Hierarchical topic structuring: from dense segmentation to topically focused fragments via burst analysis

Anca Simon, Pascale Sébillot, Guillaume Gravier

Hierarchical topic structuring: from dense segmentation to topically focused fragments via burst analysis

Anca Simon
Université de Rennes 1
IRISA & INRIA Rennes

Pascale Sébillot
INSA de Rennes
IRISA & INRIA Rennes

Guillaume Gravier
CNRS
IRISA & INRIA Rennes

anca-roxana.simon@irisa.fr
pascale.sebillot@irisa.fr
guillaume.gravier@irisa.fr

Abstract

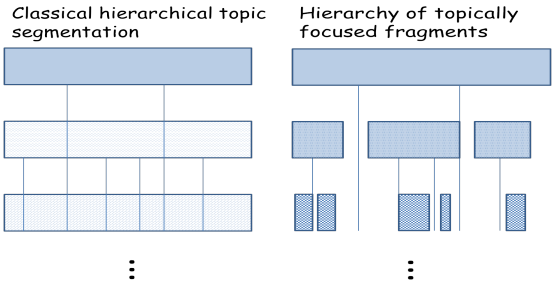
Topic segmentation traditionally relies on lexical cohesion measured through word re-occurrences to output a dense segmentation, either linear or hierarchical. In this paper, a novel organization of the topical structure of textual content is proposed. Rather than searching for topic shifts to yield dense segmentation, we propose an algorithm to extract topically focused fragments organized in a hierarchical manner. This is achieved by leveraging the temporal distribution of word re-occurrences, searching for bursts, to skirt the limits imposed by a global counting of lexical re-occurrences within segments. Comparison to a reference dense segmentation on varied datasets indicates that we can achieve a better topic focus while retrieving all of the important aspects of a text.

1 Introduction

Being aware of the topical structure of texts or automatic transcripts is known to be helpful for multiple natural language processing tasks such as summarization, question answering, etc. Various solutions have emerged to obtain such a structure, the most interesting ones being generic solutions that can be applied on any kind of textual data. These generic solutions are generally based on lexical cohesion, i.e., on identifying segments with a consistent use of vocabulary, in particular measured via word re-occurrences. Their output is a dense segmentation, i.e., contiguous segments, most of the time linear even if the structure of discourse is known to have a hierarchical form (Grosz and Sidner, 1986; Marcu, 2000).

Dense segmentation, linear or hierarchical, is however not necessarily appropriate to reflect the fact that some fragments of the data bear important ideas while others are simple fillers, i.e., they do not bring additional important information. This notion of irrelevant ideas was also mentioned in (Choi, 2000) where the author notes that skipping irrelevant fragments improves navigation. In addition, lexical re-occurrence is not sufficient for this type of segmentation, as we will demonstrate. In particular in the hierarchical case, segments get smaller as we go towards fine grain segmentation: As a consequence, there is a reduced number of words per segment and neighboring segments might refer to the same general topic and hence exhibit high lexical coherence.

To skirt these limits, we investigate a different way of organizing the topical structure of textual content. We rely on the fact that some words appear in bursts, i.e., with a frequency higher than normal at specific locations in the text. The key idea that we leverage is that the presence of lexical bursts usually indicates a strong topical focus, as we will highlight. As an alternative to dense hierarchical topic segmentation, we propose to derive a hierarchy of topically focused fragments as illustrated in Figure 1. A generic representation for classical hierarchical topic segmentation is depicted in Figure 1(a), where the main topics are divided into sub-topics, which in turn can be divided. A dense segmentation is provided at each level and the goal is to identify topic frontiers. Departing from the traditional thinking, the idea in Figure 1(b) is to spot topically focused fragments that are not necessarily contiguous and organize the fragments at various levels in a hierarchical way. Exploiting Kleinberg's algorithm (Kleinberg, 2002) to provide a hierarchy of bursty frag-



(a) (b)

Figure 1: Generic respresentations of (a) classical dense topic segmentation vs (b) topically focused fragments. Vertical lines illustrate topic and subtopic frontiers.

ments for each word, we propose an algorithm to build a topical organization of a document such as the one in Figure 1(b). As a proof of concept, evaluations are performed by qualitative and quantitative comparison to the traditional dense segmentation for which hierarchical reference segmentation exists.

