Leveraging lexical cohesion and disruption for topic segmentation

Anca Şimon
Université de Rennes 1
IRISA & INRIA Rennes

Guillaume Gravier
CNRS
IRISA & INRIA Rennes

Pascale Sébillot
INSA de Rennes
IRISA & INRIA Rennes

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

Abstract

Topic segmentation classically relies on one of two criteria, either finding areas with coherent vocabulary use or detecting discontinuities. In this paper, we propose a segmentation criterion combining both lexical cohesion and disruption, enabling a trade-off between the two. We provide the mathematical formulation of the criterion and an efficient graph based decoding algorithm for topic segmentation. Experimental results on standard textual data sets and on a more challenging corpus of automatically transcribed broadcast news shows demonstrate the benefit of such a combination. Gains were observed in all conditions, with segments of either regular or varying length and abrupt or smooth topic shifts. Long segments benefit more than short segments. However the algorithm has proven robust on automatic transcripts with short segments and limited vocabulary reoccurrences.

1 Introduction

Topic segmentation consists in evidentiating the semantic structure of a document: Algorithms developed for this task aim at automatically detecting frontiers which define topically coherent segments in a text.

Various methods for topic segmentation of textual data are described in the literature, e.g., (Reynar, 1994; Hearst, 1997; Ferret et al., 1998; Choi, 2000; Moens and Busser, 2001; Utiyama and Isahara, 2001), most of them relying on the notion of lexical cohesion, i.e., identifying segments with a consistent use of vocabulary, either based on words or on semantic relations between words. Reoccurrences of words or related words and lexical chains are two popular methods to evidence lexical cohesion. This general principle of lexical cohesion is further exploited for topic segmentation with two radically different strategies. On the one hand, a measure of the *lexical cohesion* can be used to determine coherent segments (Reynar, 1994; Moens and Busser, 2001; Utiyama and Isahara, 2001). On the other hand, shifts in the use of vocabulary can be searched for to directly identify the segment frontiers by measuring the *lexical disruption* (Hearst, 1997).

Techniques based on the first strategy yield more accurate segmentation results, but face a problem of over-segmentation which can, up to now, only be solved by providing prior information regarding the distribution of segment length or the expected number of segments. In this paper, we propose a segmentation criterion combining both cohesion and disruption along with the corresponding algorithm for topic segmentation. Such a criterion ensures a coherent use of vocabulary within each resulting segment, as well as a significant difference of vocabulary between neighboring segments. Moreover, the combination of these two strategies enables regularizing the number of segments found without resorting to prior knowledge.

This piece of work uses the algorithm of Utiyama and Isahara (2001) as a starting point, a versatile and performing topic segmentation algorithm cast in a statistical framework. Among the benefits of this algorithm are its independency to any particular domain and its ability to cope with thematic segments

of highly varying lengths, two interesting features to obtain a generic solution to the problem of topic segmentation. Moreover, the algorithm has proven to be up to the state of the art in several studies, with no need of a priori information about the number of segments (contrary to algorithms in (Malioutov and Barzilay, 2006; Eisenstein and Barzilay, 2008) that can attain a higher segmentation accuracy). It also provides an efficient graph based implementation of which we take advantage.

To account both for cohesion and disruption, we extend the formalism of Isahara and Utiyama using a Markovian assumption between segments in place of the independence assumption of the original algorithm. Keeping unchanged their probabilistic measure of lexical cohesion, the Markovian assumption enables to introduce the disruption between two consecutive segments. We propose an extended graph based decoding strategy, which is both optimal and efficient, exploiting the notion of generalized segment model or semi hidden Markov models. Tests are performed on standard textual data sets and on a more challenging corpus of automatically transcribed broadcast news shows.

The seminal idea of this paper was partially published in (Simon et al., 2013) in the French language. The current paper significantly elaborates on the latter, with a more detailed description of the algorithm and additional contrastive experiments including more data sets. In particular, new experiments clearly demonstrate the benefit of the method in a realistic setting with statistically significant gains.

The organization of the article is as follows. Existing work on topic segmentation is presented in Section 2, emphasizing the motivations of the model we propose. Section 3 details the baseline method of Utiyama and Isahara before introducing our algorithm. Experimental protocol and results are given in Section 4. Section 5 summarizes the finding and concludes with a discussion of future work.

2 Related work

Defining the concept of theme precisely is not trivial and a large number of definitions have been given by linguists. Brown and Yule (1983) discuss at length the difficulty of defining a topic and note: *"The notion of 'topic' is clearly an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else, for it is appealed to very frequently in the discourse analysis literature. Yet the basis for the identification of 'topic' is rarely made explicit"*. To skirt the issue of defining a topic, they suggest to focus on topic-shift markers and to identify topic changes, what most current topic segmentation methods do.

