BeamSeg: a Joint Model for Multi-Document Segmentation and Topic Identification

Pedro Mota
Carnegie Mellon University
Pittsburgh, PA, USA
INESC-ID
Instituto Superior Técnico
Lisboa, Portugal
pjdrm@ist.utl.pt

Maxine Eskenazi
Carnegie Mellon University
Pittsburgh, PA, USA
max@cs.cmu.edu

Luísa Coheur
INESC-ID
Instituto Superior Técnico
Lisboa, Portugal
lcoheur@l2f.inesc-id.pt

Abstract

We propose BeamSeg, a joint model for segmentation and topic identification of documents from the same domain. The model assumes that lexical cohesion can be observed across documents, meaning that segments describing the same topic use a similar lexical distribution over the vocabulary. The model implements lexical cohesion in an unsupervised Bayesian setting by drawing from the same language model segments with the same topic. Contrary to previous approaches, we assume that language models are not independent, since the vocabulary changes in consecutive segments are expected to be smooth and not abrupt. We achieve this by using a dynamic Dirichlet prior that takes into account data contributions from other topics. BeamSeg also models segment length properties of documents based on modality (textbooks, slides, *etc.*). The evaluation is carried out in three datasets. In two of them, improvements of up to 4.8% and 7.3% are obtained in the segmentation and topic identifications tasks, indicating that both tasks should be jointly modeled.

1 Introduction

Documents exhibit a content organization that aggregates related text passages in topically coherent segments. Understanding the document structure at the segment level enables efficient content navigation. This has become more relevant with the number of available documents on the Web. The current information landscape allows access to documents describing the same subject, providing alternative views or complementary information. This is advantageous in a variety of scenarios. For example, students have at their disposal several learning materials and might need to find a particular topic segment that best suits their learning needs. Finding such documents is an easy task since search engines are capable of returning documents conveying related information. However, if search engines are effective in retrieving these documents, the task of putting them into a coherent picture remains a challenge (Shahaf et al., 2012). Automatically finding document segments – text segmentation – and identifying which ones discuss the same topic – topic identification – addresses this issue (Jeong and Titov, 2010).

Text segmentation and topic identification have been used as intermediate steps in a variety of natural language processing tasks, including summarization (Radev et al., 2004), opinion mining (Murakami et al., 2009), semantic and information retrieval (Purver, 2011; Amoualian et al., 2017). The improvements they brought spurred research in text segmentation. Invariably, all works resort to the lexical cohesion theory (Halliday and Hasan, 1976), which postulates that discourse structure is correlated to the use of cohesive vocabulary. Thus, segments can be identified by detecting vocabulary changes. Most approaches either consider segmentation and identification separately and/or do not take into account all documents in the dataset (single-document approach). Recently, some works studied these phenomena in collections of related documents (Jeong and Titov, 2010; Mota et al., 2016). These multi-document models assume that documents describing the same topic have similar lexical cohesion properties; an example of this phenomenon with similar segments but in different documents is depicted in Figure 1. Thus, better likelihood estimations can be obtained if all documents are taken into account (Mota et al., 2016). In this work, we expand the multi-document lexical cohesion idea by hypothesizing that vocabulary relationships between different segments exist. For example, if a word is heavily used in one segment, it is likely that it continues to appear in the following one, though less frequently. Modeling such interactions can lever-

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 582–592
Hong Kong, China, November 3-4, 2019. ©2019 Association for Computational Linguistics

age topic segmentation algorithms. We also explore the role of modality in the multi-document scenario. Previous approaches treat all documents equally, but it is plausible that we can improve segmentation by making assumptions about the expected segment length on a document modality basis. For example, segments in slide presentations are expected to be shorter than in video lectures.

We propose BeamSeg, a Bayesian unsupervised topic modeling approach to breaking documents in coherent segments while identifying similar topics. The generative process assumes that segments can share the same topic and, consequently, are generated from the same lexical distribution. Lexical cohesion is achieved by having higher segmentation likelihoods when the probability mass is concentrated in a narrow subset of words. This is in the same spirit as topic modeling approaches such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), but here the inherent topics are constrained to the linear discourse structure. To model interactions between lexical distributions, we use a dynamic prior, which assumes that the word probabilities change smoothly across topics. To model segment length characteristics, we assign prior variables conditioned on document modality.

The linear segmentation constraint has been used to make inference tractable by exhaustively exploring the segmentation space to obtain the exact maximum-likelihood estimation (Eisenstein and Barzilay, 2008). Given a multi-document setting, this is not feasible, as segments can share topics. We address this issue using a beam search algorithm, which allows the inference procedure to recover from early mistakes. In our experiments, we show that BeamSeg is able to perform well when segmenting learning materials, where previously single-document models obtained better results (Mota et al., 2018). We also observe that topic identification is more accurately determined in a joint model, as opposed to a pipeline approach (performing the tasks sequentially), indicating that both problems should be modeled simultaneously.

