# UNSUPERVISED TOPIC MODEL FOR BROADCAST PROGRAM SEGMENTATION

*Gilles Boulianne*[⋆] *Pierre Dumouchel* [†]

[⋆] Centre de recherche informatique de Montréal (CRIM), Canada
[†] École de Technologie Supérieure, Montréal, Canada

## ABSTRACT

Several unsupervised methods have been proposed to segment a continuous text stream into individual topics. A simple HMM formulation of the most successful of these methods exposes their underlying assumptions and suggests the use of a new prior for segmentation probability. Under this formulation, we explore the space of possible modeling choices on databases of English and French TV and radio programs. We show that the proposed prior improves segmentation results and can also accommodate additional knowledge sources within the HMM efficient dynamic programming.

***Index Terms*—** Story segmentation, Bayesian methods, Graphical models, HMM, Topic models

## 1. INTRODUCTION

As audio data becomes increasingly available, the need for indexing it grows accordingly. Even though transcripts in the form of closed-captions, subtitles, or automatic speech recognition may be available, they are usually not structured, i.e. they consists in a single stream of text without individual topic or story markers. The task we examine here is to segment a text into individual semantically coherent units, such as topics or stories, into a linear sequence, for further processing or as a way to present search results to a user in reasonable chunks. In the context of TV and radio programming, unsupervised methods are clearly preferable, because of the constantly changing nature of contents, unpredictability of topics, and addition of new sources.

Approaches for segmenting text into topics fall in four broad categories [1] : methods based on lexical dissimilarity, such as *LCSeg* [2], on lexical cohesion, such as *C99* [3], on discriminative boundary detection with classifiers [4], or generative models [5][6][7][8][9].

Generative models assume that text is generated from an underlying sequence of topics. Each topic is associated with a particular probability distribution on words, which is sampled to generate words. The problem is then, given the sequence of words in a text, to infer the underlying sequence of topics. Hidden Markov models provide an efficient solution for both the inference and segmentation problems. However, the basic *HMM* approach used by [5], as well as its latent concept version *aspect HMM* [6], are entirely supervised and require an annotated corpus for training. *LDA-HMM* is a Bayesian version [7] which becomes entirely unsupervised. *TextSeg* is another generative approach that is completely unsupervised and has been proposed by [8] but does not rely on underlying topics : instead, a language model is inferred for each individual segment. To solve the problem of inferring language models based on short segments, [9] presented a Bayesian version of *TextSeg*. Currently [8] and [9] represent the most successful unsupervised models when compared on standard databases, either for manual or automatic text transcriptions [1].

In an effort to compare and analyze the assumptions behind these two models, we show that when both are formulated as a unified segmental HMM model, they differ only in their choice of output probability model and prior segmentation model. We propose a new prior for the segmentation probability and efficiently include cue word information in the segmentation model. The last section concludes with experimental results on TV and radio programs, in English and French, which explore the space of possible modeling choices, including the two original methods of [8] and [9] as special cases.

## 2. TOPIC SEGMENTATION MODEL

In the following we present a particular version of an HMM tailored to the topic segmentation task. It slightly differs from the common HMM formulation in that it models whole segments instead of individual words or sentences.

Given a text to be segmented, made of sentences $t = 1, \ldots, T$, assign a number to each possible segment boundary position (gap) so that gap $t$ falls between sentences $t$ and $t+1$. Thus gap 0 will be placed before the first sentence and gap $T$ after the last sentence. The corresponding topology for an HMM is shown in figure 1 :

- A state corresponds to a gap $t = 0, \ldots T$
- A transition between state $t'$ and state $t$ corresponds to a segment covering sentences $t' + 1$ to $t$.
- A transition (or segment) $j$ emits words $\mathbf{W}_{t'+1} \ldots \mathbf{W}_t$ from a distribution $\Theta_j$, where $\mathbf{W}_t$ is the set of words $\{W_{t,n}\}$ in sentence $t$.

This work was funded in part by NSERC Grant number EGP 417255-11.