Various characteristics can be exploited to identify thematic changes in text data. The most popular ones rely either on the lexical distribution information to measure lexical cohesion (i.e., word reoccurrences, lexical chains) or on linguistic markers such as discourse markers which indicate continuity or discontinuity (Grosz and Sidner, 1986; Litman and Passonneau, 1995). Linguistic markers are however often specific to a type of text and cannot be considered in a versatile approach as the one we are targeting, where versatility is achieved relying on the sole lexical cohesion.

The key point with lexical cohesion is that a significant change in the use of vocabulary is considered to be a sign of topic shift. This general idea translates into two families of methods, local ones targeting a local detection of lexical disruptions and global ones relying on a measure of the lexical cohesion to globally find segments exhibiting coherence in their lexical distribution.

Local methods (Hearst, 1997; Ferret et al., 1998; Hernandez and Grau, 2002; Claveau and Lefèvre, 2011) locally compare adjacent fixed size regions, claiming a boundary when the similarity between the adjacent regions is small enough, thus identifying points of high lexical disruption. In the seminal work of Hearst (1997), a fixed size window divided into two adjacent blocks is used, consecutively centered at each potential boundary. Similarity between the adjacent blocks is computed at each point, the resulting similarity profile being analyzed to find significant valleys which are considered as topic boundaries.

On the contrary, global methods (Reynar, 1994; Choi, 2000; Utiyama and Isahara, 2001; Ji and Zha, 2003; Malioutov and Barzilay, 2006; Misra et al., 2009) seek to maximize the value of the lexical cohesion on each segment resulting from the segmentation globally on the text. Several approaches have

been taken relying on self-similarity matrices, such as dot plots, or on graphs. A typical and state-of-the-art algorithm is that of Utiyama and Isahara (2001) whose principle is to search globally for the best path in a graph representing all possible segmentations and where edges are valued according to the lexical cohesion measured in a probabilistic way.

When the lengths of the respective topic segments in a text (or between two texts) are very different from one another, local methods are challenged. Finding out an appropriate window size and extracting boundaries become critical with segments of varying length, in particular when short segments are present. Short windows will render comparison of adjacent blocks difficult and unreliable while long windows cannot handle short segments. The lack of a global vision also makes it difficult to normalize properly the similarities between blocks and to deal with statistics on segment length. While global methods override these drawbacks, they face the problem of over-segmentation due to the fact that they mainly rely on the sole lexical cohesion. Short segments are therefore very likely to be coherent which calls for regularization introduced as priors on the segments length.

These considerations naturally lead to the idea of methods combining lexical cohesion and disruption to make the best of both worlds. While the two criteria rely on the same underlying principle of lexical coherence (Grosz et al., 1995) and might appear as redundant, the resulting algorithms are quite different in their philosophy. A first (and, to the best of our knowledge, unique) attempt at capturing a global view of the local dissimilarities is described in Malioutov and Barzilay (2006). However, this method assumes that the number of segments to find is known beforehand which makes it difficult for real-world usage.

3 Combining lexical cohesion and disruption

We extend the graph-based formalism of Utiyama and Isahara to jointly account for lexical cohesion and disruption in a global approach. Clearly, other formalisms than the graph-based one could have been considered. However, graph-based probabilistic topic segmentation has proven very accurate and versatile, relying on very minimal prior knowledge on the texts to segment. Good results at the state-of-the-art have also been reported in difficult conditions with this approach (Misra et al., 2009; Claveau and Lefèvre, 2011; Guinaudeau et al., 2012).

We briefly recall the principle of probabilistic graph-based segmentation before detailing a Markovian extension to account for disruption.

3.1 Probabilistic graph-based segmentation

The idea of the probabilistic graph-based segmentation algorithm is to find the segmentation into the most coherent segments constrained by a prior distribution on segments length. This problem is cast into finding the most probable segmentation of a sequence of t basic units (i.e., sentences or utterances composed of words) $W = u_1^t$ among all possible segmentations, i.e.,

$$\hat{S} = \arg\max_S P[W|S]P[S] \ . \qquad (1)$$

Assuming that segments are mutually independent and assuming that basic units within a segment are also independent, the probability of a text W for a segmentation $S = S_1^m$ is given by

$$P[W|S_1^m] = \prod_{i=1}^{m} \prod_{j=1}^{n_i} P[w_j^i|S_i] \ , \qquad (2)$$

where n_i is the number of words in the segment S_i, w_j^i is the j^{th} word in S_i and m the number of segments. The probability $P[w_j^i|S_i]$ is given by a Laplace law where the parameters are estimated on S_i, i.e.,

$$P[w_j^i|S_i] = \frac{f_i(w_j^i) + 1}{n_i + k} \ , \qquad (3)$$

where $f_i(w_j^i)$ is the number of occurrences of w_j^i in S_i and k is the total number of distinct words in W, i.e., the size of the vocabulary $\mathcal{V}$. This probability favors segments that are homogeneous, increasing when words are repeated and decreasing consistently when they are different. The prior distribution on segment length is given by a simple model, $P[S_1^m] = n^{-m}$, where n is the total number of words, exhibiting a large value for a small number of segments and conversely.