We summarize our contributions as follows:

- A novel joint model for topic segmentation and identification with a dynamic prior.
- An inference procedure based on a beam search algorithm.
- A study on how different modality-based segment length priors influence segmentation.

The source code is available in the following repository: `github.com/pjdrm/BeamSeg`.

2 Related Work

Following the lexical cohesion theory, segmentation algorithms identify spans of text with prominent vocabulary changes. The main difference between algorithms is how lexical cohesion is implemented: some resort to lexical similarity; the remaining follow a probabilistic approach.

Lexical approaches rely on a similarity metric between sentences, usually the cosine. A classic method is TextTiling (Hearst, 1997), which assumes that topic boundaries are found in consecutive sentences with a low similarity value; several other works built on this idea (Galley et al., 2003; Balagopalan et al., 2012). C99 (Choi, 2000) is another lexical approach, and uses a similarity matrix in a divisive clustering to obtain segments. MinCut (Malioutov and Barzilay, 2006) casts segmentation in a minimum cut graph partitioning problem. The graph has a node for each sentence; edges are weighted using lexical similarity. Long-distance textual relationships are modeled by connecting all sentences. Affinity Propagation Segmentation (Kazantseva and Szpakowicz, 2011) also models such relationships but uses affinity propagation clustering (Frey and Dueck, 2007). The algorithm creates a factor graph and maximizes the segment similarity sum function. Alemi and Ginsparg (2015) proposed the Content Vector Segmentation (CVS) sentence vector representation based on segment word embeddings. Using this representation in C99 improves bag-of-words results.

In another line of research, Wang et al. (2017) combined learning to rank and a convolutional neural network to learn a coherence function between text pairs; higher-ranked pairs are likely to be segments. Despite a promising approach, state-of-the-art results were not achieved. Also following an approach using neural networks, is the SECTOR algorithm (Arnold et al., 2019), which uses a topic embedding trained based on utterance topic classification. Following the network architecture from (Koshorek et al., 2018), two stacked LSTM layers are used to decode word embedding representation of utterances. To recover segmentation, a TextTiling approach is applied to the topic embedding layer. The evaluation results show that SECTOR is able to improve a C99 baseline.

Just as we introduced average **velocity** we will now describe **average acceleration**. Notice when **velocity** changes ... over **time**. And ... introduce an **average acceleration** ... The **average acceleration** between **time** t2 ... And the dimension ... secs per **time** squared.

Acceleration We say ... changing **velocity** are "accelerating" **Acceleration** is the "Rate of change of **velocity**" You hit the accelerator to speed up ... it's true you also hit ... friction is slowing ... **Average acceleration** Unit of **acceleration**: (m/s)/s=m/s2

The **acceleration** of a particle ... rate of change of **velocity** ... **time**. Average acceleration ... is v2 - v1 t2 - t1 ... **Acceleration** may be positive, negative or zero. Zero **acceleration** means we have constant **velocity**. Note that the direction and **acceleration** need not coincide.

Figure 1: Examples of segment excerpts from video, slide presentation, and PDF documents describing the accelaration topic. Words in bold depict shared vocabulary across segments.

Probabilistic approaches to segmentation follow a setup similar to the LDA model: words are assigned to topics such that probability mass is distributed on a small set of topically relevant words. In order to adapt this idea to segmentation, the model needs to be able to determine if sentences belong to the same topic (or mixture of topics). An example of such adaptation is the single-document segmentation model PLDA (Purver et al., 2006), where topic proportions are shared by sentences within the same segment. Segmentation is then determined through a binary topic shift sentence variable. Models such as TopicTiling (Riedl and Biemann, 2012), Structured Topic Model (STM) (Du et al., 2013), and NTSeg (Jameel and Lam, 2013) extend this LDA-based approach to segmentation. In all these approaches, topic identification is not possible since all segments are a mixture of topics.

In this paper, we adopt a probabilistic multi-document view on segmentation. Only two other models follow this approach: MultiSeg (Jeong and Titov, 2010) and Bayesseg-MD (Mota et al., 2016). MultiSeg uses a two-level LDA model where documents are generated using local and global topics. Local topics are specific to a document; global topics are shared between documents. Documents are mixtures of topics, but each segment is generated by a single topic, lending itself to a joint model of segmentation and topic identification. The multi-document aspect of the model stems from topic proportions being inferred from the whole dataset. In the experiments, this joint modeling outperforms a pipeline strategy that performs these tasks sequentially.

