A New Quality Measure for Topic Segmentation of Text and Speech

Mehryar Mohri [1,2]*, Pedro Moreno* [1]*, and Eugene Weinstein* [1,2]

[1] Google Inc.
76 Ninth Avenue, New York, NY 10011.
[2] Courant Institute of Mathematical Sciences
251 Mercer Street, New York, NY 10012.

Abstract

The recent proliferation of large multimedia collections has gathered immense attention from the speech research community, because speech recognition enables the transcription and indexing of such collections. Topicality information can be used to improve transcription quality and enable content navigation. In this paper, we give a novel quality measure for topic segmentation algorithms that improves over previously used measures. Our measure takes into account not only the presence or absence of topic boundaries but also the content of the text or speech segments labeled as topic-coherent. Additionally, we demonstrate that topic segmentation quality of spoken language can be improved using speech recognition lattices. Using lattices, improvements over the baseline one-best topic model are observed when measured with the previously existing topic segmentation quality measure, as well as the new measure proposed in this paper (9.4% and 7.0% relative error reduction, respectively).

Index Terms: Topic segmentation, speech recognition lattices, text similarity, speech processing.

1. Introduction

Natural language streams, such as news broadcasts and telephone conversations, are marked with the presence of underlying topics. These topics influence the statistics of the text or speech produced. Learning to identify the topic underlying a given segment of speech or text, or to detect topic changes is beneficial in a number of ways. For example, knowledge of the topic of a speech recording being transcribed by a speech recognizer can be used to improve transcription quality by using a topic-dependent language model. Topicality information can also be used to improve navigation of audio and video collections such as YouTube, by considering a common topic as a feature when creating links between items.

In this paper, we focus on topic segmentation, or the automatic detection of topic changes in text or speech. After a review of previous work, we point out major limitations of the currently accepted topic segmentation quality measure known as CoAP, including the fact that it does not take into account the word content of the segments produced by the algorithms. We then introduce a general measure of text similarity and give a topic segmentation quality measure incorporating this similarity score and overcoming many of the limitations of CoAP. In experiments over speech and text streams from the Topic Detection and Tracking (TDT) corpus, we demonstrate that our Topic Closeness Measure (TCM) is an effective indicator of segmentation quality. We additionally explore the topic segmentation task when the input to the segmentation algorithm is the output of a speech recognizer. We demonstrate that information from speech recognition lattices can help improve topic segmentation over the one-best baseline.

2. Topic Modeling and Segmentation

Much of the recent work on topic analysis has been focused on generative topic models. Let $V = \{w_1, w_2, \ldots, w_n\}$ be the vocabulary of n words. Then an *observation* a is an observed set of text or speech expressed through the empirical frequency, or expected count, $C_a(w_i)$ for each $w_i \in V$. In generative topic models, a sequence of word observations is explained by a latent sequence of topic labels. As a result, high-dimensional text can be described with a low-dimensional mixture of the topics learned. A simple generative formulation topic model is

$$z = \arg\max_z \Pr(z|a) = \arg\max_z \Pr(a|z)\Pr(z), \quad (1)$$

where a is the sequence of observed text, and z is the topic label assigned. The second equality follows by Bayes' rule and the realization that the prior over the observations $\Pr(a)$ does not change with respect to topic. Under such topic models, text is labeled by decoding a maximum *a posteriori* sequence of topics accounting for the text. In these models, a is treated as a "bag of words," meaning the order of the words in the text or speech stream underlying a is generally not considered, merely the occurrence frequency of each word within w. In practice, a can be a sentence, a window of n words, an utterance, or a single word. In Latent Dirichlet Allocation (LDA) [1], the formulation of Equation 1 is used, but the distributions $\Pr(w|z)$ and $\Pr(z)$ are modeled as multinomial distributions with Dirichlet priors. Hidden Topic Markov Models (HTMMs) [2] use an HMM structure where each state corresponds to a topic z and an underlying topic model (such as LDA or n-gram), as in [3, 4].

