Unsupervised Discourse Segmentation of Documents with Inherently Parallel Structure

Minwoo Jeong and **Ivan Titov**
Saarland University
Saarbrücken, Germany
{m.jeong|titov}@mmci.uni-saarland.de

Abstract

Documents often have inherently parallel structure: they may consist of a text and commentaries, or an abstract and a body, or parts presenting alternative views on the same problem. Revealing relations between the parts by jointly segmenting and predicting links between the segments, would help to visualize such documents and construct friendlier user interfaces. To address this problem, we propose an unsupervised Bayesian model for joint discourse segmentation and alignment. We apply our method to the "English as a second language" podcast dataset where each episode is composed of two parallel parts: a story and an explanatory lecture. The predicted topical links uncover hidden relations between the stories and the lectures. In this domain, our method achieves competitive results, rivaling those of a previously proposed supervised technique.

1 Introduction

Many documents consist of parts exhibiting a high degree of parallelism: e.g., abstract and body of academic publications, summaries and detailed news stories, etc. This is especially common with the emergence of the Web 2.0 technologies: many texts on the web are now accompanied with comments and discussions. Segmentation of these parallel parts into coherent fragments and discovery of hidden relations between them would facilitate the development of better user interfaces and improve the performance of summarization and information retrieval systems.

Discourse segmentation of the documents composed of parallel parts is a novel and challenging problem, as previous research has mostly focused on the linear segmentation of isolated texts (e.g., (Hearst, 1994)). The most straightforward approach would be to use a pipeline strategy, where an existing segmentation algorithm finds discourse boundaries of each part independently, and then the segments are aligned. Or, conversely, a sentence-alignment stage can be followed by a segmentation stage. However, as we will see in our experiments, these strategies may result in poor segmentation and alignment quality.

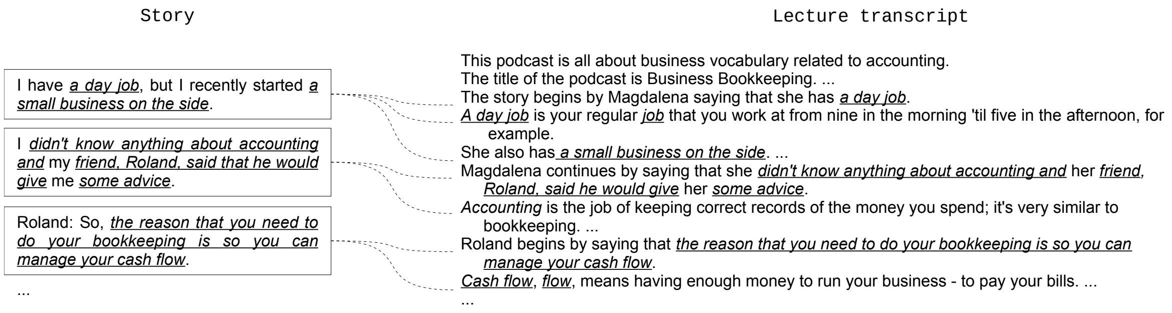
To address this problem, we construct a non-parametric Bayesian model for joint segmentation and alignment of parallel parts. In comparison with the discussed pipeline approaches, our method has two important advantages: (1) it leverages the lexical cohesion phenomenon (Halliday and Hasan, 1976) in modeling the parallel parts of documents, and (2) ensures that the effective number of segments can grow adaptively. Lexical cohesion is an idea that topically-coherent segments display compact lexical distributions (Hearst, 1994; Utiyama and Isahara, 2001; Eisenstein and Barzilay, 2008). We hypothesize that not only isolated fragments but also each group of linked fragments displays a compact and consistent lexical distribution, and our generative model leverages this inter-part cohesion assumption.

In this paper, we consider the dataset of "English as a second language" (ESL) podcast[1], where each episode consists of two parallel parts: a *story* (an example monologue or dialogue) and an explanatory *lecture* discussing the meaning and usage of English expressions appearing in the story. Fig. 1 presents an example episode, consisting of two parallel parts, and their hidden topical relations.[2] From the figure we may conclude that there is a tendency of word repetition between each pair of aligned segments, illustrating our hypothesis of compactness of their joint distribution. Our goal is

[1] http://www.eslpod.com/

[2] Episode no. 232 post on Jan. 08, 2007.

Proceedings of the ACL 2010 Conference Short Papers, pages 151–155,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics



I have *a day job*, but I recently started *a small business on the side*.

I *didn't know anything about accounting and* my *friend, Roland, said that he would give* me *some advice*.

Roland: So, *the reason that you need to do your bookkeeping is so you can manage your cash flow*.

...