The paper is organized as follows. Section 2 presents a brief overview of existing work on hierarchical topic segmentation. Section 3 shows the limits of current hierarchical segmentation strategies relying on lexical cohesion. Section 4 analyzes the distribution of reiterations via burst analysis. Section 5 describes and evaluates the algorithm to build the hierarchy of topically focused fragments. Section 6 concludes the paper.

2 Related work

Several studies for statistical laws in language have proposed burst detection models that analyze the distributional pattern of words (Sarkar et al., 2005a; Madsen et al., 2005). The quest for these models has been driven by various applications like: keyword extraction (Monachesi et al., 2006), style investigation (Sarkar et al., 2005b), etc. To our knowledge, burst detection hasn't been used before in the context of topic segmentation of textual data, most of the approaches exploiting lexical cohesion through words re-occurrences

In the case of hierarchical topic segmentation, a first approach is to apply a linear topic segmentation algorithm recursively (Carroll, 2010; Guinaudeau, 2011). One of the challenges is to decide when to stop. Additionally, a segmentation error at a higher level in the hierarchy can be propagated towards the lower levels. Hence, a few models have been proposed to explicitly model the hierarchical segment structure. HierBayes (Eisenstein, 2009) is an unsupervised algorithm formalized in a Bayesian probabilistic framework. The underlying principle is that each word in a text is represented by a language model estimated on a portion, more or less important, of the text. The drawback of this approach is that it cannot deal with segments of variable lengths and it needs prior information on the duration of the segments at each level in the hierarchy. In (Kazantseva and Szpakowicz, 2014), the authors propose to use the hierarchical affinity propagation graphical model introduced in (Givoni et al., 2011) to extract the hierarchical topic structure. Similar to Eisenstein, prior information on the granularity of the segmentation is required.

3 The limits of current hierarchical topic segmentation strategies

All of the techniques mentioned in the previous section target dense segmentation. To motivate the use of burst analysis and the introduction of a non-dense topical structure, we first show that lexical re-occurrences fail at explaining the reference hierarchical segmentation in a number of cases. We study here the behavior of two commonly-used measures of lexical cohesion on the hierarchical reference segmentation of a number of datasets.

3.1 Measures of lexical cohesion via word re-occurrences

The first measure considered is the similarity-based approach for which a cosine similarity is computed between vectors representing the content of adjacent segments. Let $\mathcal{V}$ represent the vocabulary containing each word that appears in the text to segment. For each segment S_i, the vector $\mathbf{v}_i$ contains the TF-IDF weight of each term in $\mathcal{V}$ computed over S_i, where the IDF values are computed over the entire collection for each dataset. The cosine similarity is defined as

$$C(S_{i-1}, S_i) = \frac{\sum\limits_{v \in \mathcal{V}} \mathbf{v}_{i-1}(v)\ \mathbf{v}_i(v)}{\sqrt{\sum\limits_{v \in \mathcal{V}} \mathbf{v}^2_{i-1}(v) \sum\limits_{v \in \mathcal{V}} \mathbf{v}^2_i(v)}} \ .$$

The second measure considered is a probabilistic one where lexical cohesion for a segment S_i is computed using a Laplace law as in (Utiyama and Isahara, 2001), i.e.,

$$C(S_i) = log \prod_{j=1}^{n_i} \frac{f_i(w^i_j) + 1}{n_i + k} \ ,$$

where n_i is the number of word occurrences in S_i, $f_i(w_j^i)$ is the number of occurrences of the word w_j^i in segment S_i and k is the number of words in $\mathcal{V}$. The quantity $C(S_i)$ increases when words are repeated and decreases consistently when they are different. This value obtained for a segment S_i can be seen as the capacity of a language model learned on the segment to predict the words of the segment.

