Hierarchical Topic Segmentation of TV shows Automatic Transcripts

Anca-Roxana Simon

Hierarchical Topic Segmentation of TV shows Automatic Transcripts

Internship Report

Anca-Roxana Simon
Master 2 - Research in Computer Science
Anca-Roxana.Simon@insa-rennes.fr

Supervisors: **Pascale Sébillot, Guillaume Gravier**

INSA de Rennes, IRISA, TexMex-Team

June, 2012

Abstract

The growth in the collections of multimedia documents made the development of new data access and data structuring techniques a necessity. The work presented in this report focuses on structuring TV shows and among the different kinds of structuring we approach the topic segmentation. Moreover we are interested in techniques able to provide hierarchical topic segmentation. The motivation for this research is defined by the potential impact of these techniques, since they fold perfectly on navigation and information retrieval subjects. In order to provide an automatic structuring of TV shows, that is generic, we use the words pronounced in TV shows, made available by their automatic textual transcription provided by an ASR system. The proposed topic segmentation algorithm consists in the recursive application of a modified version of TextTiling. It is based on the exploitation of a technique called vectorization, which was recently introduced for linear segmentation and outperformed the other existing techniques. We decided to study vectorization in more depth since it is a powerful technique and we tested it both for linear and hierarchical segmentation. The results obtained show that using vectorization can improve the segmentation and justify the interest of further applying such a technique.

Keywords: Automatic language processing, Automatic transcription, Hierarchical topic segmentation, Multimedia documents

Contents

1 Introduction

Existing studies show that video media has become one of the main ways for accessing information and entertainment by people [15]. In France each family had in 2007 an average of 1.8 TV, which were ON in average 6 hours/day. Also the Audiovisual National Institute (INA) has collections of multimedia data that equal more than 4 millions hours of videos, collections that grow with an average of 900,000 hours each year. After December 1st, 2008, INA started collecting 88 TV channels and 20 radio channels 365 days/year. In addition, videos, which can be accessed through the Internet, increase the volume of multimedia data available.

Because of this growth in the collections of multimedia documents, the development of new data access and data structuring techniques has become a necessity. The goal of such techniques is to facilitate the access to the information contained in documents while being generic enough for structuring different kinds of videos. Our concern is related to the structuring of these data and among the various types of structures that can be considered for organizing multimedia documents we focus on the *thematic structure* of TV shows. In order to provide such a structure, an isolation of the different subjects approached in each show needs to be performed, leading to a thematic organization of the data. Two kinds of methods for defining the thematic structure of TV shows at different levels can be of interest: linear and hierarchical topic segmentation.

These approaches have different objectives with respect to the targeted structure. *Linear topic segmentation* means dividing the data into thematically coherent segments. This allows users to search within a collection of TV shows, which have the same nature (e.g., a collection of TV-news from a certain period), for parts that contain the subject of interest. A rough structure of the data is obtained when applying this technique, because a homogeneous segment can in fact approach various aspects (sub-topics) of its main topic (e.g., a TV show concerning a war in a certain country can describe the facts of the war, recall the wars before from the same region, following with the effects of the war on the economy and politics of that country). Each sub-topic can be iteratively subdivided in other sub-topics. This process introduces the second approach for the thematic structure of TV shows, namely *hierarchical topic segmentation*. It is this kind of structuring that we are interested in. Performing hierarchical segmentation can provide a detailed organization of a TV show, offering users the possibility of zooming over a certain subject or having a general view over some fact. In addition it represents an essential step for most multimedia processing systems, which would offer a different view over their data. In [32] and [5], it is considered that the hierarchical representation of a text's topics is useful for information retrieval, text summarization, anaphora resolution and question answering. Furthermore it allows judging relevancy at different levels of details.

To obtain a thematic structure of the broadcasted TV shows, first of all, clues need to be extracted from the data. Afterwards, the methods previously presented have to be applied. Three types of sources of information can be retrieved from a video [24]:

1. visual, containing all that can be seen;
2. audio, containing the spoken words, the music, the background noise;
3. textual, containing text resources that represent the content of the multimedia document.

Some techniques use clues which are dependent on the type of data they can structure and are efficient only for certain cases (e.g., the mandatory presence of the anchor-person in a TV show that needs to be segmented [15]). Concerning our objective of automatic structuring of TV shows, we need to develop techniques that are generic, capable of segmenting and structuring any kind of TV-show. Therefore we need to use clues that are independent of the type of multimedia documents processed. For this reason we choose the speech pronounced in the TV shows, available through its automatic textual transcription provided by an *Automatic Speech Recognition system* (ASR system). Using the words pronounced during the TV shows has two advantages: words appear in any kind of TV shows and they provide access to the semantic content of these shows. Therefore our task of hierarchical topic segmentation relies on using the automatic transcripts of the words pronounced in the TV shows.

Few works, even for written text, address the problem of hierarchical segmentation. The solution we propose for providing this structuring of TV shows relies on recursively performing a linear segmentation. Such an approach already exists in the literature, the problem is that it does not perform well for levels in the hierarchy that are below the second one. In order to overcome the existing impediments for which the approach failed we take advantage of a vectorization technique, recently introduced for topic segmentation. Since we are interested in the problem of hierarchical segmentation in the context of automatic transcripts, we need to adapt the existing techniques for linear and hierarchical segmentation of written text, to the peculiarities of the transcripts, which makes the task harder.

The creation of an evaluation system is also a difficult task, since it generally consists in comparing the obtained segmentation to a reference one using a metric. Creating a reference segmentation is a time-consuming task and is affected by the lack of agreement between annotators regarding the granularity. To evaluate the proposed segmentation technique, we first use the Recall and Precision measures as employed in [15], and discuss, in a second time, another interesting approach for creating an evaluation system, presented in subsection 5.8.

The organization of the report is as follows: the specifics of the automatic transcripts obtained using an ASR system are detailed in section 2, in order to understand the peculiarities to which our methods need to adapt; then in section 3, the concept of topic segmentation is introduced, containing some fundamental theoretical notions and the techniques used for realizing it. In section 4, existing linear and hierarchical topic segmentation methods are presented, together with research work focusing on these concepts. Section 5 is developed around the proposed model comprising details regarding its implementation, evaluation and possible improvements. We end with some discussions concerning future work and several conclusions

regarding the work presented in this report.

2 Automatic TV-stream transcript characteristics

Automatic transcripts have various characteristics that differentiate them from written text. First of all, they do not contain punctuation signs or capital letters; thus they are not structured in sentences like classical texts, but in breath groups. These groups correspond to the words pronounced by a person between two breath intakes. In addition, the transcripts can contain an important number of wrongly transcribed spoken words. These errors can be due to the quality of the recording, to the presence of noise, to the difference of speaking styles or to the presence of words not contained in the dictionary of the ASR system. Indeed the role of an ASR system is to transform a speech signal into text using, among others, a dictionary. Besides these general peculiarities, TV shows also have several specifics. Contrary to radio data, where the spoken words must be necessarily understood by the listeners, for TV shows a bad reception of the spoken words can be compensated by the presence of subtitles or images, which increase even more the errors in transcription. In Figure 1 we present an example of automatic transcription and in the left side its reference transcription, extracted from [15].