The optimization of Eq. 1 can be efficiently implemented as the search for the best path in a

weighted graph which represents all the possible segmentations. Each node in the graph corresponds to a possible frontier placed between two utterances (i.e., we have a node between each pair of utterances), the arc between nodes i and j representing a segment containing utterances u_{i+1} to u_j. The corresponding arc weight is the generalized probability of the words within segment $S_{i \to j}$ according to

$$v(i,j) = \sum_{k=i+1}^{j} ln(P[u_k|S_{i \to j}]) - \alpha ln(n)$$

where the probability is given as in Eq. 3. The factor α is introduced to control the trade-off between the segments length and the lexical cohesion.

3.2 Introduction of the lexical disruption

Eq. 2 derives from the assumption that each segment S_i is independent from the others, which makes it impossible to consider disruption between two consecutive segments. To do so, the weight of an arc corresponding to a segment S_i should take into account how different this segment is from S_{i-1}. This is typically handled using a Markovian assumption of order 1. Under this assumption, Eq. 2 is reformulated as

$$P[W|S_1^m] = P[W|S_1] \prod_{i=2}^{m} P[W|S_i, S_{i-1}] \ ,$$

where the notion of disruption can be embedded in the term $P[W|S_i, S_{i-1}]$ which explicitly mentions both segments. Formally, $P[W|S_i, S_{i-1}]$ is defined as a probability. However, arbitrary scores which do not correspond to probabilities can be used instead as the search for the best path in the graph of possible segmentations makes no use of probability theory. In this study, we define the score of a segment S_i given S_{i-1} as

$$\ln P[W|S_i, S_{i-1}] = \ln P[W_i|S_i] - \lambda \Delta(W_i, W_{i-1}) \quad (4)$$

where W_i designates the set of utterances in S_i and the rightmost part reflects the disruption between the content of S_i and of S_{i-1}. Eq. 4 clearly combines the measure of lexical cohesion with a measure of the disruption between consecutive segments: $\Delta(W_i, W_{i-1}) > 0$ measures the coherence between S_i and S_{i-1}, the substraction thus accounting for disruption by penalizing consecutive coherent segments. The underlying assumption is that the bigger $\Delta(W_i, W_{i-1})$, the weaker the disruption between the two segments. Parameter λ controls the respective contributions of cohesion and disruption.

We initially adopted a probabilistic measure of disruption based on cross probabilities, i.e., $P[W_i|S_{i-1}]$ and $P[W_{i-1}|S_i]$, which proved to have limited impact on the segmentation. We therefore prefer to rely on a cosine similarity measure between the word vectors representing two adjacent segments, building upon a classical strategy of local methods such as TextTiling (Hearst, 1997). The cosine similarity measure is calculated between vectors representing the content of resp. S_i and S_{i-1}, denoted $\mathbf{v}_i$ and $\mathbf{v}_{i-1}$, where $\mathbf{v}_i$ is a vector containing the (tf-idf) weight of each term of $\mathcal{V}$ in S_i. The cosine similarity is classically defined as

$$\cos(\mathbf{v}_{i-1}, \mathbf{v}_i) = \frac{\sum\limits_{v \in \mathcal{V}} \mathbf{v}_{i-1}(v) \ \mathbf{v}_i(v)}{\sqrt{\sum\limits_{v \in \mathcal{V}} \mathbf{v}_{i-1}^2(v) \ \sum\limits_{v \in \mathcal{V}} \mathbf{v}_i^2(v)}} \ . \quad (5)$$

$\Delta(W_i, W_{i-1})$ is calculated from the cosine similarity measure as

$$\Delta(W_i, W_{i-1}) = (1 - \cos(\mathbf{v}_{i-1}, \mathbf{v}_i))^{-1} \ , \quad (6)$$

thus yielding a small penalty in Eq. 4 for highly disrupting boundaries, i.e., corresponding to low similarity measure.