Topic labeling algorithms are also topic segmentation algorithms because a topic assignment to a stream of text or speech also implies a topic-wise segmentation of the stream. Nevertheless, a number of efforts have been made to create algorithms specifically for the segmentation task. In TextTiling [5], word counts are computed for a sliding window over the input text. Text similarity is then evaluated between pairs of adjacent windows according to a cosine similarity measure, $\frac{\sum_{i=1}^n C_1(w_i)C_2(w_i)}{\sqrt{\sum_{i=1}^n C_1(w_i)^2 \sum_{i=1}^n C_2(w_i)^2}}$. The segmentation is obtained by thresholding this similarity function. In this approach, words that are naturally more prevalent in the corpus effectively receive a higher weight in the cosine score. One popular way to bypass this limitation is by using the term frequency–inverse document frequency (tf–idf) [6] to weight each word's contribution to the similarity score.

3. Measuring Topic Segmentation Quality

In order to evaluate the quality of topic segmentation algorithms, it is necessary to have a segmentation quality measure. Various measures have been proposed for measuring topic segmentation quality. The most popular measure is known as Co-occurrence Agreement Probability, or CoAP.

3.1. Co-Agreement Occurrence Probability

CoAP [7] is broadly defined as:

$$P_D(\text{ref}, \text{hyp}) = \sum_{1 \le i \le j \le n} D(i,j) \left(\delta_{\text{ref}}(i,j) \overline{\oplus} \delta_{\text{hyp}}(i,j)\right), \quad (2)$$

where $D(i,j)$ is a distance probability distribution over observations i, j; δ_{ref} and δ_{hyp} are indicator functions that are one if observations i and j are in the same topic in the reference and hypothesis segmentations, respectively; and $\overline{\oplus}$ is the exclusive NOR operation ("both or neither"). In practice, the choice of D is almost always the distribution with its mass placed entirely on one distance k. CoAP scoring is then reduced to a single fixed-size sliding window over the observations. This form of CoAP is often referred to in the literature as P_k. Various modifications of CoAP have been used in previous studies, including those assigning different weights to false positive and false negative segment boundaries (e.g., [8]).

In CoAP, every spurious or missing topic boundary is penalized equally without regard for the topics that it falsely separates or fails to correctly separate. For example, consider a segment with word distribution $z_{1,r}$ in the reference. Suppose that for a particular hypothesis segmentation, this reference chunk overlaps with two chunks with distributions $z_{1,h}$ and $z_{2,h}$. As a result, a spurious topic boundary would be detected and would be penalized by CoAP in the same way as any other boundary error. However, it is entirely possible that distributions $z_{1,h}$ and $z_{2,h}$ are both statistically very similar to $z_{1,r}$. Thus, this error should be penalized less than failing to separate $z_{1,h}$ and $z_{3,h}$, where $z_{3,h}$ is far in linguistic content from $z_{1,r}$.

Additionally, CoAP is dependent on the choice of window size k. Various heuristics exist for the choice of k. One idea used in previous work has been to set k such that the score for degenerate segmentations (e.g.., those that place every possible boundary or none at all) get a score of around 50%. This latter heuristic is the one used in the implementation of CoAP in this paper. Finally, by matching sentences i and j of Equation 2 between the reference and hypothesis, CoAP implicitly requires that the reference and the hypothesis segmentations be obtained by placing boundaries in the same stream of text, or at least two streams of text where sentence i in the reference corresponds exactly to sentence i in the hypothesis. However, when the hypothesis text is produced by a speech recognizer, the text may be different due to recognition errors, and might be broken differently into utterances and/or sentences than the reference. One way of handling this limitation used in previous work [8] and in this paper is to align the reference text with the hypothesis text temporally. This results in a rather significant mismatch between the measure used for the text case and the speech case.