Note that the two measures are complementary: One considers adjacent segments to identify topic shifts, while the other intrisically measures the cohesion of a segment. Both are nevertheless independent of the segmentation method used.

3.2 Corpora

Three datasets, previously used in the context of hierarchical segmentation, are considered in this paper: a medical textbook (Eisenstein, 2009); Wikipedia articles (Carroll, 2010); manual and automatic French TV show transcripts (Guinaudeau, 2011). All the datasets are preprocessed in the same way: Words are tagged and lemmatized with TreeTagger[1] and only the nouns, non modal verbs and adjectives are retained.

The Wikipedia corpus contains 66 articles with a hierarchy of up to 4 levels. The reference segmentation is obtained from the structures given by the author of each article. Alike, the reference segmentation considered for the medical dataset is the structure created by the author when writing the book. The book is organized as follows: It has 17 parts; each part is divided into chapters, which are in turn divided into sections. This corpus was first used by (Eisenstein and Barzilay, 2008) for linear topic segmentation and the segmentation was done at the level of sections (227 chapters and 1,136 sections). The French TV show transcripts dataset is more challenging than the two others, particularly with automatic transcripts. The corpus contains seven episodes of a report show *Envoyé Spécial*. Each report has a duration of about 2 hours and was automatically transcribed with a standard ASR system. Manual transcripts for 4 reports are also available. Note that transcripts do not respect the norms of written texts: no paragraphs; structure based on utterances (i.e., sequences of words often separated by breath intakes) rather than sentences; no punctuation signs or capital letters. Additionally, ASR transcripts contain transcription errors (word error rate ca. 30 %) which may imply a lack of word repetitions. The reference segmentation has 3 levels and was obtained through manual annotation (done by an annotator). The first level has 26 frontiers, the second one 246 and the third one 722.

Throughout this paper the highest level in the hierarchy will be denoted level 0 and represents an entire Wikipedia article/part of the medical textbook/transcript of a TV show and the lowest level will correspond to level 4/2/3 respectively.

3.3 Experimental evaluation

For the two measures, Figure 2 reports the evolution of the lexical cohesion over all segments of the second level in the reference topic hierarchy as well as global statistics for $C(S_i)$. Each row corresponds to a different dataset: First, the TV show manual transcripts (Fig. 2(a), 2(b), 2(c)) and second, the medical textbook (Fig. 2(d), 2(e), 2(f). Similar results are obtained also for the Wikipedia articles, but for brevity we do not present them here. The figures on the first column show the cohesion values obtained with the probabilistic measure for each sub-topic in the reference segmentation. The figures on the second column show general statistics (average, min and max values) for the same measure on the entire datasets. And the figures on the last column show the values obtained with the cosine similarity measure between consecutive sub-topics. For the medical textbook corpora the values on the first and last column are reported only on 4 samples for a better visibility. As it can be observed, there is a high variability in the cohesion values across sub-topics segments as well as in the similarity between consecutive segments within a document. Variability is also significantly high across documents (Fig. 2(b),2(e)), thus making it very difficult to define a threshold for segmentation purposes.

These findings point to the fact that the reference segmentation cannot be explained by the lexical cohesion measured via word re-occurrences counted globally on a segment. However, given the advantages of using lexical re-occurrences, we propose to analyze them from a different angle, by looking at the distributions of word repetitions via burst analysis. The words that are important in the process of topic segmentation are those with increased frequency for a particular segment and with insignificant appearances in the rest of the