Given the quantities defined above, the algorithm boils down to finding the best scoring segmentation as given by

$$\hat{S} = \arg\max_S \sum_{i=1}^{m} ln(P[W_i|S_i]) - \lambda \sum_{i=2}^{m} \Delta(W_i, W_{i-1}) - \alpha m ln(n) \ . \quad (7)$$

3.3 Segmentation algorithm

Translating Eq. 7 into an efficient algorithm is not straightforward since all possible combinations of adjacent segments need be considered. To do so in a graph based approach, one needs to keep separated the paths of different lengths ending in a given node. In other words, only paths of the same length ending

at a given point, with different predecessors, should be recombined so that disruption can be considered properly in subsequent steps of the algorithm. Note that, in standard decoding as in Utiyama and Isahara's algorithm, only one of such paths, the best scoring one, would be retained. We employ a strategy inspired from the decoding strategy of segment models or semi-hidden Markov model with explicit duration model (Ostendorf et al., 1996; Delakis et al., 2008).

Search is performed through a lattice $L = \{V, E\}$, with V the set of nodes representing potential boundaries and E the set of edges representing segments, i.e., a set of consecutive utterances. The set V is defined as

$$V = \{n_{ij} | 0 \leq i, j \leq N\} \quad ,$$

where n_{ij} represents a boundary after utterance u_i reached by a segment of length j utterances and $N = t+1$. In the lattice example of Fig. 1, it is trivial to see that for a given node, all incoming edges cover the same segment. For example, the node n_{42} is positioned after $u4$ and all incoming segments contain the two utterances $u3$ and $u4$. Edges are defined as

$$E = \{e_{ip,jl} | 0 \leq i, p, j, l \leq N; \\ i < j; i = j - l; L_{\min} \leq l \leq L_{\max}\} \quad ,$$

where $e_{ip,jl}$ connects n_{ip} and n_{jl} with the constraint that $l = j - i$ and $L_{\min} \leq l \leq L_{\max}$. Thus, an edge $e_{ip,jl}$ represents a segment of length l containing utterances from u_{i+1} to u_j, denoted $S_{i \to j}$. In Fig. 1, $e_{01,33}$ represents a segment of length 3 from n_{01} to n_{33}, covering utterances u_1 to u_3. To avoid explosion of the lattice, a maximum segment length $L_{\max}$ is defined. Symmetrically, a minimum segment size can be used.

The property of this lattice, where, by construction, all edges out of a node have the same segment as a predecessor, makes it possible to weight each edge in the lattice according to Eq. 4. Consider a node n_{ij} for which all incoming edges encompass utterances u_{i-j} to u_i. For each edge out of n_{ij}, whatever the target node (i.e., the edge length), one can therefore easily determine the lexical cohesion as defined by the generalized probability of Eq. 3 and the disruption with respect to the previous segment as defined by Eq. 6.

Algorithm 1 Maximum probability segmentation

Step 0. Initialization
$q[0][j] = 0 \quad \forall j \in [L_{\min}, L_{\max}]$
$q[i][j] = -\infty \quad \forall i \in [1, N], j \in [L_{\min}, L_{\max}]$

Step 1. Assign best score to each node
for $i = 0 \to t$ **do**
  for $j = L_{\min} \to L_{\max}$ **do**
    for $k = L_{\min} \to L_{\max}$ **do**
      /* extend path ending after u_i with a segment of length j with an arc of length k */

$$q[i+k][k] = \max \begin{cases} q[i+k][k], \\ q[i][j] + \\ \text{Cohesion}(u_{i+1} \to u_{i+k}) - \\ \lambda \Delta(u_{i-j} \to u_i; u_{i+1} \to u_{i+k}) \end{cases}$$

    end for
  end for
end for

Step 2. Backtrack from n_{Nj} with best score $q[N][j]$

Given the weighted decoding graph, the solution to Eq. 7 is obtained by finding out the best path in the decoding lattice, which can be done straightforwardly by scanning nodes in topological order. The decoding algorithm is summarized in Algorithm 1 with an efficient implementation in $o(NL^2_{\max})$ which does not require explicit construction of the lattice.

4 Experiments

Experiments are performed on three distinct corpora which exhibit different characteristics, two containing textual data and one spoken data. We first present the corpora before presenting and discussing results on each.

4.1 Corpora

The artificial data set of Choi (2000) is widely used in the literature and enables comparison of a new segmentation method with existing ones. Choi's data set consist of 700 documents, each created by concatenating the first z sentences of 10 articles randomly chosen from the Brown corpus, assuming each article is on a different topic. Table 1 provides

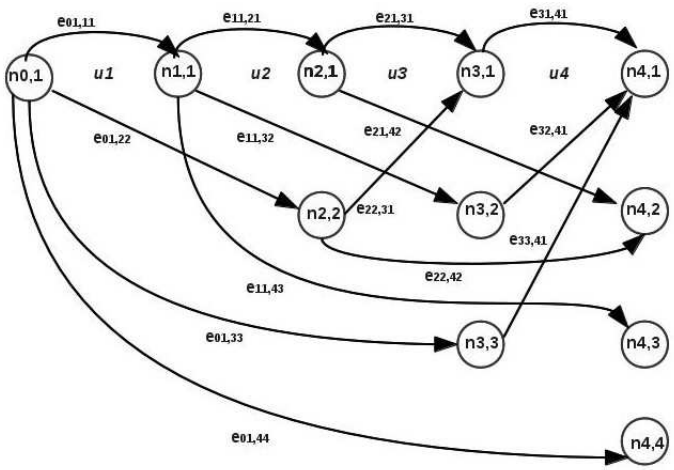


Figure 1: An example of a lattice L.