Manual transcript	**Automatic transcript**
Dix-neuf cent quatre vingt-deux, un évènement vient de se produire, il s'appelle Amandine. Trois kilos quatre, cinquante et un centimètres, le premier bébé éprouvette français est né. Ici, le bébé exploit qui a un an soufflera ce mois-ci ses vingt-cinq bougies.	dix neuf cent quatre-vingt-deux un événement vient de se produire il s'appelle *amman dina* trois kilos quatre cinquante-et-un centimètres le premier bébé *éprouvait* franais est né ici le bébé *exploite y a* un an soufflera ce mois ci ses vingt-cinq bougies

Figure 1: Manual and automatic transcript of a TV journal extracted from France 2, 7/02/2007. The words in italic correspond to transcript errors [15]

3 Topic segmentation

Topic segmentation is the division of the data into segments, based upon the topic or subject discussed[1](i.e. placement of boundaries between utterances). The resulting segments should be topically coherent. This is a difficult task, intensively studied and debated. In this section, we present the fundamental notions regarding the concept of topic segmentation and the existing techniques used for fulfilling this

[1] `http://maebmij.org/~jim/topicsegmentation/node5.html`

task. A lot of research exists on the subject of topic segmentation of audio-video or textual documents. Besides the fact that the subject of topic segmentation represents an objective for many researchers, it also serves as a starting point for research in natural language processing, for the realization of automatic summaries for example.

In subsection 3.1, the concept of theme both in the general context and in the particular case of TV-news is defined. In subsection 3.2, the features exploited by the methods used to perform topic segmentation are presented. The manner in which these features are exploited is liable to the subjective definition of theme.

3.1 The concept of theme

Topic segmentation of a text implies the detection of its overall structure into topics [31]. In what follows we will define the general notions of *theme* and of *granularity at the level of theme*, and the specific notion of *theme in the concept of TV shows*.

Theme definition

The idea of theme is difficult to define precisely. Many linguists try to characterize this notion and they offer a large number of definitions for it. The problems that arise with these definitions is that they are not clearly independent. In [4], the difficulty of defining the notion of topic is discussed at length and the authors 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."
After many trials for establishing the concept of theme, they suggest, as an alternative, investigating *topic-shift* markers. It is considered that changes in topic can be identified easier.

In [35], the authors remarked a consensus between the annotators for placing the frontiers of thematic segments, between 82% and 92%, and strong irregularities for the same task as performed by different users, between 5.5% and 41.5%. This shows how difficult it is to define a single segmentation for the same document. This difference in segmentation seems to appear because of the different perceptions of the notion of theme and of its granularity. It is not possible to find a definition that is valid for all types of texts.

Theme in the context of TV-news broadcasts

The Topic Detection and Tracking (TDT) research project is focused on finding segments, in broadcast news, that are thematically correlated. To evaluate the methods used, a guide for defining the notion of event and topic was created. An event in the TDT context is something that occurs at a specific place and time associated with some specific actions. A topic is considered to be an event together with all the events directly related to it. The main difference between event and topic is that the event is relatively short and evolves in time, while the topic is more stable and

long.

Theme hierarchy

While the notion of theme has many definitions and was the objective of numerous studies, the concept of granularity or hierarchy at the level of themes was poorly approached in the linguistic literature.

In [38], the author proposes definitions of the notion of theme and of thematic granularity based on differential semantics. For differential semantics, the sense of a text emerges from the structuring of the sememes space, where sememes represent the words of the vocabulary. The sememes are defined one with respect to another through semes, which are semantic relations. Therefore any semantic unit can be broken down into semes [20]. A generic seme indicates that the sememe belongs to a semantic class. A specific seme distinguishes a sememe from all the other sememes of the same class. Based on these notions, Rastier defined the theme as a stable structure of semes. The difference between specific theme and generic theme can be seen as a representation of different granularity, in which the specific theme would be the sub-theme of the generic one.

In [6], the author focuses his work on discourse structuring and makes the difference between theme and sub-theme in the following manner: a coherent thematic segment, considered as a sequence of sentences that are interdependent, characterizes a sub-theme if its interpretation is dependent of another thematic segment. The organizers of the TDT project have also proposed various notions for topic and event, that can define different levels of granularity, regarding a news fact.

3.2 Characteristics used for topic segmentation

In this subsection, the various features that can be exploited for identifying the topic changes in a document are presented. These features are dependent on the type of data used, for text they are generally based on the notion of lexical cohesion, while for other data, like audiovisual documents, they are especially based on prosodic cues. First the notion of lexical cohesion is described, on which most of the segmentation methods rely. Other characteristics that can be used are detailed afterwards.

Lexical cohesion

The notion of lexical cohesion is based on the distribution of words. The key point is that a significant change in the vocabulary reveals a change in the topic. Therefore cohesive segments in text can be identified based on the words they contain and their positions. There are various ways of measuring the lexical cohesion in a text. It can be directly or indirectly (i.e., using reference documents) and based on the word repetitions or on lexical chains.

For the case of TV shows topic segmentation, lexical cohesion measures are very sensitive to the specifics of the transcripts (e.g., errors, few words repetitions). For this reason in [15], the author proposed various techniques for adapting the lexical cohesion criterion to the peculiarities of the automatic transcripts of the words pronounced in TV shows.

The studies relying on lexical cohesion can be based on locally detecting the *disruptions of the lexical cohesion* (local methods) or on the *measure of the lexical cohesion* (global methods).

In cohesive and coherent texts, words are likely to have references to concepts previously mentioned or concepts related to them. These references form lexical chains. Therefore a lexical chain can link one word, the pronouns that refer to it, other words semantically close or semantically related, etc.

In [31], the author focused on the automatic creation of a table of contents using lexical chains. Therefore the table of contents is built using the hierarchical and sequential relationships between topic segments that are identified in a text. The approach used for topic segmentation relies on the distribution of *topics* and *comments* (i.e., additions to the topic) in sentences, and on patterns of thematic progression in text. In the case of automatic transcripts of spoken words in TV shows [15], a lexical chain is created between two breath groups if the same word appears in each of the two breath groups, and its apparitions are separated by at most N seconds.

Linguistic markers

In order to obtain a topic segmentation, besides taking into consideration the lexical distribution information, various markers can be considered like: prosodic cues (e.g., intonational pitch, pause, duration), discourse markers (e.g., however, first, next), etc.

Discourse markers (DMs) have been the object of numerous studies in computational linguistics, since they have an important role in various discourse processing tasks. The role of DMs as indicators of discourse structure was described in [14]. The authors considered that "certain words and sentences and more subtle cues such as intonation or changes in tense and aspect" are "among the primary indicators of discourse segment boundaries". Another important linguistic marker for spoken communication is the prosody, or intonation. This may reflect various features of the utterances as presented in [16]: the emotional state of the speaker; whether an utterance is a statement, a question or a command; whether the speaker is emphasizing, contrasting or focusing a particular item.