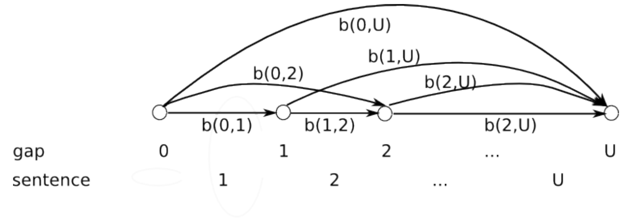

**Fig. 1**. HMM topology and recursive computation of $B(t)$.

Note that this topology is slightly different from a usual HMM in that a transition does not correspond to a single word or sentence, but rather to a sequence of sentences belonging to a single segment.

Let's define an indicator variable $z_t$ such that if sentence $t$ is in segment $j$, we have $z_t = j$. Thus the HMM output probability for segment $j$ will be :

$$p(\mathbf{W}_{t'+1\ldots t}|z_{t'+1\ldots t} = j, \Theta_j) \quad (1)$$

and the transition probability for segment $j$ will be :

$$p(z_{t'+1\ldots t} = j) \quad (2)$$

Under a bag-of-words assumption which ignores the word ordering in the segment, equation 1 can be further simplified to :

$$p(\mathbf{W}_{t'+1\ldots t}|z_{t'+1\ldots t} = j, \Theta_j) = p(\{\mathbf{W}_{t:z_t=j}\}|\Theta_j) \quad (3)$$

The optimal segmentation is the one which maximizes the joint likelihood :

$$p(\mathbf{W}, \mathbf{z}|\boldsymbol{\Theta}) = p(\mathbf{W}|\mathbf{z}, \boldsymbol{\Theta})p(\mathbf{z}) \quad (4)$$

where $\mathbf{W}$ is the set of all words over all $T$ sentences, $\mathbf{z}$ is the vector $z_{1\ldots T}$ which assigns a segment to each sentence, and $\boldsymbol{\Theta}$ is the set of language models $\{\Theta_{z_t}\}$. The last term in the right-hand side is the prior segmentation probability.

The well-known recursive solution for equation 4 allows an efficient dynamic programming implementation. Consider a segment $j$ between gap $t'$ and gap $t$, as illustrated in figure 1. This segment's contribution to the total likelihood will be :

$$b(t', t) = p(\{\mathbf{W}_{t:z_t=j}\}|\Theta_j)p(z_{t'+1\ldots t} = j) \quad (5)$$

The objective function $B(t)$ is the maximum likelihood segmentation over all segmentations ending at gap $t$ and has the recursive expression :

$$B(t) = \max_{t'<t} B(t')b(t', t) \quad (6)$$

Together, equations 5 and 6 allow us to recursively compute $B(t)$, from $t = 0$ to $t = T$, starting at $t = 0$ with initial value $B(0) = 1$.

## 2.1. Relationship to previous models

[8] directly formulates his model as an HMM with a topology defined as in the previous section. [9] starts from a model more directly related to latent topic models, but introduces a topic switching variable with a constraint which imposes a linear segmentation, with a single, different topic for each segment. Both models use the same recursive equations 5 and 6 to compute the best segmentation. However, each makes use of different values and assumptions for the output probability $p(\mathbf{W}|\mathbf{z}, \boldsymbol{\Theta})$ and prior segmentation probability $p(\mathbf{z})$. In the more general framework presented here, we can better analyze their choices and suggest new solutions.

### *2.1.1. Segmental output probabilities*

Equation 3 corresponds to a language model for each segment. [8] assumes a language model where the word probability is based on the relative word frequency in the segment, and uses Laplacian smoothing. His expression for $p(\{\mathbf{W}_{t:z_t=j}\}|\Theta_j)$ is simply :

$$p_{lm}(\{\mathbf{W}_{t:z_t=j}\}) = \prod_{i=1}^{V} \frac{c_{j,i} + 1}{\sum_{i=1}^{V} c_{j,i} + V}$$

where $c_{j,i}$ is the number of occurrences of word $i$ in segment $j$, $C_j = \sum_{i=1}^{V} c_{j,i}$ is the total number of words in the segment, and $V$ is the vocabulary size.