$z =$	3–11	3–5	6–8	9–11
# samples	400	100	100	100

Table 1: Number of documents in Choi's corpus (Choi, 2000).

the corpus statistics, where z=3–11 means z is randomly chosen in the range $[3, 11]$. Hence, Choi's corpus is adapted to test the ability of our model to deal with variable segments length, z=3–11 being the most difficult condition. Moreover, Choi's corpus provides a direct comparison with results reported in the literature.

One of the main criticism of Choi's data set is the presence of abrupt topic changes due to the artificial construction of the corpus. We therefore report results on a textual corpus with more natural topic changes, also used in (Eisenstein and Barzilay, 2008). The data set consists of 277 chapters selected from (Walker et al., 1990), a medical textbook, where each chapter—considered here as a document—was divided by its author into thematically coherent sections. The data set has a total of 1,136 segments with an average of 5 segments per document and an average of 28 sentences per segment. This data set is used to study the impact of smooth, natural, topic changes.

Finally, results are reported on a corpus of automatic transcripts of TV news spoken data. The data set consists of 56 news programs (≈1/2 hour each), broadcasted in February and March 2007 on the French television channel France 2, and transcribed by two different automatic speech recognition (ASR) systems, namely IRENE (Huet et al., 2010) and LIMSI (Gauvain et al., 2002), with respective word error rates (WER) around 36 % and 30 %. Each news program consists of successive reports of short duration (2-3 min), possibly with consecutive reports on different facets of the same news. The reference segmentation was established by associating a topic with each report, i.e., placing a boundary at the beginning of a report's introduction (and hence at the end of the closing remarks). The TV transcript data set, which corresponds to some real-world use cases in the multimedia field, is very challenging for several reasons. On the one hand, segments are short, with a reduced number of repetitions, synonyms being frequently employed. Moreover, smooth topic shifts can be found, in particular at the beginning of each program with different reports dedicated to the headline. On the other hand, transcripts significantly differ from written texts: no punctuation signs or capital letters; no sentence structure but rather utterances which are only loosely syntactically motivated; presence of transcription errors which may imply an accentuated lack of word repetitions.

All data were preprocessed in the same way: Words were tagged and lemmatized with TreeTag-

ger[1] and only the nouns, non modal verbs and adjectives were retained for segmentation. Inverse document frequencies used to measure similarity in Eq. 5 are obtained on a per document basis, referring to the number of sentences in textual data and of utterances in spoken data.

4.2 Results

Performance is measured by comparison of hypothesized frontiers with reference ones. Alignment assumes a tolerance of 1 sentence on texts and of 10 seconds on transcripts, which corresponds to standard values in the literature. Results are reported using recall, precision and F1-measure. Recall refers to the proportion of reference frontiers correctly detected; Precision corresponds to the ratio of hypothesized frontiers that belong to the reference segmentation; F1-measure combines recall and precision in a single value. These evaluation measures were selected because recall and precision are not sensitive to variations of segment length contrary to the *Pk* measure (Beeferman et al., 1997) and do not favor segmentations with a few number of frontiers as *WindowDiff* (Pevzner and Hearst, 2002) (see (Niekrasz and Moore, 2010) for a rigorous analytical explanation of the biases of *Pk* and *WindowDiff*).

Several configurations were considered in the experiments; due to space constraints, only the most salient experiments are presented here. In Eq. 7, the parameter α, which controls the contribution of the prior model with respect to the lexical cohesion and disruption, allows for different trade-offs between precision and recall. For any given value of λ, α is thus varied, providing the range of recall/precision values attainable. Results are compared to a baseline system corresponding to the application of the original algorithm of Utiyama and Isahara (i.e., setting $\lambda = 0$). This baseline has been shown to be a high-performance algorithm, in particular with respect to local methods that exploit lexical disruption. Differences in F1-measure between this baseline and our system presented below are all statistically significant at the level of $p < 0.01$ (paired t-test).