4 State of the Art

This section presents the existing work on the subject of linear and hierarchical topic segmentation. As mentioned in the introduction section, the linear segmentation technique is used for structuring textual data into successive topics, while the hierarchical structure refers to the internal architecture of the data. In addition the existing evaluation techniques for the existing segmentation methods are presented.

4.1 Linear topic segmentation

Linear topic segmentation consists in evidentiating the semantic structure of a document. Therefore the algorithms developed for this task have the objective of automatically detecting the topic frontiers, which define topically coherent segments. As presented in section 3, for the task of topic segmentation there are various characteristics that can be taken into consideration, and in what follows we first present some of the works that exploit the prosodic and discourse markers.

Work done in [14] and [25] exploits discourse markers (e.g., first, finally, also) to improve the detection of thematic frontiers in oral documents. In [22], the authors examined automatic topic segmentation based on prosodic cues for English broadcast news. In [16], the authors proposed the use of prosodic information to improve an ASR-based topic tracking system for French TV broadcast news. The problem with these linguistic markers is that they are too dependent on the type of documents considered. The prosodic cues which work for TV news do not necessarily work for other TV shows. As mentioned in the introduction section we focus on segmenting TV reports. If for TV news the segmentation can benefit from additional information like pitch or intensity, or discourse markers that are used for introducing a new topic, for TV reports this kind of information can be misleading. The TV reports are characterized by outdoor investigations, where many people are interviewed. They have various accents, different ways of presenting the information, of emphasizing on topics, etc. It is difficult to find an agreement regarding these and the linguistic cues which could help distinguish between the different topics discussed. For this reason the employment of linguistic markers for topic segmentation can make the technique specific.

We are interested in techniques that can perform a generic structuring of documents. Such techniques usually exploit the lexical cohesion (especially the techniques used on textual data), which is independent of the type of textual documents considered and does not require a learning phase. As mentioned in section 3, the methods that rely on lexical cohesion can be divided in two families: local methods [17, 21, 13, 9] and global methods [39, 7, 45, 28, 30]. There exists however approaches that combine the two methods, like in [15], where the measure of the lexical cohesion is combined with the information of the disruption of that cohesion. In what follows we will detail some of the approaches proposed for local and global methods and also their combination.

4.1.1 Local methods

These methods are based on the local detection of lexical cohesion disruption. The TextTiling algorithm, presented in [17], is considered to be a fundamental algorithm for topic segmentation based on the analysis of the word distribution in a text. A significant change in the vocabulary used is considered to be a sign of topic shift. This algorithm is based on parsing the text using a sliding window. This window covers adjacent blocks of text and is centered in a point from the text that corresponds to a

possible frontier. The content before and after each possible boundary is represented through a vector. Each vector contains words that are weighted according to their frequency *tf* (i.e., term frequency):

$$w_{t,d} = tf(t,d)$$

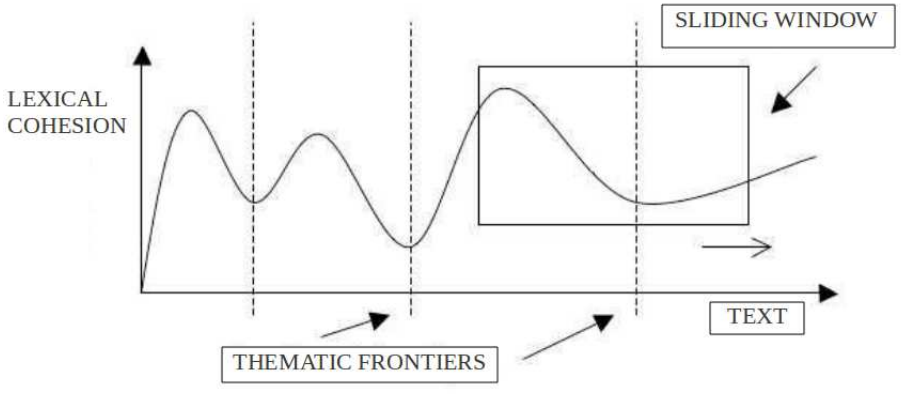
where $w_{t,d}$ is the weight associated to term[2] t of block d. A higher weight implies that the term is more relevant for the document. Different weights can be employed for this task, like tf-idf (term frequency-inverse document frequency), okapi (similar to tf-idf but takes better into account the lengths of the blocks), etc.

Then a similarity measure is computed between the two vectors. For this algorithm the cosine measure is used:

$$cos(b1,b2) = \frac{\sum_{t=1}^{n} w_{t,b1} \times w_{t,b2}}{\sqrt{(\sum_{t=1}^{n} w_{t,b1}^2) \times (\sum_{t=1}^{n} w_{t,b2}^2)}}$$

where $w_{t,b1}$ is the weight assigned to term t in block $b1$ and t ranges over all the terms in the document. As the angle between the vectors shortens, the cosine value approaches 1. This means that the two vectors are getting closer, and the similarity of what they represent increases. Thus if the similarity score between two blocks is high, then not only do the blocks have terms in common, but the terms they have in common are relatively rare with respect to the rest of the document[3]. This measure gives the lexical similarity between the two parts of the window.

Afterwards the window is shifted, in order to compute the value of the lexical cohesion at the level of another potential thematic frontier. The resulting sequence of similarity values, after being plotted and smoothed, is examined for peaks and valleys. High similarity values, implying that the adjacent blocks cohere well, tend to form peaks, while low similarity values, indicating a potential boundary between blocks, create valleys. The frontiers are identified by locating the lowermost portions of valleys in the resulted plot [17, 19, 18], as it can be seen in Figure 2 .



[2] *term* is used as a synonym for *word* and not in the context of terminology

[3] `http://www.miislita.com/information-retrieval-tutorial/cosine-similarity-tutorial.html`

Figure 2: Local thematic segmentation based on sliding window [15]

The TextTiling algorithm inspired many other research works, like [21, 13], that tried to improve the results obtained by modifying the vectorial representation and the measure of similarity computed. In order to consider the different occurrences of the terms in the data, many researchers have proposed to use *lexical chains* to represent the two blocks of the sliding window (Figure 3). A lexical chain connects the different occurrences of a word (or semantically related terms or references) as long as these occurrences are at a distance smaller than a threshold called *hiatus*. Therefore the repetitions of the words and also the locality of those repetitions can be taken into consideration. A frontier is proposed at places where few lexical chains are cut [33, 41].

