortunately, there are abundant observations for STATE=1. This allows us to estimate the conditional probability distribution over feature values for *STATE=1* independently, and restrict the procedure described above to the non-boundary states. For some features, special consideration must also be given to the mid-story state, 251.

5. SEGMENTATION

Once parameters for the HMM have been determined, segmentation is straight-forward. The Viterbi algorithm [Rabiner, `89], is employed to determine the sequence of states most likely to have produced the observations 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 the tuning of system behavior in the context of the evaluation metric.

6. VISUALIZATION

The MITRE submission was the result of a number of modifications and refinements of the HMM topology, design of the features, and the estimation procedures. This

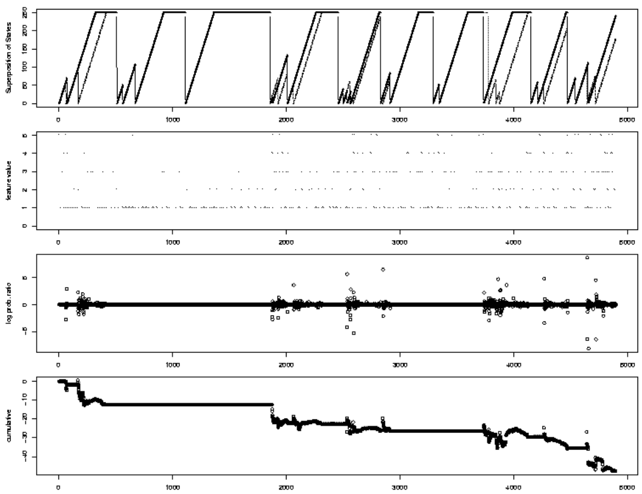

Figure 5: Visualization for X-DURATION feature.

evolution was driven by analysis of the behavior of the system, supported by visualization routines developed using the graphing capability of the R package. Figure 5 gives an example of graphs that can be used for analysis of the role of the X-DURATION feature, as part of the segmentation of a specific file. These graphs compare the maximum likelihood path produced by the HMM (lighter, thinner line) to the path that would be produced by a perfect system (thicker, darker line) -- one privy to ground-truth, as given by the boundary file. The top graph shows the states traversed by the two systems. The graph below shows the value of the X-DURATION feature corresponding to each word of the broadcast. The third graph shows, on a log scale, how many times more likely it is that this value would be generated from the true state than in the state predicted by the system; with positive values corresponding to the true state being more likely to produce the observation, and negative values corresponding to situations for which the observation is more likely from the system's state. The final graph shows the cumulative sum of the values from the graph above it. 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. This is because the system always chooses the maximum likelihood path. This is the path with greater likelihood than all others according to the HMM model, including the true path, to which the observations actually do correspond. These graphs help to direct attention to potential weaknesses in the modeling. For example, towards the end of the story, we see a cluster of three points (lower right hand corner of the probability ratio graph) corresponding to observations less likely to have been generated by the true state sequence, than by the state sequence produced by the system. This is an indication that this feature may have contributed significantly to the mistaken placement of a boundary, after the last true boundary, as shown in the top graph.

7. RESULTS

As mentioned above, the process of getting the core HMM segmentation mechanism to an acceptable base-level of performance proved to be much more difficult than we had anticipated. As the Oct. 10 submission deadline drew near, key issues in the implementation of the system still needed to be worked out. Our submission system did not perform well compared to segmentation scores reported last year at TDT-3. In the intervening time we have continued work on our approach, focusing on the analysis of system performance on the ABC_WNT news source. At this point, preliminary results for this part of the corpus are encouraging. After training on all but 15 files of the development set, a test on the remaining 15 produced a normalized Cseg score of .31, comparable to the scores of .31 and .33 achieved by IBM and CMU respectively, at TDT-3.

8. FUTURE WORK

Analysis of segmentation performance on sources other than ABC_WNT will be the initial focus of continued work on the fine-grained HMM approach to broadcast news segmentation. Analysis of each feature individually will be required to obtain insights into the modeling questions that require specific attention, taking full advantage of the visualization techniques, discussed above, to study the segmentation produced by the system on specific files. In support of this effort, and continued efforts to improve performance for all sources, we will be looking at devising theoretically motivated methods for evaluating the value of a given feature and summarizing its impact on segmentation performance over a large number of files. We also plan to take a closer look at the problem of probability estimation. A number of parameters must be set before estimation can proceed. We have found, for example, that local linear fitting with a rectangular kernel seems to give best performance, but time has not permitted the methodical comparison of alternatives such as higher degree polynomial fitting, and differently shaped kernel functions. We would also like to more carefully examine the question of choosing the best bandwidth for smoothing, and the measurement of goodness-of-fit. Finally, our original plan for TDT-2000 included experimentation, in the context of our HMM approach, with the use of more sophisticated textual features, such as the use of log-linear modeling for the determination of more complex trigger cues as developed in [Beeferman, Berger, & Lafferty '99], as well as the exploitation of features extracted directly from the audio signal, continuing work on prosodic modeling reported at the TDT-2 conference in [Stolcke, et al. ‘98]. This continues to be our goal, even though we were unable to explore these possibilities for this evaluation.

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. [Lau, Rosenfeld & Roukos, `93] Lau, Raymond, Rosenfeld, Ronald and Roukos, Salim. Trigger-Based Language Models: A Maximum Entropy Approach. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. II 45-48, Minneapolis, MN, April, 1993.
6. [Loeder, `99] Loader, C. *Local Regression and Likelihood*. Springer, Murray Hill, N.J., 1999.
7. [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.
8. [Venables & Ripley, `99] Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S-PLUS*. Springer, Murray Hill, N.J., 1999.
9. [Stolcke, et al. ‘98] Stolcke, Anderas, Shriber, Elizabeth. Hakkani-Tur, Dilek, Tur, Gokhan, Rivlin, Ze’ev and Sonmez, Kemal. Combining Words and Speech Prosody for Automatic Topic Segmentation, TDT-2 Topic Detection and Tracking Conference, Gathersburg, Md, February, 1999.