This podcast is all about business vocabulary related to accounting.
The title of the podcast is Business Bookkeeping. ...
The story begins by Magdalena saying that she has *a day job*.
A day job is your regular *job* that you work at from nine in the morning 'til five in the afternoon, for example.
She also has *a small business on the side*. ...
Magdalena continues by saying that she *didn't know anything about accounting and* her *friend, Roland, said he would give* her *some advice*.
Accounting is the job of keeping correct records of the money you spend; it's very similar to bookkeeping. ...
Roland begins by saying that *the reason that you need to do your bookkeeping is so you can manage your cash flow*.
Cash flow, *flow*, means having enough money to run your business - to pay your bills. ...
...

Figure 1: An example episode of ESL podcast. Co-occurred words are represented in *italic* and underline.

to divide the lecture transcript into discourse units and to align each unit to the related segment of the story. Predicting these structures for the ESL podcast could be the first step in development of an e-learning system and a podcast search engine for ESL learners.

2 Related Work

Discourse segmentation has been an active area of research (Hearst, 1994; Utiyama and Isahara, 2001; Galley et al., 2003; Malioutov and Barzilay, 2006). Our work extends the Bayesian segmentation model (Eisenstein and Barzilay, 2008) for isolated texts, to the problem of segmenting parallel parts of documents.

The task of aligning each sentence of an abstract to one or more sentences of the body has been studied in the context of summarization (Marcu, 1999; Jing, 2002; Daumé and Marcu, 2004). Our work is different in that we do not try to extract the most relevant sentence but rather aim to find coherent fragments with maximally overlapping lexical distributions. Similarly, the query-focused summarization (e.g., (Daumé and Marcu, 2006)) is also related but it focuses on sentence extraction rather than on joint segmentation.

We are aware of only one previous work on joint segmentation and alignment of multiple texts (Sun et al., 2007) but their approach is based on similarity functions rather than on modeling lexical cohesion in the generative framework. Our application, the analysis of the ESL podcast, was previously studied in (Noh et al., 2010). They proposed a supervised method which is driven by pairwise classification decisions. The main drawback of their approach is that it neglects the discourse structure and the lexical cohesion phenomenon.

3 Model

In this section we describe our model for discourse segmentation of documents with inherently parallel structure. We start by clarifying our assumptions about their structure.

We assume that a document $\boldsymbol{x}$ consists of K parallel parts, that is, $\boldsymbol{x} = \{\boldsymbol{x}^{(k)}\}_{k=1:K}$, and each part of the document consists of segments, $\boldsymbol{x}^{(k)} = \{\boldsymbol{s}_i^{(k)}\}_{i=1:I}$. Note that the effective number of fragments I is unknown. Each segment can either be specific to this part (drawn from a *part-specific* language model $\phi_i^{(k)}$) or correspond to the entire document (drawn from a *document-level* language model $\phi_i^{(doc)}$). For example, the first and the second sentences of the lecture transcript in Fig. 1 are part-specific, whereas other linked sentences belong to the document-level segments. The document-level language models define topical links between segments in different parts of the document, whereas the part-specific language models define the linear segmentation of the remaining unaligned text.

Each document-level language model corresponds to the set of aligned segments, at most one segment per part. Similarly, each part-specific language model corresponds to a single segment of the single corresponding part. Note that all the documents are modeled independently, as we aim not to discover collection-level topics (as e.g. in (Blei et al., 2003)), but to perform joint discourse segmentation and alignment.

Unlike (Eisenstein and Barzilay, 2008), we cannot make an assumption that the number of segments is known a-priori, as the effective number of part-specific segments can vary significantly from document to document, depending on their size and structure. To tackle this problem, we use Dirichlet processes (DP) (Ferguson, 1973) to define priors on the number of segments. We incorporate them in our model in a similar way as it is done for the Latent Dirichlet Allocation (LDA) by Yu et al. (2005). Unlike the standard LDA, the topic proportions are chosen not from a Dirichlet prior but from the marginal distribution $GEM(\alpha)$ defined by the *stick breaking* construction (Sethuraman, 1994), where α is the concentration parameter of the underlying DP distribution. $GEM(\alpha)$ defines a distribution of partitions of the unit interval into a countable number of parts.

The formal definition of our model is as follows:

- Draw the document-level topic proportions $\boldsymbol{\beta}^{(doc)} \sim GEM(\alpha^{(doc)})$.
- Choose the document-level language model $\phi_i^{(doc)} \sim Dir(\gamma^{(doc)})$ for $i \in \{1, 2, \ldots\}$.
- Draw the part-specific topic proportions $\boldsymbol{\beta}^{(k)} \sim GEM(\alpha^{(k)})$ for $k \in \{1, \ldots, K\}$.
- Choose the part-specific language models $\phi_i^{(k)} \sim Dir(\gamma^{(k)})$ for $k \in \{1, \ldots, K\}$ and $i \in \{1, 2, \ldots\}$.
- For each part k and each sentence n:
  - Draw type $t_n^{(k)} \sim Unif(Doc, Part)$.
  - If ($t_n^{(k)} = Doc$); draw topic $z_n^{(k)} \sim \boldsymbol{\beta}^{(doc)}$; generate words $\boldsymbol{x}_n^{(k)} \sim Mult(\phi_{z_n^{(k)}}^{(doc)})$
  - Otherwise; draw topic $z_n^{(k)} \sim \boldsymbol{\beta}^{(k)}$; generate words $\boldsymbol{x}_n^{(k)} \sim Mult(\phi_{z_n^{(k)}}^{(k)})$.