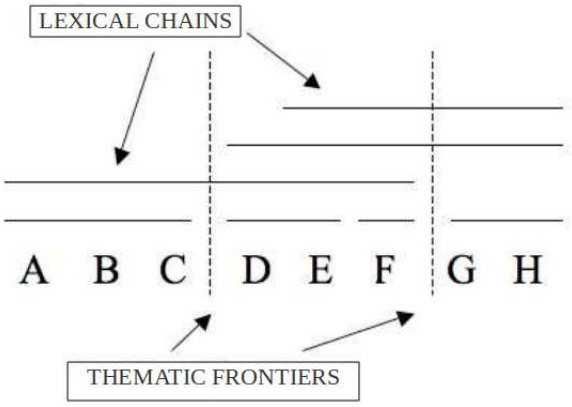


Figure 3: Local thematic segmentation based on lexical chains

Instead of directly comparing the representations of adjacent segments, [9] proposes a technique that uses indirect comparison for evidentiating the semantic similarity between two blocks of utterances, even if they do not share common vocabulary. This technique is called *vectorization*. It has been introduced and implemented in [10], in a standard Information Retrieval (IR) scenario and it has proven to have low complexity and accurate results. The principle of the proposed technique in [9] is the following:

- each document in the collection is compared with the same m pivot documents (i.e., reference documents) by computing a proximity score;
- for each document in the collection, the m scores obtained are gathered into a m-dimension vector representing the document (Figure 4);
- the comparison between two documents can then be performed by comparing their associated vectors (e.g., using a *L2* distance).

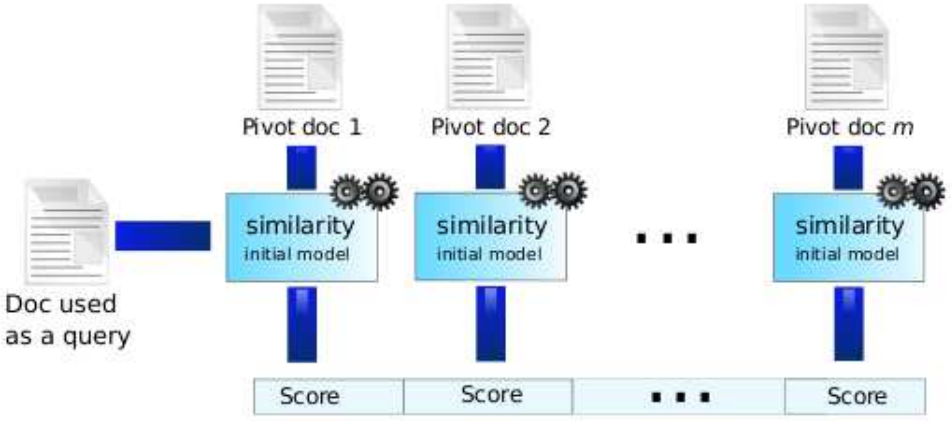


Figure 4: Vector design from pivot-documents [9]

The advantage of using the vectorization technique is that it offers two interesting properties. One is related to complexity reduction, respectively the construction of vectors associated to each document can be done offline and when a request needs to be processed, instead of computing its similarity with all documents it will be computed only with the m pivot documents. This property is useful in the IR context; however it does not apply for the segmentation task. The most interesting property is that two documents will be considered similar if they are similar to the same pivot documents. This indirect comparison overcomes the drawbacks represented by a poor repetition of the vocabulary and by the presence of synonyms.

This vectorization technique results in a change of the representation space, as opposed to other existing work consisting in a dimension reduction or in the approximation of the original distance as proposed in [3]. It is not about orthogonalisation as for Latent Semantic Indexing/Analysis (LSI/LSA) approaches, where the lexical vectors are projected into a latent concept space. The goal of LSI/LSA is to maintain dependencies between related terms and overcome the lack of word repetitions by generalizing words into semantic concepts probabilities [37]. The properties of vectorization make it more interesting than the techniques based on dimensional reduction.

In [9], the authors propose a topic segmentation of TV shows, based on automatic transcripts of the speech pronounced in the programs, by using mathematical morphology and vectorization. Their topic segmentation system is based on mathematical morphology principles, used for image segmentation. If for images the pixel represents the base element, for text the base element is represented by the sentence and it is described by the contained words. Therefore the texts are flows of utterances, which imply a 1-D representation, differently from the 2- or 3-D used for images. Still the watershed technique can be applied. Regarding the vectorization technique used, the m pivot documents are segments built from random splits of the broadcasts.

4.1.2 Global methods

These methods consist in a global comparison between all the regions of the document, searching to maximize the value of the lexical cohesion globally as defined by the frontiers. A lot of research work has been using these techniques, and various representations for the documents are considered. In [39] and [7], a matrix representation is considered, while in [45] and [28] a graph representation is used (Figure 5).

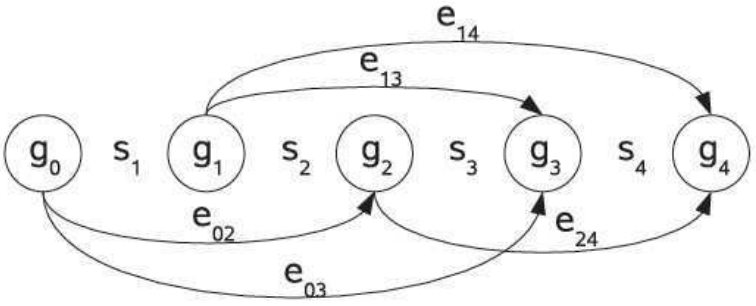


Figure 5: Graphic representation

The two approaches in [39] and [7] both rely on the computation of similarities between the candidate segments and then on a clustering based on the resulting similarity matrix. When considering a graph-based representation of the document to segment, the nodes of the graph represent potential frontiers and the edges thematic segments. In [28], the authors consider the segmentation task as a task to partition the graph, using the notion of normalized cut. In [45], the topic segmentation is provided by finding the best path in a weighted graph. The weight associated with each edge corresponds to a measure of the lexical cohesion of the segment it represents. The value of the lexical cohesion for a segment S_i is seen as the measure of the capacity of a language model, learnt on the segment S_i, to predict the words contained in the segment. In [30] the authors propose an extension of this algorithm which performs, in addition to segmentation, topic labeling. For this, they use a probabilistic topic modeling, Latent Dirichlet Allocation (LDA). They first identify the latent topics in a document and then modify the segmentation algorithm in [45] by associating with each edge in the graph a vector containing the probability of the segment corresponding to that edge to be connected with the latent topics detected. LDA is used for computing this probability. The work in [30] is different from other approaches that use LDA for segmentation [42] also because the data for training is not the same as the one used for segmentation, which helps estimating the LDA parameters more reliably. For this reason the performance reported in [42] is not

significantly better than that of TextTiling[17].

Global methods can also be applied for finding topics through generative models. LDA, previously mentioned, is such a technique. The documents can be modeled as being generated from some sequence of topics. These topics have their own characteristic word distribution. Therefore by infering the most likely sequence of topics from the observed words, the positions of the boundaries between them can be derived. The basis for such techniques is represented by Hidden Markov Models (HMMs). In [23], the authors applied a HMM topic model for segmenting broadcast news and the results obtained were promising. However, for this task they needed a segmented training dataset to estimate the topic transition probability and the topic language models. Other researchers use a probabilistic form of latent concept model within a HMM. This way, segments which may be related to multiple underlying concepts are taken into account [11, 44, 34]. The limitation of parametric topic models (LDA) is represented by the difficulty of determining the number of topics for a corpus (i.e., computing optimal number of topics is time-consuming and the optimal number of topics varies for each corpus)[4].