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

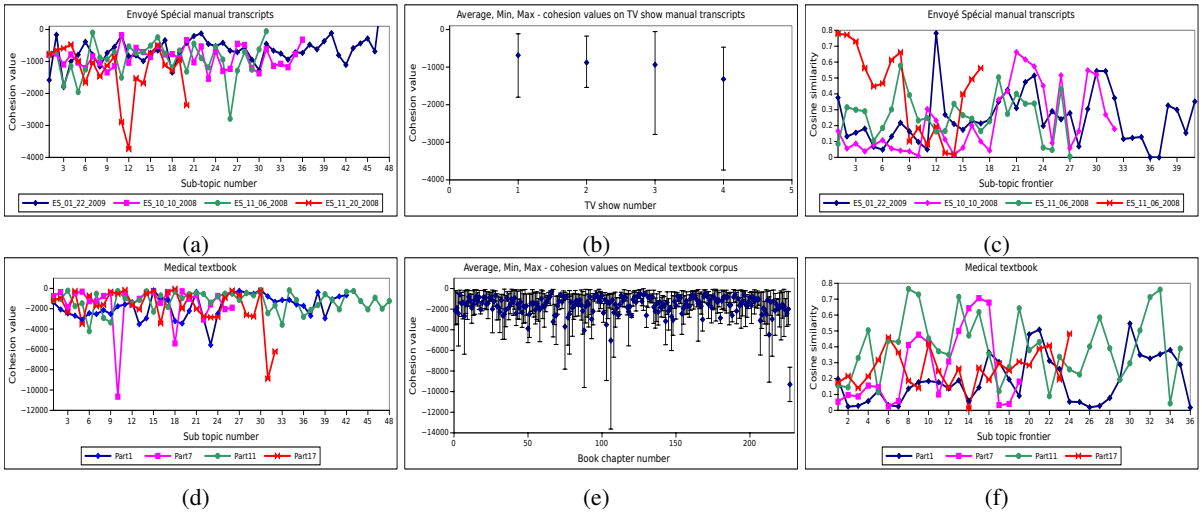


Figure 2: Lexical cohesion measures for each dataset. Each row correspond to a dataset, from top to bottom: TV shows, medical textbook. Columns correspond to, from left to right: $C(S_i)$, distribution of $C(S_i)$ per document, $C(S_{i-1}, S_i)$. Only a fraction of the results are presented for the textbook for legibility reasons.

segments. Note that the second point is usually not taken into account in existing segmentation algorithms. Such words can be captured through burst analysis. In the following sections, we thus analyze the relevance of bursts in the context of hierarchical topic segmentation.

4 Distribution of lexical reiterations through burst analysis

The burst of a given word corresponds to a period where the word occurs with increased frequency with respect to normal behavior. Thus a burst signals both the existence of lexical disruption and of fragments of text that are cohesive: A fragment with one or more words bursts has a more consistent use of vocabulary, with concepts repeated locally in the fragment, apart from the rest of the text; also a fragment with bursty words can be differentiated from other fragments in the text since the burst of a word signals a high frequency of that word in a restricted interval and therefore increases the disruption with adjacent fragments.

4.1 Kleinberg's algorithm

At the core of the analysis of the distribution of word re-occurrences, we rely on Kleinberg's algorithm (Kleinberg, 2002) to identify word bursts, together with the intervals where they occur[2]. The algorithm relies on an infinite-state automaton where the states $i \in \mathbb{N}^+$ correspond to the frequency at which an individual word repeats. Arbitrarily, state 0 accounts for normal behavior while increasing values of i correspond to increasing levels of burstiness. State transitions thus correspond to points in time when there is a important change in the occurrence frequency of a word. The algorithm outputs a hierarchy of burst intervals for each word, taking one word at a time, by searching for the state sequence that minimizes a cost function. For more details, see (Kleinberg, 2002). The interval of a burst at level j in the hierarchy of bursts is the maximal interval during which the optimal state sequence is in state j or higher, i.e., $k > j$, thus forming a hierarchical organization of burst intervals. In other words, a word considered bursty on a time interval $[a, b]$ with a burstiness level of i is simultaneously considered as busty at a level $i-1$ on an interval $[c, d]$, with $[a, b] \subset [c, d]$. This hierarchy is illustrated in Figure 4 for one word: The word occurs with a burstiness level of 1 on the first utterances, with an important amount of occurrences at the very beginning yielding a short interval at level 2 included in the interval at level 1. Long bursts intensifying into briefer ones can be seen as imposing a fine-grain organization within the text according to a natural tree structure.