3.2. New Topic Segmentation Quality Measure

We next describe our new Topic Closeness Measure (TCM), which overcomes the limitations just mentioned, and as we shall see in the experimental section, correlates with CoAP in empirical trials. To incorporate word content information into our topic segmentation quality measure, we need to quantify the similarity between chunks of text or speech. One rudimentary similarity function is the cosine distance between word frequencies. Alternatively, if the word frequencies are viewed as a probability distribution, a number of probability distance functions can be used, including the symmetrized relative entropy or KL-divergence. However, these distance measures are all limited in that they are based on evaluating the divergence in the frequency or probability of a given word between the two segments. For example, if the first segment being considered has many occurrences of "sport", then a segment making no mention of "sport" but mentioning "baseball" frequently would be assigned the same similarity score as a segment not mentioning anything relevant to sports at all.

Clearly, a measure of closeness between words is required. One powerful indicator of word similarity is co-occurrence in speech or text segments known to be topic-coherent. A measure that captures this intuition is *mutual information*. Let V be the vocabulary, and $x, y \in V$ be two words. If T is a large training corpus, then let $C_T(x,y)$, $C_T(x)$, and $C_T(y)$ be the empirical probabilities of x and y appearing together, and that of x and y appearing, in T, respectively. The pointwise mutual information (PMI) between x and y is then defined as $\text{PMI}(x,y) = \log \frac{C_T(x,y)}{C_T(x)C_T(y)}$. The definition of "appearing together" can be interpreted to mean proximity in the word stream [9]. However, since topic segmentation is our task, we assume that our training corpus T is pre-segmented into topic-coherent chunks, and we say that x and y appear together when they appear in the same chunk.

The logarithm in the PMI is customarily used due to connections with well-understood quantities in information theory, such as entropy. However, since logarithm is a monotone function, dropping it in the above formula does not change the ordering of word pairs and enables the similarity measure K_{norm} given below to be a positive definite symmetric kernel. Thus, our similarity between words (sometimes referred to as *interest*) shall be evaluated as

$$\text{sim}(x,y) = \frac{C_T(x,y)}{C_T(x)C_T(y)}. \quad (3)$$

Our goal is to design a segmentation quality measure that penalizes segments spanning multiple topics while rewarding segments that respect topic boundaries. In the following measure, we match segments between the reference and the hypothesis segmentation. The intuition is that those segments in the hypothesis that span multiple reference segments will likely get a low similarity score when compared to either reference segment, while hypothesis segments respecting reference segment boundaries will receive a high similarity score.

We will evaluate the total similarity of a pair of observations a and b as $K(a,b) = \sum_{w_1 \in a, w_2 \in b} C_a(w_1)\ C_b(w_2)\ \text{sim}(w_1, w_2)$. Let A and B be the column vectors of empirical word frequencies such that $A_i = C_a(w_i)$ and $B_i = C_b(w_i)$ for $i = 1, \ldots, n$. Let $\mathbf{K}$ be the matrix such that $\mathbf{K}_{i,j} = \text{sim}(w_i, w_j)$. The similarity score can then be written as a matrix operation, $K(a,b) = A^\top \mathbf{K} B$. We normalize to ensure that the score is in the range $[0,1]$ and that for any input, the self-similarity is 1,

$$K_{\text{norm}}(a,b) = \frac{A^\top \mathbf{K} B}{\sqrt{(A^\top \mathbf{K} A)(B^\top \mathbf{K} B)}}. \quad (4)$$