4.1.3 Combination of a local and a global method

In [15], the author proposes the use of a global method based on lexical cohesion, in order to deal with the size variability of the segments and the use of a local method to detect the disruptions of the lexical cohesion. She considers that by combining these two approaches the thematic segmentation obtained can provide segments that are both very coherent, and very different from each others. The global method used is the algorithm proposed in [45], which does not need a priori information regarding the number of segments expected. For the detection of the disruption of the lexical cohesion, the technique proposed in [9] was employed. The topic segmentation technique was adapted to the peculiarities of automatic transcripts. Other works that combine the characteristics that can be used for topic segmentation are [26, 27]. Their approaches use supervised classifiers and take advantage of linguistic markers making their solutions more specific. In [43] the authors use a Naive Bayes classifier, and lexical cohesion scores are combined with syntactic features and information regarding the speaker's identity. Using combined approaches can improve the accuracy of the segmentation since they can bring different insights to the problem. Therefore it is the direction most recent systems have taken [37].

4.1.4 Linear topic segmentation evaluation methods

In order to evaluate the quality of a linear topic segmentation produced by a method, the authors usually compare the obtained segmentation with a reference one using a metric. Therefore a reference segmentation needs to be made, which is time-consuming and is affected by the disagreement between the annotators. As men-

[4] `http://www.cs.princeton.edu/~blei/kdd-tutorial.pdf`

tioned in section 3, theme identification is subjective. For these reasons researchers have proposed various solutions.

In [7], the author proposes an evaluation of segmentation methods by creating a corpus of 700 documents made by the concatenation of portions of texts, taken from articles, randomly chosen from the Brown corpus. Therefore different authors can compare their methods with existing ones, based on a common corpus test. The drawback of this solution is that the thematic changes happen very suddenly, which is not the case in classic documents.

To choose the metric, evaluation measures from the IR field can be used, like precision and recall. The recall corresponds to the portion of reference frontiers detected by the method. The precision corresponds to the ratio of produced frontiers that belong to the reference segmentation. In [1], the author mentions several limitations of the precision and recall measurements and proposes another measure: *Pk*. This solution is based on a sliding window of size k, which parses the reference segmentation and the hypothetical segmentation proposed by the system under evaluation. *Pk* consists in evaluating the similarity between the two segmentations inside the window.

The authors of [36] analysed *Pk* and observed several limitations. One of them is that *Pk* is too sensitive to the variations of the segments dimensions. Therefore they have proposed *WindowDiff*, which considers the number of frontiers between two sentences separated by a distance k. This approach is more resistant to the dimensions of segments. Because both *Pk* and *WindowDiff* employ the use of a sliding window, lower weights are given to the frontiers close to the beginning or ending of a document. In addition, the author of [5] considers that *WindowDiff* penalizes the false positives and false negatives equally; thus segmentations with fewer number of frontiers are favored.

4.2 Hierarchical topic segmentation

Compared to the important literature about linear segmentation few research has been done on the subject of hierarchical topic segmentation and its evaluation. It is widely agreed [14, 5, 12] that the discourse structure often displays a hierarchical form, while a lot of disagreement appears regarding its units and elementary relations. The main problem regarding this kind of segmentation technique is that, at each step, the number of words considered decreases and makes the lexical-based methods difficult to apply.

To obtain a hierarchical structuring, the authors of [45] propose to apply their algorithm for linear thematic segmentation iteratively over the thematically homogeneous segments obtained after a previous iteration. In [15], the author takes into consideration their suggestion and applies the iterative technique. For this, the linear topic segmentation algorithm is modified to reflect the distribution of the vocabulary at different levels in the hierarchy (adjusting the lexical cohesion computation and considering lexical chains). This work helps defining a hierarchical thematic segmentation for two levels, remaining a challenge to define it at all the levels in the hierarchy.

The authors of [32] and [12] consider that the distribution of words in a text is a very important clue to extract the hierarchical structure. In [32], they use this distribution to construct lexical chains for inferring the cohesive structure of the text. A chain is built for each important content term (i.e., terms remaining after removing stopwords[5]), containing the term and its related terms (including the resolved anaphors). The authors consider that the starts, interruptions and terminations of lexical chains give valuable information on topic boundaries. Therefore they propose a bottom-up approach that consists in connecting the segments thematically homogeneous that have a hierarchical link. In [12], an unsupervised method for hierarchical topic segmentation is presented. This segmentation is formalized using a Bayesian probabilistic framework. It is based on the hypothesis that each word of the text is represented by a language model estimated on a portion, more or less important, of the text. This proposed approach is the only one that allows obtaining a hierarchical thematic segmentation without having to perform a linear segmentation before. However, it does not fit with our objectives of obtaining a generic and unsupervised segmentation. Indeed the algorithm proposed requires an input parameter that specifies the expected dimensions of the segments at each level of hierarchy. Another drawback is that the sizes of the segments returned are very regular.

The hierarchical topic segmentation can also be viewed as a clustering task. Therefore the algorithm proposed in [47], based on an agglomerative hierarchical clustering, could be exploited for extracting such a structure. In [40], the authors combine LSI with a technique used for signal processing (scale-space segmentation). The documents are represented as matrices due to LSI [8]. Each sentence S_i from a text is represented by a vector corresponding to the i^{th} column of the matrix. Singular value decomposition allows reducing the number of dimensions to k by retaining only the k most important words. After that the authors apply scale-space segmentation, which consists in smoothing each dimension of the vectors independently. This is done by using a Gaussian kernel associated with multiple values of scale σ. The importance of the thematic frontiers is defined for each vector by analysing the smoothing at different levels of scale. The difference of importance between frontiers will determine the hierarchical aspect of the segmentation. There also exist generative models that address the problem of learning topic hierarchy from data, through hierarchical topic models [46, 2, 29]. Therefore an adaptation of such techniques for obtaining hierarchical topic segmentation could provide promising results.

4.2.1 Hierarchical topic segmentation evaluation methods

The evaluation of hierarchical thematic segmentation faces to the same problems encountered for evaluating the linear thematic segmentation (producing reference segmentation, defining a metric). The problem of subjectivity regarding the concept of theme is however more accentuated than for the linear segmentation. As mentioned in section 3, the definition of a theme at different levels of granularity is

[5]words that appear frequently in the data without providing important information

a challenging task. Several studies, that described hierarchical discourse segmentation algorithms have been presented in the first part of this subsection, but none of them rigorously evaluated the segmentation in its hierarchical form.

In [40], the authors evaluated their algorithm by visual comparison with the ground truth (in this case, it is represented by chapter headings and sub-headings and their titles). In [32], the results obtained from the segmentation are used in a system for automatic summary generation, leading to an indirect evaluation of the segmentation. In [12], the author evaluated the algorithm against three recursive segmentation algorithms on a corpus that had just two levels of segment depth. Only in [5] an evaluation method dedicated to the hierarchical thematic segmentation is proposed. An extension of the *Pk* measure is presented. This measure computes, for each level i in the hierarchy, the value of *Pk* between the reference frontiers and the hypothesizes frontiers at a level greater or equal to i. This constraint makes the evaluation system not able to characterize the behavior of a real segmentation. In [15], the author evaluates the segmentation at each level separately. The drawbacks are that a global error cannot be computed (an error at a higher level in hierarchy should influence the segmentations at a lower level) and making reference segmentations for each level is difficult.