4.2 A case analysis of bursts

We conducted a case-study to assess if the concept of bursts is relevant or not to produce traditional dense segmentations. For each segment at each

[2]We use Jeff Binder's open-source implementation, available at http://cran.r-project.org/web/packages/bursts

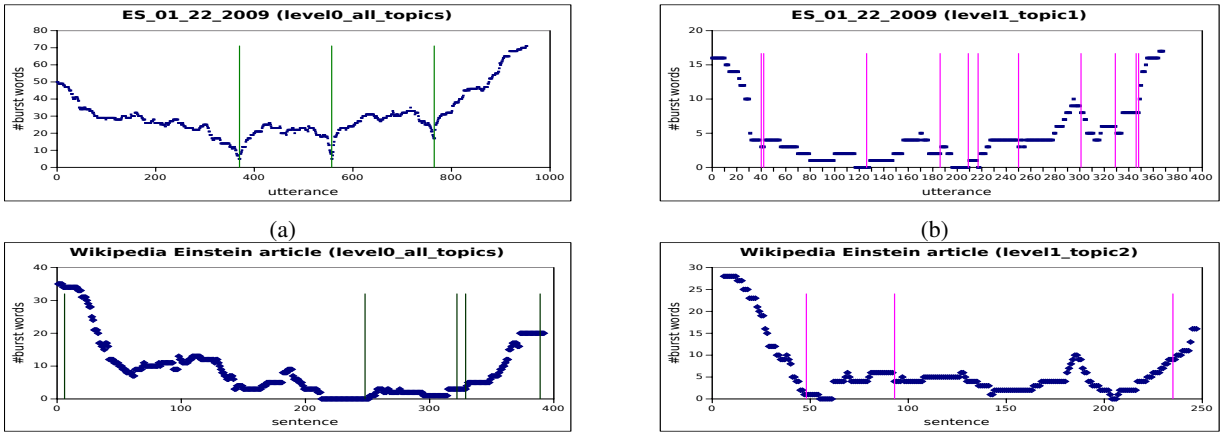


Figure 3: Number of bursty words for each utterance on a TV show (top) and on a Wikpedia article (bottom). Burst intervals are computed either from dense topic segments taken at level 0 (left), or from the level 1 subtopics of the first level-0 topic (right). Vertical lines indicate reference segment boundaries.

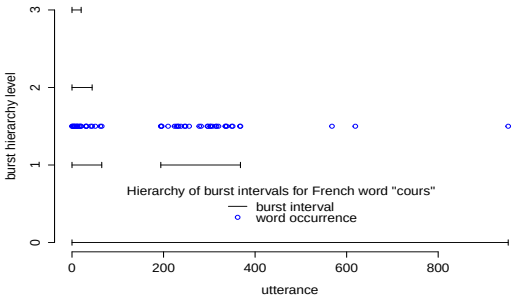


Figure 4: Sample output of Kleinberg's algorithm: The y-axis depicts the burstiness level while utterance number are on the x-axis; Circles indicate occurrences of the word considered. There are two bursts of level 1, the first one coming along with a burst of level 2 for a fraction of its time.

level of the reference dense topic segmentation, a hierarchy of burst intervals as the one illustrated in Figure 4 is computed for each word. Given the set of burst intervals, we count for each utterance the number of words within the utterance which appear as bursty at that position. Figure 3 presents the counts for bursts computed at two levels (level 0 and level 1) in the reference hierarchical topic segmentation for a sample from the TV show transcripts and one from a Wikipedia article. The reference frontiers are marked with vertical lines. For brevity, figures for the medical corpus, which are similar to those of the TV show transcripts, are not presented. We expect that local minima in the plots, i.e., utterances that contain few bursy words, are indicators of topic shifts.