The priors $\gamma^{(doc)}$, $\gamma^{(k)}$, $\alpha^{(doc)}$ and $\alpha^{(k)}$ can be estimated at learning time using non-informative hyperpriors (as we do in our experiments), or set manually to indicate preferences of segmentation granularity.

At inference time, we enforce each latent topic $z_n^{(k)}$ to be assigned to a contiguous span of text, assuming that coherent topics are not recurring across the document (Halliday and Hasan, 1976). It also reduces the search space and, consequently, speeds up our sampling-based inference by reducing the time needed for Monte Carlo chains to mix. In fact, this constraint can be integrated in the model definition but it would significantly complicate the model description.

4 Inference

As exact inference is intractable, we follow Eisenstein and Barzilay (2008) and instead use a Metropolis-Hastings (MH) algorithm. At each iteration of the MH algorithm, a new potential alignment-segmentation pair $(\boldsymbol{z}', \boldsymbol{t}')$ is drawn from a proposal distribution $Q(\boldsymbol{z}', \boldsymbol{t}'|\boldsymbol{z}, \boldsymbol{t})$, where $(\boldsymbol{z}, \boldsymbol{t})$ is the current segmentation and its type. The new pair $(\boldsymbol{z}', \boldsymbol{t}')$ is accepted with the probability

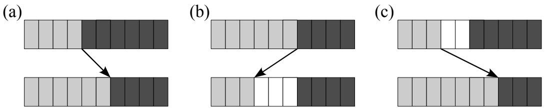


Figure 2: Three types of moves: (a) shift, (b) split and (c) merge.

$$\min\left(1, \frac{P(\boldsymbol{z}', \boldsymbol{t}', \boldsymbol{x})Q(\boldsymbol{z}', \boldsymbol{t}'|\boldsymbol{z}, \boldsymbol{t})}{P(\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{x})Q(\boldsymbol{z}, \boldsymbol{t}|\boldsymbol{z}', \boldsymbol{t}')}\right).$$

In order to implement the MH algorithm for our model, we need to define the set of potential *moves* (i.e. admissible changes from $(\boldsymbol{z}, \boldsymbol{t})$ to $(\boldsymbol{z}', \boldsymbol{t}')$), and the proposal distribution Q over these moves. If the actual number of segments is known and only a linear discourse structure is acceptable, then a single move, *shift* of the segment border (Fig. 2(a)), is sufficient (Eisenstein and Barzilay, 2008). In our case, however, a more complex set of moves is required.

We make two assumptions which are motivated by the problem considered in Section 5: we assume that (1) we are given the number of document-level segments and also that (2) the aligned segments appear in the same order in each part of the document. With these assumptions in mind, we introduce two additional moves (Fig. 2(b) and (c)):

- *Split move*: select a segment, and split it at one of the spanned sentences; if the segment was a document-level segment then one of the fragments becomes the same document-level segment.
- *Merge move*: select a pair of adjacent segments where at least one of the segments is part-specific, and merge them; if one of them was a document-level segment then the new segment has the same document-level topic.

All the moves are selected with the uniform probability, and the distance c for the shift move is drawn from the proposal distribution proportional to $c^{-1/c_{max}}$. The moves are selected independently for each part.

Although the above two assumptions are not crucial as a simple modification to the set of moves would support both introduction and deletion of document-level fragments, this modification was not necessary for our experiments.

5 Experiment

5.1 Dataset and setup

Dataset We apply our model to the ESL podcast dataset (Noh et al., 2010) of 200 episodes, with an average of 17 sentences per story and 80 sentences per lecture transcript. The gold standard alignments assign each fragment of the story to a segment of the lecture transcript. We can induce segmentations at different levels of granularity on both the story and the lecture side. However, given that the segmentation of the story was obtained by an automatic sentence splitter, there is no reason to attempt to reproduce this segmentation. Therefore, for quantitative evaluation purposes we follow Noh et al. (2010) and restrict our model to alignment structures which agree with the given segmentation of the story. For all evaluations, we apply standard stemming algorithm and remove common stop words.