Another natural choice for the output probability distributions $\{\Theta_{z_t}\}$ is a multinomial language model $W_t \sim \text{Multinomial}(\boldsymbol{\beta}_{z_t})$. Using a point estimate for $\{\boldsymbol{\beta}_{\mathbf{z_t}}\}$ results in :

$$p_{mult}(\{\mathbf{W}_{t:z_t=j}\}|\beta_0) = \prod_{i=1}^{V} \frac{c_{j,i} + \beta_0}{\sum_{i=1}^{V} c_{j,i} + V\beta_0}$$

As pointed out by [9], Laplacian smoothing as used in [8] is a special case of this expression, with $\beta_0 = 1$.

Instead of using a point estimate, [9] proposes a Bayesian approach to marginalize over all possible language models, assuming a Dirichlet prior for the multinomial. The corresponding Dirichlet compound multinomial distribution is conditioned on a prior parameter $\beta_0$ :

$$p_{dcm}(\{\mathbf{W}_{t:z_t=j}\}|\beta_0) = \frac{\Gamma(V\beta_0)}{\Gamma(C_j + V\beta_0)} \prod_{i=1}^{V} \frac{\Gamma(c_{j,i} + \beta_0)}{\Gamma(\beta_0)}$$

Note that in the equation given here, the numerator inside the product term differs from the numerator given in [9].

### *2.1.2. Segmental transition probabilities*

For the prior segmentation distribution $p(\mathbf{z})$ in equation 4, [9] simply assumes that all valid segmentations have probability one, i.e. $p(\mathbf{z}) = 1$, and ignores the term. As this tends to produce stretches of unreasonably small segments, a more recent work [10] introduces a prior probability based on the expected segment duration. [8] has also introduced a segment penalty corresponding to $p(z_{t'+1...t}) = \frac{1}{N}$ where $N$ is the number of words in the segment. Complexity theory is invoked to justify this value, and the suggested penalty reflects the amount of bits needed to encode a segmentation.

Assuming an ergodic HMM model, we propose here a simple choice for a non-informative prior distribution on $p(\mathbf{z})$, ignoring the left-to-right structure of figure 1 for simplification. We allow a transition from every node to every other node with equal probability, thus the probability that a segment starts at $t'$ and ends at $t$ is constant $p(z_{t'+1...t}) = \frac{1}{T}$. Note that this prior is determined solely by the number of sentences in the text. It is not a parameter of the model to be set externally, such as the expected segment duration parameters $\mathbf{d}$ proposed in [10].

More informative priors $p(\mathbf{z})$ can also be considered, taking account of other information about the segmentation, with the advantage of avoiding hard decisions, since decisions consider both the current data and the prior evidence. Any prior probability that answers the Markov property, i.e. can be computed by considering only the segment and its starting point, can be used in the recursive solution.

Examples of suitable informative priors include information about pauses or prosody that provide a probability of a boundary at certain gaps. Another example is the use of linguistic markers. We propose to use cue words which can signal a change between topics (such as "welcome", "thank you", "this was ..."). This was implemented by incrementing the segment prior probability whenever its first or last sentence contains one of these markers. Since this change is only dependent upon the current segment's content, it can be easily incorporated into the dynamic programming search.

The use of prior $p(\mathbf{z})$ to model cue words contrasts with the approach of [9] in which output probabilities for cue words are obtained with a distribution that is shared across topics, which renders dynamic programming inference inapplicable.