Results in Figures 3(a) and 3(b) are obtained on the reference transcript of one TV show, for level 0 and level 1. Clearly, local minima in the plot for level 0 can be associated with the reference frontiers: The number of bursts shared between the utterances at these points are considerably fewer than at any other point. Thus, at this level, the topical segments can be easily identified relying on bursts information. The same analysis for level 1 shows that local minima are neither easy to identify in this case, nor do they correspond with reference frontiers (see, Figure 3(b)). Results on a Wikipedia article in Figures 3(c) and 3(d) show that in this type of documents the topic shifts are not as obvious to identify as in the case of the TV show at level 0.

By looking specifically at each segment and analyzing the bursts in the segment, two types of bursts can be distinguished: Bursts that are specific to each of the segments' sub-segments and bursts that are shared between the segments' sub-segments. The number of specific bursts for a sub-segment is given by the number of burst intervals contained between the boundaries of that sub-segment, while the number of bursts shared between sub-segments is given by the number of burst intervals crossing over the frontier between the sub-segments. For example, the French TV show has an average number of specific (resp. shared) bursts of 51 (resp. 6.75) at level 0 while the figures decrease to resp. 2.91 and 1.58 at level 1. When going to lower levels, the number of specific bursts decreases and approaches the number of bursts shared. Thus similar observations as the ones drawn from the counts of bursts (Figure 3) can be made.

This case study leads to several important obser-

vations: Frontiers can be identified when there are few bursts across a position and many before/after that position; words that are bursty at one level in the topic hierarchy (i.e., specific at this level) can become general for lower levels in the hierarchy; when going to lower levels in the topic hierarchy, the number of bursts decreases; there are segments with no bursty words. Thus burst analysis is relevant in the context of hierarchical topic segmentation, but an appropriate way to exploit it has to be proposed; we address this open issue in the following section.

5 Hierarchical structure of topically focused fragments

Burst modeling has the effect of exposing salient words (i.e., keywords) with different (burst hierarchy) levels. We propose to take advantage of this fact to spot salient topics and sub-topics. Thus, we do not focus anymore on producing a dense hierarchy of segments but instead we aim to derive a hierarchy of topically focused, i.e., salient, fragments which are not necessarily contiguous.

5.1 The algorithm

We propose a new algorithm that generates a hierarchy of salient topics using an agglomerative clustering of burst intervals. The result is a set of nested topically focused fragments which are hierarchically organized. Note that contrary to the segments obtained with traditional topic segmentation, the fragments resulting from clustering burst intervals are not necessarily contiguous, and they have a stronger focus. Obtaining this structuring of the data brings several advantages: It is a representation of the entire document; it is highly informative since the words included are assumed to be the most informative ones in the document; the bursty words present in the resulting fragments offer an approximation of what the document is about and facilitate its understanding; relevant information is given at various levels of detail.

The clustering algorithm exploits the output of Kleinberg's burst detection algorithm which provides for each word a hierarchy of burst intervals. The key idea is to iteratively group together burst intervals from distinct words at each level of the hierarchy of bursts based on their overlaps, thus yielding a nested set of clusters. We first group the two most overlapping intervals to form a new interval (or fragment) and proceed until no more overlapping intervals appear. Details are given in Algorithm 1. For each level $l \in [1, L]$ in the hierarchy of bursts H, the burst intervals contained at this level for each word w form a collection of intervals I_{l_w}. Each interval $I_{l_w}(i)$ in the collection has a start $S_{l_w}(i)$ and an end $E_{l_w}(i)$ point. An exhaustive comparison between the intervals in H is done independently for each level. If two burst intervals ($I_{l_u}(i)$, $I_{l_v}(j)$) overlap, they are merged together and a new interval is obtained ($I_{l_{u,v}}(t)$) and added to the collection. This step is done until there are no more overlapping intervals. In the end the fragments corresponding to the final intervals are extracted to represent the salient fragments at level l. The hierarchy of topically focused fragments is created using a mapping across levels of the fragments obtained. An example of such a hierarchy, of two levels, is presented in Figure 5. The limits of the fragments formed are given by the starting and ending utterance/sentence positions and their content is represented by a sample of the bursty words that contributed in forming them. These fragments pertain the most relevant information in the data at various levels of detail. The solution we propose to create the hierarchy of topically focused fragments has the advantage of deriving the hierarchy directly, without any prior on the duration of fragments (segments in case of traditional segmentation) and number of levels in