Choi's corpus. Figure 2 reports results obtained on Choi's data set, each graphic corresponding to a specific variation in the size of the thematic segments forming the documents (e.g., 9 to 11 sentences for the top left graphic). Results are provided for different values of λ in terms of F1-measure boxplots, i.e., variations of the F1-measure when α varies (same range of variation for α considered for each plot), where the leftmost boxplot, denoted by UI, corresponds to the baseline. Box and whisker plots graphically depicts the distribution of the F1-measures that can be attained by varying α, plotting the median value, the first and third quartile and the extrema.

z	τ	F1 gain	Confidence interval 95 %	
			UI	Combined
3-5	0	-0.2	[66.6,74.26]	[75.23,78.08]
3-5	1	0.7	[72.25,83.4]	[87.88,92.13]
3-11	1	0.23	[68.5,79.3]	[86.6,87.43]
6-8	1	0.4	[68.48,80.99]	[76.9,85.17]
9-11	0	1.6	[64.35,75.16]	[81.31,84.86]
9-11	1	1.4	[68.39,80.39]	[84.37,88.9]

Table 2: Gain in F1-measure for Choi's corpus when using lexical cohesion and disruption, and the corresponding 95 % confidence intervals for the F1-measure. Results are reported for different tolerance τ. UI denotes the baseline and Combined the proposed model.

Figure 2 shows that, whatever the segments length, results globally improve according to the importance given to the disruption (λ variable). Moreover, the variation in F1-measure diminishes when disruption is considered, thus indicating the influence of the prior model diminishes. When the segments size decreases (see Figs. 2(b), 2(c), 2(d)), the difference in the maximum F1-measure between our results and that of the baseline lowers, however still in favor of our model. This can be explained by the fact that our approach is based on the distribution of words, thus more words better help discriminate between potential thematic frontiers. Finally, using too large values for λ can lead to under-segmentation, as can be seen in Fig. 2(d) where, for $\lambda = 3$, the variation of F1-measure increases and the distribution becomes negatively skewed (i.e., the median is closer to the third quartile than to the first).

These results are confirmed by Table 2 which presents the gain in F1-measure (i.e., the difference between the highest F1-measure obtained when

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

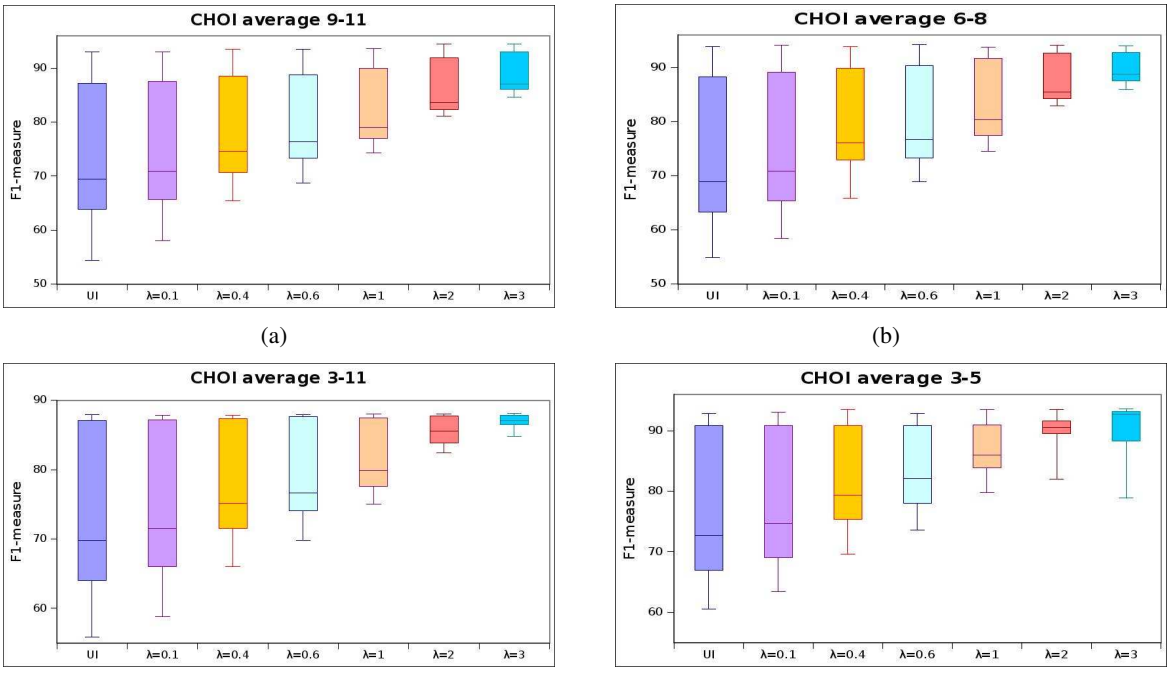


Figure 2: F1-measure variation obtained on Choi's corpus. In each graphic, the leftmost boxplot UI corresponds to results obtained by using the sole lexical cohesion (baseline), while the λ value is the importance given to the lexical disruption in our approach. Results are provided for the same range of variation of factor α, allowing a tolerance of 1 sentence between the hypothesized and reference frontiers.