The other multi-document model, Bayesseg-MD, is an extension of Bayesseg (Eisenstein and Barzilay, 2008). In Bayesseg, sentences from the same segment are assigned the same topic. The inference procedure affords an exact maximum-likelihood estimation by exploring the segmentation space with a dynamic programming algorithm. This approach cannot be applied to multi-document segmentation since the hidden topic variables are integrated out; other single-document models following this approach also have this problem (Eisenstein, 2009; Malmasi et al., 2017). Bayesseg-MD sidesteps this problem by using lexically similar sentences from other documents. The word counts of such sentences are added to the segment likelihood estimation to reduce data sparseness. Despite using all documents for segment likelihood estimations, topic identification is not available. In this paper, we address these issues by designing an inference algorithm that explicitly tracks segment topic assignments.

3 BeamSeg Model

We implement lexical cohesion in a Bayesian setting in a generative process where segments with the same topic are drawn from the same multinomial language model. Thus, all u utterances with a topic k have their bag-of-words representation $\mathbf{x}_u$ drawn from language model ϕ_{z_u}; z_u is the hidden topic variable of u. We constrain the model to yield linear segmentations by having topics occurring at most once per document. This induces higher likelihood segmentations to have language models concentrating probability mass on a small subset of the vocabulary. Conversely, low likelihood segmentations spread the probability mass on a broad set of words. This modeling behavior is attuned to the lexical cohesion theory. Multi-document segmentation emerges by assuming that topics are shared across documents.

During inference, we want to find the hidden set of language models Φ and the topic vector assignment $\mathbf{z}$ that maximize the likelihood of the joint distribution of the model. Since we only care about segmentation, this process can be simplified by analytically marginalizing out the hidden language models Φ. This enables search to be carried out only in the segmentation space. Therefore, inference amounts to finding the segmentation $\hat{\mathbf{z}} = \operatorname{argmax}_{\mathbf{z}} p(\mathbf{X}|\mathbf{z})p(\mathbf{z})$. Using the marginalized joint likelihood, an approximation of $\hat{\mathbf{z}}$ is obtained using a beam search algorithm.

3.1 Language Models

Using the previous setup, we define the joint likelihood as follows:

$$p(\mathbf{X}|\mathbf{z}, \Phi) = \prod_{k}^{K} p(\phi_k|\beta) \prod_{u}^{U} p(\mathbf{x}_u|\phi_{z_u}), \quad (1)$$

where $\mathbf{X}$ is the set of all U utterances in the dataset; K is the number language models; and β are the Dirichlet prior parameters from which Φ is drawn.

The marginalization process is performed by appealing to the conjugacy between multinomial language models and the Dirichlet prior. This allows the conjugate Dirichlet distribution to integrate to one, leaving the marginalized joint likelihood expression with the normalizing constants:

$$\begin{aligned} p(\mathbf{X}|\mathbf{z}) &= \int p(\mathbf{X}|\mathbf{z}, \Phi)p(\Phi|\beta)d\Phi \qquad (2) \\ &= \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(n_{U,w}^k + \beta)}{\Gamma(n_U^k + \beta)}, \end{aligned}$$

where W is the vocabulary set; $n_{U,w}^k$ is number of times word w is assigned topic k in all U utterances of the document collection; n_U^k is number of times topic k appears in U; and the symbol Γ refers to the Gamma function. The resulting expression in Equation 2 corresponds to the product of the individual topic likelihoods, comprised of segments from different documents.

3.2 Segment Length Prior

The $\hat{\mathbf{z}} = \operatorname{argmax}_{\mathbf{z}} p(\mathbf{X}|\mathbf{z})p(\mathbf{z})$ expression we want to maximize to obtain the most likely segmentation puts a prior, $p(\mathbf{z})$, on the segment length of documents. Given the approach of searching the segmentation space only during inference, we do not require the mathematical conveniences of conjugacy for the segment length prior. In this perspective, we can plug in different distributions to see how they behave during the segmentation task. One of such distribution is the Beta-Bernoulli, which has been used before in a probabilistic approach to segmentation (Purver et al., 2006):

$$p(\mathbf{z}) = \left(\frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2}\right)^D \prod_{d=1}^{D} \frac{\Gamma(n_1^d + \gamma)\Gamma(n_0^d + \gamma)}{\Gamma(U_d + 2\gamma)}, \quad (3)$$

where D is the number of document in the dataset, U_d is the total number of utterances in document d, n_1^d is the number of segments in d, n_0^d the number of non-segment boundary utterances in d, and γ the hyperparameter of the Beta distribution.

We also propose a Gamma-Poisson distributed segment length prior. In this setup, we assume that the document topic shift probabilities are drawn from a Gamma prior parameterized by α and β:

$$p(\mathbf{z}) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^D \prod_{d=1}^{D} \frac{\Gamma(n_1^d + \alpha)}{(U_d + \beta)^{n_1^d + \alpha}} \quad (4)$$

Applying priors based on document modality can be done by assuming they are known *a priori*, which is the approach we take. It is only necessary to have dedicated hyperparameters for each modality and apply them accordingly when computing segmentation likelihood. This means we are encoding in the model our prior beliefs about the segment length of each modality. Nonetheless, if the lexical cohesion in a hypothesized segment is strong enough, the model will identify it even if the length is not inline with the prior.