In this section we have presented the existing methods, and in the following section we will describe the work done during the internship.

5 Contribution

Regarding linear topic segmentation, several solutions exist for written text and some are applied to automatic transcripts of speech. As for hierarchical segmentation, few solutions exist even for written text. Therefore the objective of our work is to provide techniques for performing hierarchical topic segmentation of automatic transcripts of TV shows. What we want is a technique that is generic, and the vectorization technique (c.f., section 4.1.1) seems to be a good approach for this task. This technique is powerful not only because it can perform well even when the segments are small and are characterized by a small number of words, but also because it allows changing and adapting the pivot-documents to the different topics and levels in the hierarchy.

In subsection 5.1 we describe the way in which we use this technique and we position our work with respect to the existing approaches on this subject. The validation of the model follows for justifying the choice of the proposed model. Then the experiments performed for evaluating the model are detailed. In the last part a discussion regarding the advantages and drawbacks of the approach considered and some techniques that could improve it are provided.

5.1 Positioning and the proposed model

Among the approaches presented in section 4 for hierarchical topic segmentation, the most promising one is the one proposed by Eisenstein [12], which aims at directly obtaining a hierarchical segmentation without passing through a linear one first. However this method does not respond to the objectives set for the work presented in this report because it is not generic and does not imply unsupervised structuring as it requires an input parameter that specifies the expected dimensions of the segments at each level of the hierarchy. Moreover, evaluating the algorithm on our corpus of TV reports for the first two levels of the hierarchy, we observe that it does not perform well: for the first level the segments obtained correspond to the reference segmentation; however for the second one the segments returned are very regular in duration, characteristic that does not correspond to our data. Therefore we considered using another approach, based on a recursive application of a linear segmentation algorithm.

Since we are interested in developing a generic technique we do not take into consideration specific clues, like linguistic markers, instead we use the lexical cohesion criterion. Using this criterion, which performs well for linear segmentation is not straightforward when applied for the hierarchical one. At lower levels in the hierarchy the segments do not share that much vocabulary, therefore it is difficult to connect segments belonging to the same topic. In order to overcome this, we use a vectorization technique, like the one proposed in [9], which uses pivot-documents to compute the similarities between segments. For the work described in this report we propose to study the contribution of vectorization for the hierarchical segmentation of multimedia documents. We differentiate our work apart from [9] in three ways: by performing hierarchical topic segmentation and, applying a different algorithm for this task, TextTiling, and by exploring the use of external pivot-documents in addition to a more refined study in the case of internal pivot documents.

For the model considered in this work, the vectorization technique is integrated in a topic segmentation algorithm, a modified version of TextTiling. The reason behind this is that we are interested in the impact of using pivot documents for obtaining a hierarchical segmentation. Therefore we analyze in more depth the selection of pivot documents and how they can improve the results even for the basic approach for topic segmentation, TextTiling. In order to obtain a hierarchical segmentation the algorithm is applied recursively. Beside the use of a different segmentation algorithm our work differs in the way the pivot documents are selected. In [9] the authors constructed the pivot documents using the same data as for segmentation. They performed a division of the reports in multiple segments of varying sizes. There are several questions that arise when performing such a selection of pivot documents, respectively: Could this kind of division perform well for finding topics at a lower level? For a lower level should the data be divided even more? If so, how much can we divide it, in order to be able to overcome the small number of available words at lower levels? Given these questions we consider that using an external corpus for the vectorization can help the topic segmentation to be effective at different levels, by overcoming the decreased number of words at these levels and by allowing a change of pivot documents at any of these levels.

Our approach consists in doing the vectorization with using two corpora. The first corpus contains the data to segment and consist of 7 samples of Envoyé Spécial (2008-2009, 2 hours long each), a French TV report broadcast. The manner in which the pivot-documents are selected from this corpus is different than in [9]. The second corpus is represented by articles from a French journal and it is used for the vectorization (Le Monde, 2006). These corpora have different properties and they will be detailed in the following subsections. A reference segmentation of the TV reports is also available, in which reports were manually divided into themes and sub-themes. Another important aspect is defining the concept of theme and granularity at the level of theme. As mentioned in section 3, finding a definition for these concepts that is accepted unanimously is difficult since these are notions liable to subjectivity. Regarding the subject of interest for us, we consider the definition of topic and sub-topic as proposed in [15] . Therefore the topic corresponds to a news report and the sub-topics to different points of view on the subject approached in the news report.

The following subsections consist in providing details for the pre-processings required for the task and in providing an answer for the main question: How to select pivot-documents pertinents for the task?

5.2 Data pre-processing

Before providing details regarding the segmentation task, respectively how it was employed, the data under consideration, the TV reports and le Monde corpus need to be prepared. Since the lexical cohesion characteristic is considered for topic segmentation the word distribution needs to be analyzed. Several words appear frequently in the data without providing important information. It is the case of, for example, modal or auxiliary verbs (e.g. être, avoir, falloir, etc.). Therefore a filter on the existing words is applied and only the most informative ones are retained. The words are tagged with the TreeTagger[6] tool and only the substantives, verbs (other than modal and auxiliary verbs) and adjectives are kept. In addition, in order to favor word repetitions, words are lemmatised with the same tool.

5.3 Vectorization principle, pivot document selection

The principle of vectorization was previously described in section 4. For this reason only a brief recall of the main idea is provided and we will focus on how this technique was used in this work.

The principle is relatively simple, instead of directly comparing two documents using the similar words they share, they are compared based on their similarity with "external" documents: the more they resemble to the same documents, the more similar they are. The vectorization principle can be applied with TV reports as in [9]. Instead of dealing with documents, report segments are considered; therefore segments similar to the same m pivot documents are similar to each other.

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

As previously mentioned we select the pivot documents (PDs) for vectorization differently than in [9]. In this paper, the authors selected PDs from random splits of the data to segment. Creating PDs from the TV reports seems a reasonable decision since this assures that there is vocabulary shared between them and the reports. The problem of this approach happens when going at inferior levels of the hierarchy, where the available vocabulary is not sufficient. Since the authors did propose this technique for the linear segmentation, they were not therefore concerned with that problem. When applying their technique for obtaining the segmentation at an inferior level in the hierarchy, we observed that it does not provide as good results as for the first level. Therefore the first problems that arised concerned the selection of PDs. From where should they be built? (since using the same data is not sufficient for providing hierarchical segmentation) What size should they have? How much vocabulary should they share with the TV reports? Can we select PDs that have topics in common with those from the TV reports? Therefore we considered selecting PDs in two different ways. Our first choice is using an external corpus because we consider that it can overcome the lack of words at inferior levels in the hierarchy. This method is detailed in subsection 5.3.1. The second choice is closer to the one employed in [9], using the TV reports and is presented in subsection 5.3.2; the manner in which the PDs are created is different and the reason for using the reports is that the amount of shared vocabulary is high.