combining lexical cohesion and disruption and the highest value for the baseline) for each of the four sets of documents in Choi's corpus, together with the 95 % confidence intervals: The effect of using the disruption is higher when segment size is longer, whether evaluation allows or not for a tolerance τ between the hypothesized frontiers and the reference ones. A qualitative analysis of the segmentations obtained confirmed that employing disruption helps eliminate wrong hypothesis and shift hypothesized frontiers closer to the reference ones (explaining the higher gain at tolerance 0 for 9-11 data set). When smaller segments—thus few word repetitions—and no tolerance are considered (e.g., 3–5), disruption cannot improve segmentation. Our model is globally stable with respect to segment length, with relatively similar gain for 3–11 and 6–8 data sets in which the average number of words (distinct or repeated) is close.

Results discussed up to now are optimistic as they correspond to the best F1 value attainable computed a posteriori. Stability of the results was confirmed

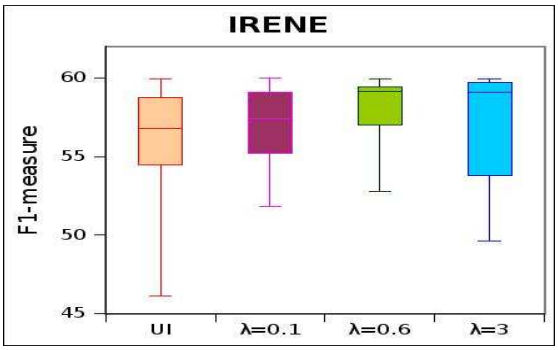
$z =$	3–5	3–11	6–8	9–11
UI	91.9	87.0	93.1	92.8
Combined	92.9	87.5	93.5	94.0

Table 3: F1 results using cross-validation on Choi's data set.

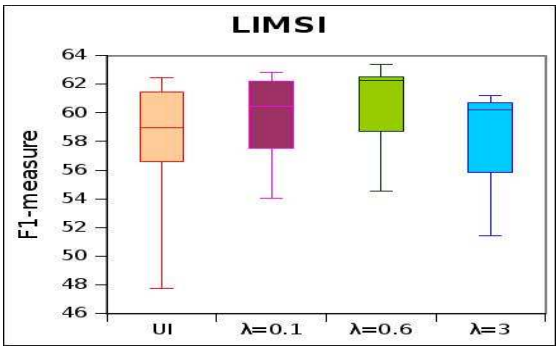
using cross-validation with 5 folds (10 folds for z=3–11): Parameters λ and α maximizing the F1-measure are determined on all but one fold, this last fold being used for evaluation. Results, averaged over all folds, are reported in Tab. 3 for the baseline and the method combining cohesion and disruption.

Medical textbook corpus. The medical textbook corpus was previously used for topic segmentation by Eisenstein and Barzilay (2008) with their algorithm BayesSeg[2]. We thus compare our results with those obtained by BayesSeg and by the baseline. When considering the best F1-measure (i.e., the best F1-measure which can be achieved by varying α and

[2]The code and the data set are available at http://groups.csail.mit.edu/rbg/code/bayesseg/



(a)



(b)

Figure 3: Boxplots showing F1-measure variation on transcripts obtained using IRENE and LIMSI automatic speech recognition systems.

λ), we achieved an improvement of 2.2 with respect to BayesSeg when no tolerance is allowed, and of 0.5 when the tolerance is of 1 sentence. The corresponding figures with respect to the baseline are 0.6 and 0.4. When considering the F1-measure value for which the number of hypothesized frontiers is the closest to the number of reference boundaries, improvement is of resp. 1.5 and 0.5 with respect to BayesSeg, -0.1 and 0.4 with respect to the baseline. These results show that our model combining lexical cohesion and disruption is also able to deal with topic segmentation of corpora from a homogeneous domain, with smooth topic changes and segments of regular size.

One can argue that the higher number of free parameters in our method explains most of the gain with respect to BayesSeg. While BayesSeg has only one free parameter (as opposed to two in our case), the number of segments is assumed to be provided as prior knowledge. This assumption can be seen as an additional free parameter, i.e., the number of segments, and is a much stronger constraint than we are using. Moreover, cross-validation experiments on the Choi data set show that improvement is not due to over-fitting of the development data thanks to an additional parameter. Gains on development set with parameters tuned on the development set itself and with parameters tuned on a held-out set in cross-validation experiments are in the same range.