3.3 Dynamic Language Model Prior

The previous priors assume that language model's draws are independent of each other, and, thus cannot encode relationships between them. This is not a reasonable assumption in datasets with documents following an overarching subject. We hypothesize that in these cases, language models change smoothly across topics by establishing a dynamic between the previous and the current prior parameters. This time series modeling of topics can be found in other works (Blei and Lafferty, 2006b,a; Du et al., 2013; Jahnichen et al., 2018). In BeamSeg, we adopt a similar perspective to topic tracking (Watanabe et al., 2011) for modeling such interactions. We factor the β in $\alpha_k \hat{\phi}_{k'}$, a precision and mean language model word probabilities parameters. Assuming some ordering between the topics, k indexes the topic parameters, and k' the parameters of the previous topic. The α_k precision represents the persistence of word usage throughout topics; $\hat{\phi}_k$ models the language model dynamics by assuming that the mean word probabilities at k are the same as those at k'. This entails that there is a single chain of language models, which contrasts with the multiple chains in the original topic tracking model.

To compute the likelihood of the joint under this prior, it is necessary to determine the parameters for all $k \in K$. This is a two-fold process, where we first update the α_k precision parameter using the expression derived from Minka (2000):

$$\alpha_k = \alpha_k \frac{\sum_{w}^{W} \hat{\phi}_{k'w}(\Psi(n_{kw} + \alpha_k \hat{\phi}_{k'w}) - \Psi(\alpha_k \hat{\phi}_{k'w}))}{\Psi(n_k + \alpha_k) - \Psi(\alpha_k)}, \tag{5}$$

where n_{kw} is the number of times word w appear in k; n_k is the total number of words in k; and Ψ is the digamma function. Then, we update the mean word probability parameters:

$$\hat{\phi}_{kw} = \frac{n_{kw} + \alpha_k \hat{\phi}_{k'w}}{n_k + \alpha_k} \tag{6}$$

The update equations are sequentially applied according to a fixed topic order. By following this process, we model long-range dependencies by taking into account the data contribution at each k. Finally, we plug-in the obtained prior parameters in the join likelihood formula in Equation 2.

3.4 Beam Search for Inference

Following Bayesseg (Eisenstein and Barzilay, 2008), inference is viewed as an optimization problem, where the target segmentation maximizes the objective function defined by the joint likelihood. Contrary to Bayesseg, we assume that language models aggregate segments from different documents, making an exhaustive exploration of the segmentation space intractable. To address this problem we combine beam search and a greedy segmentation procedure. Other considered inference alternatives include Gibbs sampling (Bishop, 2006) and Variational Inference (Ghahramani et al., 2008). The difficulty in applying Gibbs sampling is its slow convergence to the stationary distribution, due to the tight coupling of the variables induced by the linear segmentation constraint. A similar problem occurs in the variational inference procedure from Eisenstein (2009), where variational parameters and segmentation are iteratively estimated.

We define $\mathbf{z}_j^*$ as the segmentation that maximizes the objective function up to utterance j. Considering the topic assignment $z_j = k$ and the previous segmentation $\mathbf{z}_{j-1}$, the value for the objective function is written,

$$s(k, j, \mathbf{z}_{j-1}) = p(\{\mathbf{x}_0 ... \mathbf{x}_j\} | \mathbf{z}_{j-1}, z_j = k) \tag{7}$$

Using a recursive definition, we obtain the optimal segmentation using:

$$\mathbf{z}_j^* = \underset{k \in K}{\operatorname{argmax}}\, s(k, j, \mathbf{z}_{j-1}^*) \tag{8}$$

This is a greedy approach since it makes incremental decisions to find the highest likelihood segmentation. This is an error-prone procedure since we should take into account subsequent utterances to discover higher likelihood segmentations. Moreover, once a mistake is made, we cannot recover from it. To address this problem, we add a beam search feature to the algorithm. This is achieved by keeping track of all topic assignments, instead of just the highest likelihood one. At the end of each recursive step, we prune the top-n segmentations.

4 Experiments

We now describe the experimental setup and report the results for the target tasks.