5.3.1 Pivot document selection using an external corpus

For this task a corpus which contains all articles from 2006 of the French newspaper "Le Monde" was used. The articles available in the newpaper cover a wide range of topics that are also treated in the reports. The generality of the topics in the articles varies and this ensures that a hierarchy of topics can be created. Of course this does not ensure that all existing topics in the reports are found in the newspaper and also that all sub-topics of a certain topic are available. Finding specific topics that are shared between the PDs and the TV reports is not the main focus, instead searching for similarities regarding the words that characterize the topics is. Therefore if a segment in the TV reports does not share vocabulary with the PDs, it means that it is not similar to another segment which shared vocabulary with the PDs. In addition we can take advantage of the generality of the topics at the first level in the TV reports since they can be found with a high probability in the newspaper's articles.

In order to provide a generic technique for the task of topic segmentation, a random selection on the articles for creating the PDs was considered. A question that arises is: Does selecting randomly a certain number of PDs or PDs that have certain characteristics help improve the segmentation task while maintaining the method generic? If all the available articles would have been considered and not just the m PDs randomly selected, the complexity of the computations would be significantly increased ($\#of\ articles \approx 34,000$).

Since there are high chances to have articles that share the same topic, two manners to deal with this are considered. One consists in performing a clustering

to concatenate them, therefore a certain topic will be characterized by more words, giving a higher chance to match the words used in the reports for the same topic. The steps for the selection of the m PDs and the clustering are provided below:

1: select m random PDs
2: part-of-speech tag the PDs
3: keep lemmas of substantives,verbs and adjectives for each PD
4: create m_c clusters, where $m\_c = m$ initially
    4.1: initialize clusters with the m_c PDs
    4.2: for each pair (PD_i, PD_j) with $i \neq j$
        $similarity = cosine(PD_i, PD_j)$
            if $similarity \geq 0.5$ then
                $PD_{new} = concatenate(PD_i, PD_j)$
                $m_c - -$
    goto 4.1 until convergence
5: $if = m_c < m$
    goto 1.,$m = m - m_c$

The second approach consists in selecting the PDs that are very different from the point of view of their topic. Therefore the remaining PDs will be different and their size will not be modified. This gives the possibility of having PDs that are more specific in terms of topic. For example if we have two articles that speak about economy, but actually one speaks about the economical crises and the other one about the economical impact of a war in a country, if we consider the first technique there is a high probability they will be concatenated if they use the same vocabulary to describe the news. When using the second approach none of the two articles will be taken into consideration. This approach could be advantageous for lower levels in the hierarchy where the topics are very specific. The drawback is that topics that could help the segmentation may be eliminated. The modifications necessary for the previous algorithm to keep only the PDs that have different topics regard step 4.2:

4.2: for each pair (PD_i, PD_j) with $i \neq j$
    $similarity = cosine(PD_i, PD_j)$
    if $similarity \geq 0.5$ then
        $eliminate PD_i, PD_j$
        $m_c = m_c - 2$

Another solution that could be employed to help the segmentation at lower levels could be a hierarchical clustering of the articles.

Before going forward and perform the topic segmentation after the PDs selection, an analysis on the data was performed to check if the TV reports and the newspaper articles selected share vocabulary. This was done because the type of the vocabulary in the two corpora differs: the one employed in the French newspaper is closer to literary French while in the TV reports a more informal language is used. The PDs for this tasks were randomly selected using the first technique (the second is more suitable for lower levels in the hierarchy, as mentioned above). The results

of the statistics are presented in Table 1. It can be observed that TV reports share an important proportion of their vocabulary with PDs (for example ES_09_26_2008 shares 62% of its vocabulary with the vocabulary created from 100 PDs). The last row in the table corresponds to the number of terms shared between the PDs and ALL_ES, where ALL_ES represents a concatenation of all the TV reports transcripts (i.e., the reports were pieced together by connecting the end of each report with the start of another). The values in this case are lower since there is a high amount of vocabulary shared between the TV reports. As increasing the number of PDs selected the amount of terms shared increases. This analysis was done considering only the lemmas[7]; their frequency was not taken into consideration. We consider that the

TV reports	100 PDs	200 PDs	300 PDs
ES_09_26_2008	62%	72.4%	76.5%
ES_10_10_2008	60%	70.3%	73.3%
ES_11_06_2008	58%	67.9%	71%
ES_11_20_2008	60%	69%	71%
ES_11_29_2008	59%	68.8%	71.9%
ES_01_15_2009	60%	70%	73.6%
ES_01_22_2009	61%	70.4%	74.7%
ALL_ES	30%	38.6%	41.44%

Table 1: Shared vocabulary between TV reports and 3 corpora composed by 100, 200 or 300 PDs

the two corpora share enough vocabulary to continue investigating the contribution of vectorization when using an external corpus.

5.3.2 Pivot document selection using the same corpus

In [9], the authors selected the PDs from the TV reports. They used a window of three different sizes to split the data into segments that have either very small, medium or big dimension with respect to the total number of breath groups. Their approach performed well for segmenting the first level of the hierarchy. In our case, the data is split using windows that have 7 different sizes, in the range of 30 to 300. We are interested in analyzing the impact of using the TV reports corpus for the vectorization and, for this reason, we perform different combinations of the segments obtained after splitting:

1. PDs from all TV reports
2. PDs from all the TV reports except the one on which the segmentation is performed
3. PDs from one TV report

[7]Lemma - a word considered as its citation form together with all the inflected forms. http://www.thefreedictionary.com/lemmas

4. PDs from 2 TV reports

The PDs were created after performing a random selection on the segments obtained after splitting. As opposed to what was done in [9], we do not use just the data that we want to segment to build the PDs, we use other reports as well. This way the importance of sharing vocabulary between the PDs and the data to segment can be observed. The selection is done by simply choosing m random PDs. Similar to the statistics done for the case of PDs selected from newspaper articles, a statistics was done for the case of selecting PDs from the TV reports. In the table below the results obtained are presented. For each TV report, the results are lower than those obtained in the previous case, while for the concatenation of TV reports (ALL_ES) the amount of vocabulary shared is higher. The results are justified by the fact that the number of word occurrences were not taken into consideration and the fact that the TV reports share a lot of their vocabulary, therefore choosing a higher number of PDs does not lead to a high increase in the number of vocabulary shared. We consider that the amount of shared vocabulary is sufficient to continue

TV reports	PDES_600	PDES_100
ES_09_26_2008	47.16%	44.26%
ES_10_10_2008	45.82%	39.99%
ES_11_06_2008	46.57%	42.83%
ES_11_20_2008	45.91%	38.30%
ES_11_29_2008	48.10%	39.62%
ES_01_15_2009	45.91%	36.83%
ES_01_22_2009	46.416%	36.09%
ALL_ES	59.19%	41.7%

Table 2: Shared vocabulary between TV reports and PDs constructed from TV reports

investigating this path of creating PDs also.

Now that the manner in which we select the PDs was detailed, we will describe the segmentation methods, with or without the use of PDs.