TV news transcripts corpus Figure 3 provides results, in terms of F1-measure variation, for TV news transcripts obtained with the two ASR systems. On this highly challenging corpus, with short segments, wrongly transcribed spoken words, and thus few word repetitions, the capabilities of our model to overcome the baseline system are reduced. Yet, an improvement of the quality of the segmentation of these noisy data is still observed, and general conclusions are quite similar—though a bit weaker—to those already made for Choi's corpus. Results are confirmed in Table 4 which presents the gain in F1-measure of our model together with the 95 % confidence interval, where F1-measure values correspond to that of segmentations with a number of hypothesized frontiers the closest to the reference. The two first lines show that the gain is smaller for IRENE transcripts which have a higher WER, thus fewer words available to discriminate between segments belonging to different topics. The impact of transcription errors is illustrated in the last three lines, when segmenting six TV news for which manual reference transcripts were available (line 3), where the higher the WER, the smaller the F1-measure gain.

5 Conclusions

We have proposed a method to combine lexical cohesion and disruption for topic segmentation. Experimental results on various data sets with various characteristics demonstrate the impact of taking into account disruption in addition to lexical cohesion. We observed gains both on data sets with segments of regular length and on data sets exhibiting segments of highly varying length within a document. Unsurprisingly, bigger gains were observed on doc-

Corpus	F1 gain	Confidence interval 95 % UI	Combined
IRENE	0.3	[54.4,57.6]	[56.92,59]
LIMSI	0.86	[56.7,60.2]	[59.44,61.95]
MANUAL (6)	0.77	[70.39,72.29]	[71.7,73.29]
IRENE (6)	0.2	[56.81,60.94]	[59.51,63.43]
LIMSI (6)	0.5	[64.27,68.64]	[67.7,71.56]

Table 4: Gain in F1-measure for TV news corpus automatic and manual transcripts when using lexical cohesion and disruption, and the corresponding 95 % confidence intervals. Last three rows report results on only 6 shows for which manual reference transcripts are available.

uments containing relatively long segments. However the segmentation algorithm has proven to be robust on automatic transcripts with short segments and limited vocabulary reoccurrences. Finally, we tested both abrupt topic changes and smooth ones with good results on both. Further work can be considered to improve segmentation of documents characterized by small segments and few words repetitions, such as using semantic relations or vectorization techniques to better exploit implicit relations not considered by lexical reoccurrence.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *2nd Conference on Empirical Methods in Natural Language Processing*, pages 35–46.

Gillian Brown and George Yule. 1983. Discourse analysis. *Cambridge University Press*.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *1st International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 26–33.

Vincent Claveau and Sébastien Lefèvre. 2011. Topic segmentation of TV-streams by mathematical morphology and vectorization. In *12th International Conference of the International Speech Communication Association*, pages 1105–1108.

Manolis Delakis, Guillaume Gravier, and Patrick Gros. 2008. Audiovisual integration with segment models for tennis video parsing. *Computer Vision and Image Understanding*, 111(2):142–154.

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

Olivier Ferret, Brigitte Grau, and Nicolas Masson. 1998. Thematic segmentation of texts: Two methods for two kinds of texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 392–396.

Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. 2002. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1–2):89–108.

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

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, June.

Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. 2012. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech and Language*, 26(2):90–104.

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

Nicolas Hernandez and Brigitte Grau. 2002. Analyse thématique du discours : segmentation, structuration, description et représentation. In *5e colloque international sur le document électronique*, pages 277–285.

Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. 2010. Morpho-syntactic post-processing of n-best lists for improved French automatic speech recognition. *Computer Speech and Language*, 24(4):663–684.

Xiang Ji and Hongyuan Zha. 2003. Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 322–329.

Diane J. Litman and Rebecca J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 108–115.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappe. 2009. Text segmentation via topic modeling: an analytical study. In *Proc. ACM conference on Information and knowledge management*, pages 1553–1556.

Marie-Francine Moens and Rik De Busser. 2001. Generic topic segmentation of document texts. In *24th*

International Conference on Research and Developement in Information Retrieval, pages 418–419.

John Niekrasz and Johanna D. Moore. 2010. Unbiased discourse segmentation evaluation. In *Spoken Language Technology*, pages 43–48.

Mari Ostendorf, Vassilios V. Digalakis, and Owen A. Kimball. 1996. From HMM's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378.

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

Jeffrey C. Reynar. 1994. An automatic method of finding topic boundaries. In *32nd Annual Meeting on Association for Computational Linguistics*, pages 331–333.

Anca Simon, Guillaume Gravier, and Pascale Sébillot. 2013. Un modèle segmental probabiliste combinant cohésion lexicale et rupture lexicale pour la segmentation thématique. In *20e conférence Traitement Automatique des Langues Naturelles*, pages 202–214.

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.

Kenneth H. Walker, Dallas W. Hall, and Willis J. Hurst. 1990. *Clinical Methods: The History, Physical, and Laboratory Examinations*. Butterworths.