4.1 Datasets

Currently, there are two multi-document segmentation datasets with different document modalities. One of the datasets is comprised of learning materials describing the subject of Adelson-Velsky and Landis' (AVL) trees (Mota et al., 2016). The available modalities are video transcripts, PPT, and HTML. In total, the dataset contains 10 documents, 85 segments, and 17 topics. The other dataset also contains learning materials but from the Physics domain (Mota et al., 2018). In addition, this dataset also has PDF modality. The dataset has 141 documents, 739 segments, and 135 topics from 7 different Physics subjects. This dataset does not provide topic identification labels for the segments. Therefore, we manually annotated it with this information. In this context, we made an inter-annotator agreement study for the 'Introduction to Kinematics' subject with two annotators. A 0.69 Fleiss-kappa (Shrout and Fleiss, 1979) agreement value was obtained, showing that annotators had a similar perception of whether segments share the same topic. Most of the disagreement cases are due to considering textual and plot-based explanations as different topics.

In addition to the previous datasets, we also used Biography documents from Jeong and Titov (2010). The dataset contains 116 documents regarding 29 personalities; 4 documents per personality with a total of 240 segments; the number of topics is 405; all documents have the same HTML modality. The Biography domain has different topic development characteristics from the previous domains. The documents have fewer and shorter segments when compared with the AVL and Physics domains, leaving less room for topics to be described. All datasets were preprocessed by stemming and stop words were removed.

4.2 Segmentation Experiments

In the experiments, we benchmark syntax similarity and probabilistic approaches: C99, CVS, Bayesseg, PLDA, Bayesseg-MD, and MultiSeg. The hyperparameter tuning of the models is done on a development set. In the Biography dataset, we use documents from one of the personalities. For MultiSeg we use the configurations provided by the authors. In the Physics domain, we use ten documents from one of the subjects. The obtained tuning is also used for the AVL trees domain since both datasets have pedagogical content. The Gibbs sampling for PLDA run for 20000 iterations with a burn-in period of 1000 and a lag value of 200. In BeamSeg, we investigate the role of two factors in segmentation: using the dynamic *vs.* an independent language model prior, and using a modality-based segment duration prior *vs.* using a single prior variable. The beam size was set to 200.

To measure performance, we use the standard Window Difference (WD) metric (Pevzner and Hearst, 2002). WD slides a window through a document and penalizes segmentations according to the difference between the number of expected segment boundaries and the predicted ones. This gives partial credit to near-miss situations. The metric is calculated as follows:

$$\text{WD} = \frac{1}{N-k} \sum_{i=1}^{N-k} |ref - hyp| \neq 0, \qquad (9)$$

where N is the length of the document and k the window size. WD is a penalty score between 0 (the best value) and 1. For consistency, we take the output segmentations from all systems and evaluate it using the same software (the python module `segeval` (Fournier, 2013)).

The WD average results for the baseline are in Table 1. In the Biography dataset, MultiSeg is the best performing model, improving the WD of Bayesseg-MD by 0.05. In the AVL dataset, the best results are obtained by Bayesseg-MD. The difference to the second best result, Bayesseg, is 0.02. For the Physics dataset, the single-document model Bayesseg achieves the best results with a WD difference of 0.01. These results show that the performance of the algorithms varies across the different datasets. This suggests that the different modeling approaches do not generalize well to the different characteristics of the datasets. The Biography dataset is characterized by short segments, which does not leave much room for lexical cohesion to be observed. This contrasts with the AVL and Physics datasets where the segments are longer and describe an overarching topic.

Table 1: Segmentation baseline average WD results.

	Bio	**AVL**	**Physics**
C99	0.61	0.59	0.54
PLDA	0.58	0.55	0.49
CVS	0.54	0.45	0.43
Bayesseg	0.53	0.39	**0.42**
Bayesseg-MD	0.42	**0.37**	0.43
MultiSeg	**0.37**	0.41	0.44

The results using different prior configurations are in Table 2. In the table, the LMP and SLP columns correspond to the language model and segment length priors. In the Biography dataset, we can see that using the dynamic LMP instead of the independent improves the the Beta-Bernoulli and Gamma-Poisson results by 0.01 and 0.09, respectively. In the AVL dataset, the dynamic LMP improves the best WD results of the independent LMP by 0.02. When comparing the scope results of the dynamic LMP in the AVL dataset, we observe further improvements when

Table 2: BeamSeg average WD results. The SLP column depicts the Beta-Bernoulli (BB), and Gamma-Poisson (GP) distributions. The scope indicates if the SLP is modality-based (M) or if there is one variable for the whole dataset (D). The Biography dataset has one modality, and, thus, only the D scope exists.

LMP	SLP	Scope	Bio	AVL	Physics
Ind	BB	D	0.54	0.39	0.45
		M	–	0.40	0.42
	GP	D	0.58	0.40	**0.40**
		M	–	0.43	0.42
Dyn	BB	D	0.53	0.44	0.54
		M	–	0.38	0.42
	GP	D	0.49	0.38	0.47
		M	–	**0.37**	**0.40**

Table 3: Number of exact segment boundary matches between hypothesis and reference segmentations.