It can be shown that this general measure of text similarity is a positive definite symmetric (PDS) kernel, and thus it can be used in future discriminative learning for topic segmentation and labeling algorithms. However, in this paper our primary use for this similarity score is to create our segmentation quality measure. Let k and l be the number of segments in the reference and hypothesis segmentation, respectively. Additionally, let $R_1, \ldots, R_k$ and $H_1, \ldots, H_l$ be the normalized column count vectors of the segments in the reference and hypothesis segmentation, respectively. $Q(i,j)$ quantifies the overlap between the two segments i, j. In this work, $Q(i,j)$ is the indicator variable that is one when reference segment i overlaps with hypothesis segment j, and zero otherwise. However, various other functions can be used for O, such as the duration of the overlap or the number of overlapping sentences or utterances. Similarly, other similarity scoring functions can be incorporated

in place of K_{norm}. The TCM score between the reference segmentation R and the hypothesis segmentation H is defined as

$$\text{TCM}(R,H) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} Q(i,j) K_{\text{norm}}(r_i, h_j)}{\sum_{i=1}^{k}\sum_{j=1}^{l} Q(i,j)}. \quad (5)$$

Like CoAP, TCM is in the range $[0,1]$, and is symmetric in the sense the if the reference and hypothesis segmentations are exchanged the score is the same. Further, since TCM makes use of the general text content similarity measure of Equation 4, it considers not only where the topic boundaries lie but also the closeness of the content of the segments being separated by the boundaries. Additionally, the use of TCM is not dependent on the window size parameter k used in previous measures. TCM does consider the placement of topic boundaries, and accordingly, accomplishes the goal of CoAP – to penalize false positive and false negative segmentations. For example, adding a spurious boundary (i.e., one that separates two segments of the same topic) in a hypothesis segmentation would add one to l and would thus be penalized by the extra contribution to the normalization term $\sum_{i=1}^{k}\sum_{j=1}^{l} Q(i,j)$. Deleting a boundary between two different-topic segments is also penalized because the similarity score K_{norm} between the combined segment and the overlapping reference segments would be decreased.

4. Lattice-based Topic Analysis

Now that we have defined a general quality measure that applies to both speech and text topic segmentation algorithms, we explore the application of topic models to the output of a speech recognizer. There is a significant literature on topic analysis of spoken language (e.g., [3, 4]). However, the majority of the approaches use only the one-best recognition hypothesis as input to a topic labeling and/or segmentation algorithm. Since lattices carry more information than just the one-best hypothesis, we are interested in using them to improve the quality of these algorithms. A recent work [10] demonstrated an improvement using word and phoneme lattices for topic identification, or labeling pre-segmented utterances in isolation. In this work we focus exclusively on word lattices.

In our topic segmentation and labeling algorithms, we use two information sources derived from lattices, expected counts and confidence scores. Each word found in a lattice is associated with a total posterior probability, or expected count, accumulated over all the paths that contain that word. If V is the vocabulary the count of the word x according to a stochastic lattice automaton A is $C(x) = \sum_{u \in V^*} |u|_x [\![A]\!](u)$, where $|u|_x$ is the number of occurrences of word x in string u and $[\![A]\!](u)$ is the probability associated by A to string u. The set of expected counts for the words found in a lattice can be computed efficiently [11]. We also compute word-level confidence scores for the one-best hypothesis using a logistic regression classifier. The classifier takes two features as input, the first being the word expected counts just mentioned. The second feature is a likelihood ratio between the standard recognizer with full context-dependent acoustic models and a simple recognizer with context-independent models. Since the input to generative topic models is a sequence of bag-of-word observations to be labeled with topics, it is straightforward to incorporate lattice counts and confidence scores into the generative model as a prior weighting on the input word frequencies.

5. Experiments

We have applied HTMM to learn a topic model over the English speech portion of the TDT corpus of broadcast news speech [12]. In total, there were 447 news show recordings of 30-60 minutes per show, for a total corpus size of around