## 3. EXPERIMENTS

Table 2 summarizes the corpora used in the experiments. TDT4 TV is an extract of the standard TDT4 NIST corpus containing only TV broadcast news and is used here to provide a reference for comparison with other published work on the same database. CD is an in-house Canadian database which includes TV and radio programming, covering a variety of subjects, from broadcast news to talk shows. CD EN contains texts in English language, and CD FR in French. Texts are uncorrected closed-captions as they were captured with each program. To provide the ground-truth, all CD data was manually segmented into topics by human experts, librarians specialized in searching newspapers, newswires and TV and radio transcripts.

| Corpus | Sentences | Shows | Stories |
|---|---|---|---|
| TDT4 TV | 69013 | 276 | 7026 |
| CD EN | 17744 | 62 | 1124 |
| CD FR | 8629 | 36 | 559 |

**Table 2**. Corpora used in experiments.

The experiments consisted in running an implementation of the proposed HMM segmenter on each of the three corpora, for the alternative output probability distributions from section 2.1.1, and the segmentation probability distributions from section 2.1.2. These combinations include the models originally proposed in [8] and [9]. We checked that our implementation results for [8] matched exactly the results from the author's publicly available code on the same data.

For cue word experiments, 11 French and 7 English cue words were manually selected from different transcribed shows, from broadcasters that do not appear in the test corpora.

### 3.1. Performance metrics

Commonly used performance metrics include FA/FR, $WD$ and $P_k$ [1], but $P_k$ has been the most popular measure in the past and there is an abundant litterature providing values of $P_k$ for various systems. $P_k$ measures the average probability of segmentation error. It uses a a window of $k$ sentences that is slided over the text, at each step checking if the hypothesized segmentation is correct about the two ends of the window belonging (or not) to the same segment.

The value of $k$ used for measuring $P_k$ was chosen as the one which produced scores as close as possible to $P_k = 0.5$ for trivial segmentations placing boundaries at every gap, or without any boundaries, or with boundaries placed at random (with equal probability at each gap equal to the inverse average reference segment length). Thus one value of $k$ was determined for each reference segmentation using solely the reference.

Although many limitations of $P_k$ have been pointed out in past studies and alternative measures have been proposed, here we encountered a major problem with segmentations that can contain segments of unrealistic length while having a good $P_k$. This was also observed in [10] where the gap between $P_k$ and $WD$ metrics was explained by the generation of spurious short segments. In an effort to quantize this phenomenon, we provide an additional measure of the quality of the segmentation, the Pearson statistic $X^2$ for the segment length, which indicates how different the segment length distributions are in the hypothesized and reference

| | | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|---|
| Transition probability | | $\frac{1}{N}$ | $\frac{1}{N}$ | $\frac{1}{T}$ | $\frac{1}{T}$ | 1 | $\frac{1}{T}$ | $\frac{1}{T} \cdot p(cue)$ |
| Output probability | | $p_{lm}$ | $p_{mult}$ | $p_{lm}$ | $p_{mult}$ | $p_{dcm}$ | $p_{dcm}$ | $p_{dcm}$ |
| TDT4 TV | $\beta_0$ | 1.0 | 0.8 | 1.0 | 0.8 | 0.6 | 0.2 | 0.2 |
| | $P_k$ | 0.25 | 0.24 | 0.23 | 0.23 | **0.20** | 0.21 | **0.20** |
| | $X^2$ | 1.86 | 1.70 | 1.62 | 1.47 | 1.28 | 0.99 | **0.80** |
| CD EN | $\beta_0$ | 1.0 | 1.3 | 1.0 | 1.4 | 0.7 | 0.4 | 0.4 |
| | $P_k$ | 0.30 | 0.27 | 0.29 | 0.26 | **0.22** | 0.24 | 0.25 |
| | $X^2$ | 2.34 | 2.61 | 1.89 | 2.37 | 2.26 | 1.80 | **1.16** |
| CD FR | $\beta_0$ | 1.0 | 1.5 | 1.0 | 1.7 | 1.2 | 0.6 | 0.7 |
| | $P_k$ | 0.28 | 0.22 | 0.29 | 0.21 | **0.17** | 0.19 | 0.18 |
| | $X^2$ | 2.10 | 2.21 | 1.77 | 2.42 | 3.11 | 1.83 | **1.34** |

**Table 1**. Topic segmentation results according to choices of model output and transition probabilities.

segmentations :

$$X^2 = \sum_{i=1}^{n} \frac{(S_i - R_i)^2}{R_i}$$

where $S_i$ and $R_i$ are the relative frequencies of segment length $i$ for the hypothesized segmentation and reference segmentation, respectively. Segment length is measured as the number of sentences in a segment. More similar hypothesized and reference distributions means $X^2$ closer to zero.