LMP	SLP	Scope	Bio	AVL	Physics
Ind	BB	D	88	1	16
		M	–	1	8
	GP	D	15	1	5
		M	–	1	20
Dyn	BB	D	147	3	34
		M	–	4	39
	GP	D	244	2	19
		M	–	5	46

using the modality-based SLP; the results differences are 0.06 and 0.01, respectively. In the Physics dataset, a dynamic LMP combined with the modality-based Gamma-Poisson SLP obtains the best results tied with the independent LMP and dataset-based Gamma-Poisson SLP. It should be noted that the former configuration better generalizes across the different datasets since it obtains better results in the Biography and AVL datasets; the WD differences are 0.09 and 0.03, respectively. Looking at the scope results of the dynamic LMP, we observe that the Beta-Bernoulli and the Gamma-Poisson perform better when using the modality prior (0.12 and 0.07 improvements).

WD is a metric that assesses the overall quality of a segmentation, accounting for different types of errors. This can make the WD scores of two very different segmentations to be close, which is the case of the previous results. For example, a segmentation that has no segments and another that only has misplaced segments will have similar WD scores despite being different. To show that the different prior configurations output significantly different segmentations, we provide the counts of the exact segment boundary matches in Table 3. From these results, we can observe that using a dynamic LMP can increase the number of boundary up to 229. A similar observation can be made when comparing the dataset and modality scopes, where the increases are up 27 segments. These increases in exact boundary matches show that despite the small differences in WD the impact on how the segmentation looks like is significant. Therefore, we conclude that using a dynamic LMP with a modality Gamma-Poisson SLP is necessary to achieve the best results.

Comparing BeamSeg's results to the baseline, we see that in the Biography dataset MultiSeg performs better by a 0.12 margin. The main difference between the segmentation of the two approaches is that BeamSeg outputs fewer segments, which is a disadvantage since this dataset has a high number of short segments. In the AVL dataset, the performance is similar to Bayesseg-MD. Looking at the individual documents shows that BeamSeg has better results in five out of ten documents, one tie, and two documents where the WD difference is 0.01. This leaves Bayesseg-MD to perform significantly better only in two documents. Therefore, BeamSeg is more consistent in this dataset. In the Physics dataset, BeamSeg improves the Bayesseg baseline by 4.8%. Taking into account the result analysis, we conclude that BeamSeg's performance depends on the characteristics of the datasets. In datasets where topic development is prominent across the segments (AVL and Physics), BeamSeg is the model with the most consistent results. This is only possible when using a dynamic LMP and a modality Gamma-Poisson SLP, showing that both modeling aspects are relevant to obtain the best segmentation.

To understand the behavior of the priors we provide a segmentation example in Figure 2. From the example, we see that the main difference between the independent and dynamic LMPs is the number of segments. In the independent LMP, the number of segments is low, especially when using the dataset SLP. For the modality SLP, the number of segments is higher but they tend to be misplaced. When using dynamic LMP, the behavior changes at the SLP level. The dataset SLP outputs more segments than the modality version. However, most segments do not match the reference. The modality SLP finds fewer segments, but they tend to be more accurate. This makes sense since the over-segmentation of the dataset SLP might be related to the bias towards documents with short segments, and the modality prior is able to adjust to a wider variety of documents.

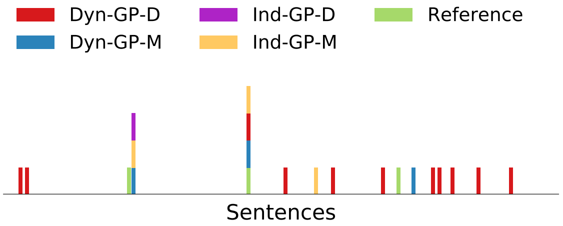


Figure 2: Physics document segmentation using different prior configurations in BeamSeg. Bars with the same color represent segments of the same prior configuration. The names of the configurations start with the LMP type, followed by the SLP, and its scope.

4.3 Topic Identification Experiments

We use the previous datasets to evaluate topic identification and compare multi-document joint models to a pipeline approach. In the pipeline approach, we evaluate clustering and graph-community detection algorithms. The clustering algorithms take the golden standard segments and identify segments sharing the same topic if they are assigned the same cluster. Several clustering algorithms are surveyed (Aggarwal and Reddy, 2014): DBSCAN, Mean Shift, and NMF. For the graph-community detection approach, word communities are obtained from the segments. Then, based on lexical similarity, segments are assigned to one of the communities (Mota et al., 2018). If two segments are assigned to the same community, they share the same topic. Several graph-community detection algorithms are surveyed (Fortunato, 2010): Bigclam, Label Propagation, CNM, Walktraps, and Leading Eigenvector. For conciseness, we only report the results of the best algorithms.