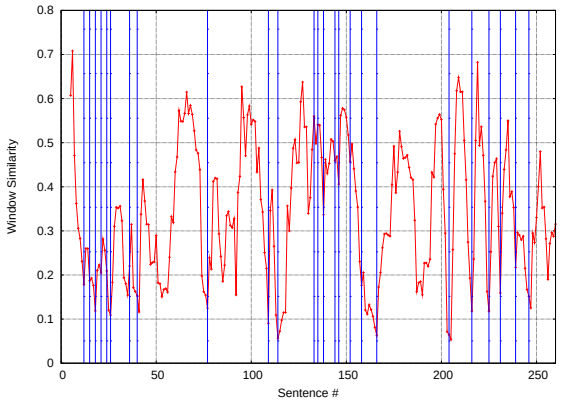

Figure 1: The window distance $K_{\text{norm}}(w_t, w_{t+\Delta t})$ for a representative show. The vertical lines are true story boundaries from the human-labeled corpus. A line at sentence t means that sentence $t+1$ is from a new story.

311 hours. For development and testing, we used 41 and 69 shows picked from the Voice of America English News Program (VOA_ENG) and MS-NBC News With Brian Williams (MNB_NBW), containing 957 and 1,674 stories, respectively. The 337 shows from other sources were used for training. The training shows contained 6,310 stories, and were annotated with human story segmentations and transcriptions for each story. Certain stories were also annotated by hand with topics such as "Earthquake in El Salvador," but these labels were not used in the model training. The HTMM was trained with 20 topics.

5.1. Text Similarity Evaluation

To evaluate our co-occurrence based similarity score empirically, we computed K_{norm} between all pairs of test and development stories with human topic labels. With 291 stories, there were 3,166 same-topic story pairs and 39,172 different-topic pairs in our experiment. The average pairwise similarity between between different-topic story pairs was 0.2558 and that between same-topic story pairs was 0.7138, or around 2.8 times greater. This indicates that our text similarity measure is a good indicator of topical similarity between two segments of text or speech.

The following experiment explores the correlation between K_{norm} and true segmentation boundaries. For our text test set, we processed each show's transcription by sliding a window of $\Delta t = 6$ sentences along the text, accumulating the word frequencies within each window. This value for Δt was selected to yield good performance on the development set of a new topic segmentation algorithm that is being developed in ongoing research. For each sentence t, let w_t be the window ending at sentence t. We computed the distance between all pairs $K_{\text{norm}}(w_t, w_{t+\Delta t})$ and plotted this distance. Figure 1 displays this plot for a representative show. As this figure illustrates, true topic boundaries are extremely well correlated with local minima in the similarity score. Similar trends are observed with other shows in the corpus.

5.2. Topic Segmentation Results

For our text-only experiments we used the human news show transcriptions. For the speech experiments, the audio for each show was first automatically segmented into utterances, while removing most non-speech audio, such as music and silence[13]. Each utterance was transcribed using the Google large-vocabulary continuous speech recognizer. This recognizer (the baseline system of [13]) used standard PLP cepstral fea-

Table 1: Topic segmentation quality as measured with CoAP and TCM.

Condition	CoAP	TCM	CoAP	TCM
	(Text Training)		(Speech Training)	
Text Random	50.4%	58.4%	-	-
Text Full	50.4%	51.8%	-	-
Text None	49.6%	56.2%	-	-
One-best Random	50.8%	48.8%	-	-
One-best Full	51.0%	43.0%	-	-
One-best None	49.1%	52.9%	-	-
Text	66.9%	72.6%	-	-
One-best	65.0%	61.5%	67.3%	62.8%
Counts	65.5%	62.4%	69.7%	64.1%
Confidence	68.3%	64.2%	68.8%	64.9%

tures, a vocabulary of about 71K words, GMM-based triphone HMM acoustic models, and smoothed 4-gram language models pruned to about 8M n-grams. Both the acoustic and language models were trained on standard Broadcast News (BN) corpora. The word error rate of this recognizer on the 1997 BN evaluation set was 17.7%. The vocabulary for the HTMM algorithm consisted of a subset of $8{,}821$ words. This was constructed by starting with the set of words seen in the recognizer transcription of the training data, applying Porter stemming [14], removing a stoplist of function and other words not likely to indicate any topic, and keeping only those words occurring more than five times. Since our topic model EM training algorithm begins with random values, we ran 20 trials of model training and testing and picked the model that had the best performance on the development data set.