Algorithm 1 Create a hierarchy of topically focused segments.

for each level l **do**
  Step 0. Initialize segment clusters
  for all word w **do**
    $I_{l_w} = \{I_{l_w}(1), I_{l_w}(2), ...I_{l_w}(n_{l_w})\}$
    where $I_{l_w}(i) = [S_{l_w}(i), E_{l_w}(i)]$
  end for
  Step 1. Agglomerative clustering
  repeat
    for all $I_{l_u}(i), I_{l_v}(j) \in I_{l_w}, \forall u, v, \forall i, j, i \neq j$ **do**
      if $I_{l_u}(i) \cap I_{l_v}(j) \neq \emptyset$ **then**
        $I_{l_{u,v}}(t) = [min(S_{l_u}(i), S_{l_v}(j)), max(E_{l_u}(i), E_{l_v}(j))]$
        $add(I_{l_{u,v}}(t), I_{l_w})$
        $remove(I_{l_u}(i), I_{l_w})$
        $remove(I_{l_v}(j), I_{l_w})$
      end if
    end for
  until convergence
end for
Step 2. Mapping across levels
for $l = L \rightarrow 1$ **do**
  $I_{l_w}(i)$ mapped to $I_{l-1_w}(j)$ such that $I_{l_w}(i) \subset I_{l-1_w}(j)$
end for

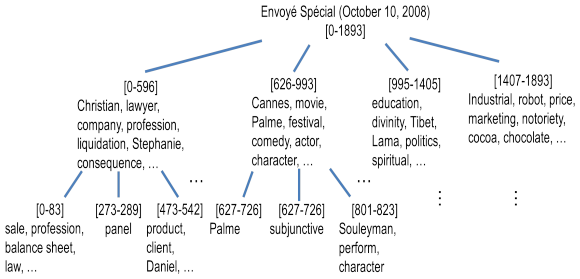


Figure 5: A two-level hierarchy of topically focused fragments obtained with a TV show. At each level, fragments are represented by their limits in terms of utterance number (in brackets) and characterized with the bursty words (translated from French) that helped form the fragments.

the hierarchy, unlike traditional hierarchical topic segmentation strategies.

5.2 Evaluation and discussion

Currently, there is no metric to evaluate the structure resulting from the above algorithm, the measures traditionally used for hierarchical topic segmentation being inappropriate for at least two reasons: 1- The structure that our algorithm outputs is a hierarchy of topically focused fragments and not a dense hierarchy of segments (cf. Figure 1); 2- there is no groundtruth for this hierarchy of topically focused fragments, which is required for the metrics used to evaluate traditional segmentations. Moreover building such a groundtruth is not an easy task: the topically focused fragments are obtained in a data-driven, bottom-up, manner that does not necessarily reflect a prior organization as would be provided by human experts; introducing this new way of thinking is indeed the main goal pursued by the paper. In addition of being costly, annotating new data requires that clear, shared, annotation guidelines be defined first. This last point requires a good understanding and characterization of what our approach can yield, which is exactly what this paper intends to provide. Therefore, to prove the relevance of our approach and provide a good insight into the hierarchical fragments that it outputs, we believe that it is important to see how focused fragments compare with traditional dense segmentation before moving further into annotating data with this new paradigm.