To measure the performance, we use the standard B^3 clustering metric (Amigó et al., 2009). B^3 decomposes uses item-wise precision and recall. Precision represents how many items in the same cluster belong to its class. Recall represents how many items from a class appear in the cluster. The final B^3 value combines precision and recall:

$$B^3 = \frac{1}{0.5(\frac{1}{Pre}) + 0.5(\frac{1}{Rec})} \qquad (10)$$

The baseline results are depicted in Table 4. In this benchmark, the pipeline approach performs better than the joint model in all datasets. The differences range between 0.04 and 0.14 in B^3 score. The DBSCAN clustering approach obtains the best performance in the Biography dataset by a 0.09 margin.The Louvain graph-community detection approach obtains the best results in the AVL and Physics datasets with result differences to DBSCAN of 0.04 in both cases.

Table 4: Topic identification baseline results.

	Bio	AVL	Physics
DBSCAN	**0.66**	0.33	0.34
Louvain	0.57	**0.37**	**0.38**
MultiSeg	0.52	0.29	0.30

Table 5 shows the results for different prior configurations. In the Biography domain, we observe that the dynamic LMP improves the results of both SLPs; 0.03 and 0.16, for the Beta-Bernoulli and Gamma-Poisson, respectively. In the AVL datasets, three different configurations obtain the best performance. In the Physics dataset, the dynamic LMP modality Gamma-Poisson SLP performs better. In this case, using a modality SLP instead of the dataset affords a 0.11 improvement. Comparing the independent and dynamic LMPs, we see that the former improves the results by 0.05. This shows that both modeling aspects are contributing for the best results.

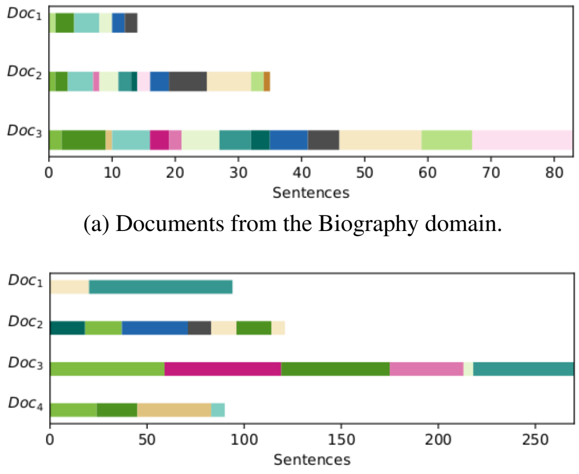
Table 5: BeamSeg topic identification results.

LMP	SLP	Scope	Bio	AVL	Physics
Ind	BB	D	0.51	0.35	0.36
		M	–	**0.39**	0.38
	GP	D	0.37	0.38	0.35
		M	–	0.36	0.37
Dyn	BB	D	0.54	**0.39**	0.30
		M	–	0.32	0.34
	GP	D	0.53	0.38	0.31
		M	–	**0.39**	**0.41**

Comparing BeamSeg's best results to the baseline, we observe that it is only outperformed by DBSCAN in the Biography dataset (a 19.7% difference). DBSCAN obtains better results by putting segments it cannot group in individual clusters, which keep the larger clusters clean. In BeamSeg, the number of identified topics (clusters) is lower, a 385 difference to DBSCAN, which ends up forcing wrong topic segment assignments. In the AVL dataset, BeamSeg improves the Louvain baseline by 5.1%. The topic identification behavior of both approaches is different from the Biography dataset. Louvain only outputs 7 clusters whereas the reference has 17 topics. This is related to the topic development aspect across segments, which makes them hard to distinguish. BeamSeg obtains a higher B^3 score because it is able to identify 15 topics, a number closer to the reference, and, consequently, assign topics more appropriately. In the Physics dataset, BeamSeg improves the baseline by 7.3%. The topic identification patterns are similar to the ones observed in the AVL dataset with BeamSeg outputting more topics than Louvain, 70 and 48 topics, respectively. Another observation is that the performance differences between the Biography and the other datasets are related to the topic structure complexity. In the Biography dataset, there is a tendency for the topic order to persist across documents, whereas in the other datasets the interweaving of the topics is not as regular. This is depicted in Figure 3, where color changes represent a topic changes and similar topics have the same color. In Figure 3a (Biography domain) we can see that the colors sequences in different documents are similar whereas in Figure 3 (Physics domain) the sequence is not constant. Connecting the topic structure differences with the topic order assumptions in BeamSeg explains the performance differences.