The results of the experiment are given in Table 1. We trained two separate HTMM topic models, the first using the reference text as training data (Text Training), and the second using the one-best transcription of the training data (Speech Training). We tested on the reference text (Text), as well as three different varieties of speech transcriptions, transcriptions only (One-best), and speech transcriptions weighted with lattice counts (Counts) and confidence scores (Confidence). The first six rows give scores for degenerate segmentations with random segment boundaries (Random), all possible boundaries (Full), and no boundaries at all (None).

These results show that TCM is an effective measure of topic segmentation quality. Qualitatively, its output is generally correlated with that of CoAP. Segmentations produced by the topic model significantly outperform degenerate segmentations by both measures. Lattice counts yield a 2.3% and 3.5% relative improvement with text and speech training, respectively, in TCM error compared to the one-best baseline, and 1.4% and 7.3% in terms of CoAP. Confidence scores yield even larger improvements with both measures, 9.4% and 4.6% relative by CoAP and 7.0% and 5.6% by TCM. One interesting comparison to make is that between the Text case and the One-best case. Certainly we can expect topic segmentation on the reference transcriptions to be a much easier task than that on the output of a speech recognizer, due to the transcription errors present in the latter. Indeed, error reductions from One-best to Text are achieved, but 5.4% as measured by CoAP, and 28.8% by TCM. This asymmetry can possibly be attributed to the mismatch between the CoAP used for text and that used for speech mentioned in Section 3.1.

6. Conclusion and Future Work

In this paper, we have made several contributions to topic analysis of spoken language. The first is to give a new measure of topic segmentation quality that overcomes major limitations of past evaluation techniques. Unlike previous quality measures, TCM applies generally to either speech or text sources, does not depend on a fixed window size, and considers similarity between segments labeled as topic-coherent, rather than simply the presence or absence of a segment boundary in the same places as in the reference. In empirical trials, TCM is correlated with the previous measures. Additionally, the general text similarity measure underlying TCM is empirically correlated with the ground truth topic boundaries and topic labels. We have also demonstrated that a topic segmentation and identification algorithm can be improved by using lattice information.

We are currently working on a topic segmentation algorithm that explicitly attempts to maximize TCM by placing topic boundaries at points in the observation stream where text similarity is low. We believe that such an algorithm will outperform the ones used in the present work.

7. Acknowledgments

The authors would like to thank Christopher Alberti, Michiel Bacchiani, Eugene Ie, Hank Liao, Ashok Popat, Olivier Siohan, and all the members of the Google speech recognition group for their help and advice.

8. References

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.

[2] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Hidden topic markov models," in *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, San Juan, Puerto Rico, 2007.

[3] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, "Event tracking and text segmentation via hidden markov models," in *Proceedings of ASRU*, 1998.

[4] D. M. Blei and P. J. Moreno, "Topic segmentation with an aspect hidden markov model," in *Proceedings of SIGIR*, 2001.

[5] M. A. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[6] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[7] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, no. 1-3, pp. 177–210, 1999.

[8] NIST, "Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan," http://www.nist.gov/speech/tests/tdt/1998/doc/tdt2.eval.plan.98.v3.5.pdf, 1998.

[9] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[10] T. J. Hazen and A. Margolis, "Discriminative feature weighting using MCE training for topic identification of spoken audio recordings," in *Proceedings of ICASSP*, Las Vegas, Nevada, 2008.

[11] C. Allauzen, M. Mohri, and B. Roark, "Generalized algorithms for constructing statistical language models," in *Proceedings of ACL*, 2003, pp. 40–47.

[12] J. Kong and D. Graff, "TDT4 Multilingual Broadcast News Speech Corpus," http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005S11, 2005.

[13] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proceedings of ICASSP*, Taipei, Taiwan, 2009.

[14] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.