We thus report a number of measures relying on existing dense annotations: At each level, fragments are compared to their counterpart in the dense segmentation, after mapping. Conversely, dense segments are mapped to topically focused fragments. Two measures are defined: $M1$, the proportion of topically focused fragments belonging to a unique reference segment; $M2$, the percentage of reference segments which have at least one matching topically focused fragment. The values obtained with these measures both for a dense segmentation resulting from applying HierBayes and a hierarchy of topically focused fragments (HTFF) are reported in Table 1. Similar results are obtained on the medical corpus. For HierBayes we report only the results at two levels since trying to obtain more levels worsened the segmentation, resulting in the same segments at all levels. As going to lower levels with HTFF it is expected to have such a small coverage of the segments since their number is considerably high and the average number of bursts is ≈ 1. Results demonstrate that the fragments we extract in a bottom-up manner usually have an equivalent in a dense segmentation and have a stronger focus than their counterpart.

		HTFF		HierBayes	
Data-set	level	M1	M2	M1	M2
ES(manual)	1	0.75	1	0.51	1
	2	0.56	0.74	0.15	1
	3	0.47	0.17	–	–
ES(auto)	1	0.73	1	0.48	1
	2	0.46	0.62	0.1	1
	3	0.51	0.11	–	–
Wikipedia	1	0.22	0.97	0.29	1
	2	0.62	0.66	0.42	1
	3	0.69	0.29	–	–
	4	0.49	0.06	–	–

Table 1: The values obtained with M1 and M2 measures on two data sets after applying HierBayes and HTFF.

6 Conclusion

In this paper, we have investigated the relevance of bursts to organize the topical structure of textual content. We have shown that global measures of lexical re-occurrence are not adequate to detect topic shifts while the temporal distribution of word re-occurrences provides strong cues. As a consequence, we have proposed an algorithm to extract a hierarchy of topically focused fragments using agglomerative clustering of burst intervals. Comparison of this novel structure to a reference dense segmentation on several datasets has indicated that we can achieve a better topic focus than the one provided by the reference dense segmentation while retrieving the important aspects of a text.

References

Lucien Carroll. 2010. Evaluating hierarchical discourse segmentation. In *11th International Conf. of the North American Chapter of the Association for Computational Linguistics*, pages 993–1001.

Freddy Y. Y. Choi. 2000. A speech interface for rapid reading. In *Proceedings of IEE colloquium: Speech and Language Processing for Disabled and Elderly People*, pages 1–4.

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

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of 10th International Conf. of the North American Chapter of the Association for Computational Linguistics*, pages 353–361.

Inmar E. Givoni, Clement Chung, and Brendan J. Frey. 2011. Hierarchical affinity propagation. In *Proceedings of the 27th Conf. on Uncertainty in Artificial Intelligence*, pages 238–246.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Camille Guinaudeau. 2011. *Structuration automatique de flux télévisuels*. Ph.D. thesis, INSA de Rennes.

Anna Kazantseva and Stan Szpakowicz. 2014. Hierarchical topical segmentation with affinity propagation. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 37–47.

Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *8th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, pages 91–101.

Rasmus E. Madsen, David Kauchak, and Charles Elkan. 2005. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 545–552, New York, NY, USA. ACM.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA.

Paola Monachesi, Lothar Lemnitzer, and Kiril Simov. 2006. Language technology for elearning. In Wolfgang Nejdl and Klaus Tochtermann, editors, *Innovative Approaches for Learning and Knowledge Sharing*, volume 4227 of *Lecture Notes in Computer Science*, pages 667–672. Springer Berlin Heidelberg.

Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. 2005a. A bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 48–55. Association for Computational Linguistics.

Avik Sarkar, Anne De Roeck, and Paul Garthwaite. 2005b. Team re-occurrence measures for analyzing style. In S. Argamon, J. Karlgren, and J.G. Shanahan, editors, *Proceedings of the SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access.*, pages 28–36. ACM Press, August.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *39th Annual Meeting on the Association for Computational Linguistics*, pages 499–506.