5 Conclusions and Future Work

In this work, we propose BeamSeg, an unsupervised Bayesian algorithm that jointly segments documents and identifies topical relationships using a beam search procedure to find high likelihood segmentations during inference. Relationships between topics are modeled using a dynamic prior encoding that word distributions change smoothly in documents with an overarching subject. BeamSeg also models segment length properties based on document modality. To evaluate segmentation, single and multi-document algorithms were used as a baseline. For topic identification, we compared BeamSeg to MultiSeg, another joint model, as well as a pipeline approach. In both tasks, BeamSeg obtains the best results in two of the datasets used for evaluation. The conclusion from the evaluation is that BeamSeg is effective in datasets with prevalent topic development throughout document segments. To achieve the best performance, it is necessary to use a combination of a dynamic LMP with a modality Gamma-Poisson SLP. Therefore, the proposed modeling assumptions fit the data well. This supports the hypothesis that lexical cohesion is a cross-document phenomenon and can be used to leverage multi-document segmentation and topic identification.



(a) Documents from the Biography domain.

(b) Documents from the Physics domain.

Figure 3: Topic identification examples.

Regarding future work, one of the concerns is that the proposed inference procedure is a maximum likelihood estimation approach. Ideally, we want to access the full posterior distribution since it finds more accurate parameters. Another concern is the raw assumption that there is a shared topic ordering among all documents. We believe that addressing these issues will allow BeamSeg to improve its results and to consistently perform in datasets with different characteristics.

Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019, and through the Carnegie Mellon Portugal Program under Grant SFRH/BD/51917/2012.

References

Charu C. Aggarwal and Chandan K. Reddy, editors. 2014. *Data Clustering: Algorithms and Applications*. CRC Press.

Alexander A Alemi and Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings. *arXiv e-prints*, page arXiv:1503.05543.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Hesam Amoualian, Wei Lu, Éric Gaussier, Georgios Balikas, Massih-Reza Amini, and Marianne Clausel. 2017. Topical coherence in lda-based models through induced segmentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1799–1809.

Sebastian Arnold, Rudolf Schneider, Philippe Cudr-Mauroux, Felix A. Gers, and Alexander Lser. 2019. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Arun Balagopalan, Lalitha L. Balasubramanian, Vidhya Balasubramanian, Nithin Chandrasekharan, and Aswin Damodar. 2012. Automatic keyphrase extraction and segmentation of video lectures. In *Proceedings of the International Conference on Transformations in Engineering Education*, pages 152–162.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York.

D. Blei and J. Lafferty. 2006a. Correlated Topic Models. In *Advances in Neural Information Processing Systems*, volume 18, pages 147–154. MIT Press.

David M. Blei and John D. Lafferty. 2006b. Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference*, pages 26–33.

Lan Du, Wray L. Buntine, and Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200.

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, pages 353–361.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343.

Santo Fortunato. 2010. *Physics Reports*, 486(3-5):75 – 174.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1702–1712.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 562–569.

Z. Ghahramani, J. Sung, and S. Bang. 2008. Latent-space variational bayes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 30:2236–2242.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Patrick Jahnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. 2018. Scalable generalized dynamic topic models. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 1427–1435.

Shoaib Jameel and Wai Lam. 2013. An unsupervised topic segmentation model incorporating word order. In *Proceedings of the International Conference on Research and Development in Information Retrieval*, pages 203–212.

Minwoo Jeong and Ivan Titov. 2010. Multi-document topic segmentation. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 1119–1128.

Anna Kazantseva and Stan Szpakowicz. 2011. Linear text segmentation using affinity propagation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 284–293.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 469–473.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Shervin Malmasi, Mark Dras, Mark Johnson, Lan Du, and Magdalena Wolska. 2017. Unsupervised text segmentation based on native language characteristics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1457–1469. Association for Computational Linguistics.

Thomas P. Minka. 2000. Estimating a dirichlet distribution. Technical report.

Pedro Mota, Luísa Coheur, and Maxine Eskénazi. 2018. Efficient navigation in learning materials: An empirical study on the linking process. In *Artificial Intelligence in Education*, pages 230–235.

Pedro Mota, Maxine Eskenazi, and Luísa Coheur. 2016. Multi-document topic segmentation using bayesian estimation. In *Proceedings of the International Workshop on Semantic Multimedia*, pages 443–447.

Koji Murakami, Shouko Masuda, Suguru Matsuyoshi, Eric Nichols, Kentaro Inui, and Yuji Matsumoto. 2009. Annotating semantic relations combining facts and opinions. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 150–153. Association for Computational Linguistics.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Matthew Purver. 2011. Topic segmentation. In *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317.

Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*, pages 17–24.

Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938.

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of Association for Computational Linguistics Student Research Workshop*, pages 37–42.

Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012. Trains of thought: Generating information maps. In *Proceedings of the International Conference on World Wide Web*.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*.

Liang Wang, Sujian Li, Yajuan Lv, and Houfeng WANG. 2017. Learning to rank semantic coherence for topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1340–1344.

Shinji Watanabe, Tomoharu Iwata, Takaaki Hori, Atsushi Sako, and Yasuo Ariki. 2011. Topic tracking language model for speech recognition. *Computer Speech Language*, 25(2):440–461.