Evaluation metrics To measure the quality of segmentation of the lecture transcript, we use two standard metrics, P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002), but both metrics disregard the alignment links (i.e. the topic labels). Consequently, we also use the macro-averaged F_1 score on pairs of aligned span, which measures both the segmentation and alignment quality.

Baseline Since there has been little previous research on this problem, we compare our results against two straightforward unsupervised baselines. For the first baseline, we consider the pairwise sentence alignment (SentAlign) based on the unigram and bigram overlap. The second baseline is a pipeline approach (Pipeline), where we first segment the lecture transcript with BayesSeg (Eisenstein and Barzilay, 2008) and then use the pairwise alignment to find their best alignment to the segments of the story.

Our model We evaluate our joint model of segmentation and alignment both with and without the split/merge moves. For the model without these moves, we set the desired number of segments in the lecture to be equal to the actual number of segments in the story I. In this setting, the moves can only adjust positions of the segment borders. For the model with the split/merge moves, we start with the same number of segments I but it can be increased or decreased during inference. For evaluation of our model, we run our inference algorithm from five random states, and take the 100,000th iteration of each chain as a sample. Results are the average over these five runs. Also we perform L-BFGS optimization to automatically adjust the non-informative hyperpriors after each 1,000 iterations of sampling.

Method	P_k	WD	$1 - F_1$
Uniform	0.453	0.458	0.682
SentAlign	0.446	0.547	0.313
Pipeline (I)	0.250	0.249	0.443
Pipeline ($2I$+1)	0.268	0.289	0.318
Our model (I)	0.193	0.204	0.254
+split/merge	**0.181**	**0.193**	**0.239**

Table 1: Results on the ESL podcast dataset. For all metrics, lower values are better.

5.2 Result

Table 1 summarizes the obtained results. 'Uniform' denotes the minimal baseline which uniformly draws a random set of I spans for each lecture, and then aligns them to the segments of the story preserving the linear order. Also, we consider two variants of the pipeline approach: segmenting the lecture on I and $2I + 1$ segments, respectively.[3] Our joint model substantially outperforms the baselines. The difference is statistically significant with the level $p < .01$ measured with the paired t-test. The significant improvement over the pipeline results demonstrates benefits of joint modeling for the considered problem. Moreover, additional benefits are obtained by using the DP priors and the split/merge moves (the last line in Table 1). Finally, our model significantly outperforms the previously proposed supervised model (Noh et al., 2010): they report micro-averaged F_1 score 0.698 while our best model achieves 0.778 with the same metric. This observation confirms that lexical cohesion modeling is crucial for successful discourse analysis.

6 Conclusions

We studied the problem of joint discourse segmentation and alignment of documents with inherently parallel structure and achieved favorable results on the ESL podcast dataset outperforming the cascaded baselines. Accurate prediction of these hidden relations would open interesting possibilities

[3] The use of the DP priors and the split/merge moves on the first stage of the pipeline did not result in any improvement in accuracy.

for construction of friendlier user interfaces. One example being an application which, given a user-selected fragment of the abstract, produces a summary from the aligned segment of the document body.

Acknowledgment

The authors acknowledge the support of the Excellence Cluster on Multimodal Computing and Interaction (MMCI), and also thank Mikhail Kozhevnikov and the anonymous reviewers for their valuable comments, and Hyungjong Noh for providing their data.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Computational Linguistics*, 34(1–3):177–210.

David M. Blei, Andrew Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3:993–1022.

Hal Daumé and Daniel Marcu. 2004. A phrase-based hmm approach to document/abstract alignment. In *Proceedings of EMNLP*, pages 137–144.

Hal Daumé and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of ACL*, pages 305–312.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, pages 334–343.

Thomas S. Ferguson. 1973. A Bayesian analysis of some non-parametric problems. *Annals of Statistics*, 1:209–230.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*, pages 562–569.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Marti Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of ACL*, pages 9–16.

Hongyan Jing. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4):527–543.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL*, pages 25–32.

Daniel Marcu. 1999. The automatic construction of large-scale corpora for summarization research. In *Proceedings of ACM SIGIR*, pages 137–144.

Hyungjong Noh, Minwoo Jeong, Sungjin Lee, Jonghoon Lee, and Gary Geunbae Lee. 2010. Script-description pair extraction from text documents of English as second language podcast. In *Proceedings of the 2nd International Conference on Computer Supported Education*.

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

Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Bingjun Sun, Prasenjit Mitra, C. Lee Giles, John Yen, and Hongyuan Zha. 2007. Topic segmentation with shared topic detection and alignment of multiple documents. In *Proceedings of ACM SIGIR*, pages 199–206.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of ACL*, pages 491–498.

Kai Yu, Shipeng Yu, and Vokler Tresp. 2005. Dirichlet enhanced latent semantic analysis. In *Proceedings of AISTATS*.