### 3.2. Results

Table 1 presents the results obtained on the three corpora, each column corresponding to a particular combination of output probability model and prior segmentation model. For each corpora, the best value of $P_k$ and $X^2$ is boldfaced (smaller values indicate better results). The value of $\beta_0$ given in the table was determined by an oracle, i.e. chosen for producing the best $P_k$ on the test set. In a realistic test scenario $\beta_0$ would have to determined by another method such as cross validation, but here the $\beta_0$ selection method is comparable for all models. Column A corresponds to the original *TextSeg* [8] model, and column E to the model of [9]. The model of [10] requires an additional parameter $\mathbf{d}$ that controls the granularity of segmentation and has to be set to the expected segment duration, and is not compared here.

It first appears that column E is best, as it consistently gives lower values of $P_k$. However corresponding $X^2$ values are higher than for other $p_{dcm}$ models, meaning that its segment length distribution is farther from the reference. Overall, column F and G provide low values both for $P_k$ and $X^2$.

As a choice for output probabilities, $p_{dcm}$ outperforms $p_{mult}$ which itself outperforms $p_{lm}$. This is clear when comparing columns C, D, and F, across all corpora (for the same segmentation prior). The proposed prior $\frac{1}{T}$ slightly but consistently outperforms the complexity penalty $\frac{1}{N}$ of [8], but when used in conjunction with $p_{dcm}$, it improves $X^2$ but not $P_k$.

Finally, column G shows that including other knowledge sources such as cue words can be effective ; while not providing significant improvements on $P_k$, it provides the best segmentations in terms of length distribution relative to the reference, as measured by $X^2$.

## 4. CONCLUSION

In this work, we formulated a unified segmental model which includes both [9] and [8] as special cases, and showed how they differ in their choice of output probability model and prior segmentation model. We proposed a new prior which preserves the low segmentation error rate of both methods but improves quality in terms of segment length distribution. This new prior also provides a mechanism to include other sources of knowledge, such as cue words, within the HMM dynamic programming algorithm, thus overcoming a major limitation of [9]. The proposed non-informative prior generally provides a lower error rate, and when linguistic cues are incorporated, it obtains the best results in terms of combined segment error rate and segment length distribution, and can be used directly in the HMM model without compromising its efficiency.

## 5. REFERENCES

[1] M. Purver, *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, chapter 11 : Topic Segmentation, pp. 219–317, Wiley, 2011.

[2] M. Galley, K. McKeown, and E. Fosler-Lussier, “Discourse Segmentation of Multi-Party Conversation,” *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, pp. 562–569, 2003.

[3] F.Y.Y. Choi, “Advances in domain independent linear text segmentation,” *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pp. 26–33, 2000.

[4] D. Beeferman and A. Berger, “Statistical Models for Text Segmentation,” *Machine Learning*, vol. 34, no. 1, pp. 177–210, 1999.

[5] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt, “A hidden Markov model approach to text segmentation and event tracking,” *Proceedings ICASSP 1998*, vol. 1, pp. 333–336 vol. 1, 1998.

[6] D. M. Blei and P.J. Moreno, “Topic segmentation with an aspect hidden Markov model,” *Proceedings of the 24th annual international ACM SIGIR*, pp. 343–348, 2001.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[8] M. Utiyama and H. Isahara, “A statistical model for domain-independent text segmentation,” *Proceedings ACL 2001*, pp. 491–498, 2001.

[9] J. Eisenstein and R. Barzilay, “Bayesian unsupervised topic segmentation,” *Proceedings of EMNLP*, pp. 334–343, 2008.

[10] J Eisenstein, “Hierarchical text segmentation from multi-scale lexical cohesion,” *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 353–361, 2009.