5.4 Topic segmentation using TextTiling

In this subsection the modified version of the TextTiling algorithm used for topic segmentation is presented. The basic principle of the method, as described in section 4, is maintained. Modifications regard the way the similarity is computed. In what follows the modifications and the motivation behind them are detailed.

In [9], the authors have used a technique similar to TextTiling for the vectorization part, while for topic segmentation they have used a new technique inspired from the image segmentation field. We have considered using TextTiling for topic segmentation but with several modifications that were employed in [9].

The report to segment is parsed using a sliding window and a gradient is computed between the utterances contained in the left side of the window and the ones

in the right side. For this, the cosinus measure was used and the utterances are weighted by $\sqrt{TF}$ (i.e., the square root of the number of occurrences of the word in the breath group). This means that each side of the window is represented by a vector in which each dimension represents a word. Topic boundaries are predicted in the points of local minima, meaning that the gradient is small. Inspired from some image gradient computation methods, more importance is given to the words that are closer to the candidate frontier, as was done in [9]. This can be obtained through a convolution with a kernel (e.g., Gaussian kernel). The way this is employed depends on the data representation for the similarity computation. In our case, when computing the $\sqrt{TF}$, a word occurrence in the closest breath group to the candidate edge has the weight equal to 1 and less if it is further. For this a linear penalty is applied. The motivation behind this is that favoring words closer to a potential frontier make the gradient value at that point either smaller or greater. Therefore a better differentiation between segments is provided.

To perform the topic segmentation, a prediction is done on the reports, using the gradient values computed as previously described. First a smoothing is applied on these values following with an examination of the peaks and valleys from the plot. Afterwards frontiers are predicted in the lowermost portions of the valleys. In Figure 6, the predicted frontiers obtained after computing the gradient for one TV report are presented. In addition the reference frontiers are positioned.

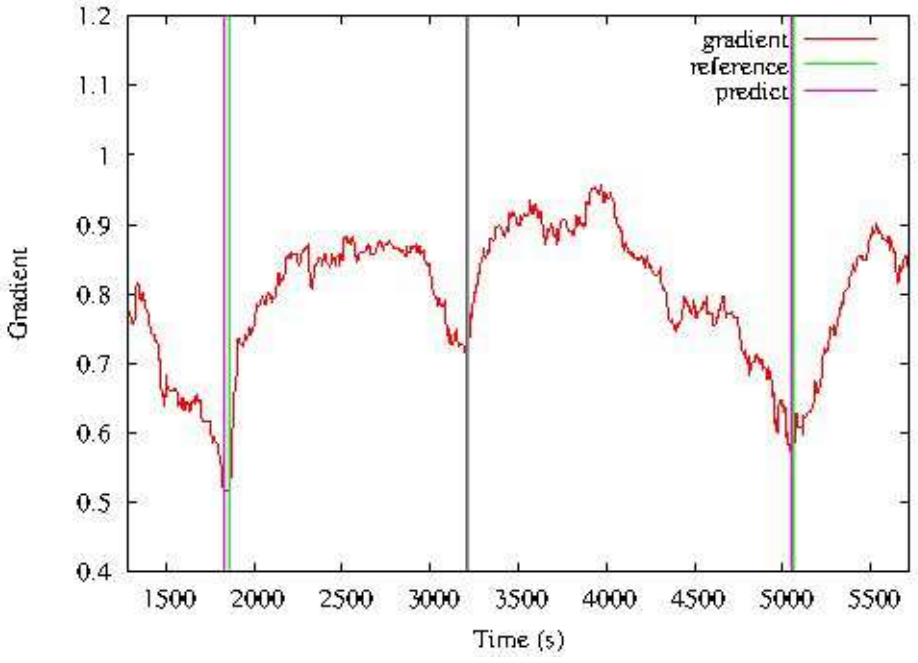


Figure 6: Illustration of segmentation process (gradient vs. time of utterances)

In order to obtain a hierarchical topic segmentation this modified version of TextTiling is applied recursively.

5.5 TextTiling with vectorization

In this subsection the integration of the vectorization technique with the topic segmentation method is detailed. For this task the modifications brought to TextTiling concern the gradient computation. We have considered two different ways of modifying the lexical cohesion computation:

1. Indirect method:
Instead of directly computing the cosine similarity between the report's blocks of breath groups (as done with TextTiling), a similarity between the blocks and the PDs is computed. This leads to a new vectorial representation of the blocks. The value of the lexical cohesion associated with each potential frontier i is defined by:
$\nabla(i) = cosine(Vect(Cprev(i-1), P, \sqrt{(TF)}/cosine), Vect(Cnext(i), P, \sqrt{(TF)}/cosine))$,
where $Vect(Cprev(i-1), P, \sqrt{(TF)}/cosine)$ is the representative vector for the breath groups before the potential frontier i. It contains the similarity values between the block of breath groups $Cprev(i-1)$ and the pivot-documents P. The similarity values are obtained using the cosine and the wieghts of the terms in the breath groups and PDs are computed using $\sqrt{(TF)}$. Each block is denoted by $Cprev(i)$ which denotes the result of the convolution operator (c.f. subsection 5.4) applied on the block.

2. Combine indirect with direct method:
The idea is to combine the previously described method with the one presented when describing the functionality of TextTiling. Therefore the vectorial representation of each segment in the TV report will contain the similarity values computed using the PDs and also the vectorial representation of the segment itself.
$Vect(Cprev(i-1), P, \sqrt{(TF)}/cosine; Cprev(i-1), \sqrt{(TF)})$ contains now also the vectorial representation of the segments. For a clearer view on this technique we provide the figure below, where Swi represents the score associated to each word in the block of breath groups:

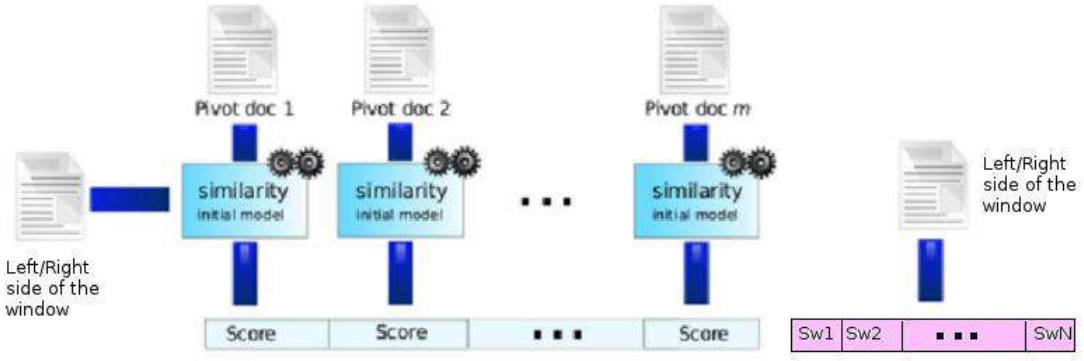


Figure 7: Combination of the direct and the indirect method

By combining the two methods, the similarity score between the blocks of breath groups contained in the reports can be improved. The direct method takes into consideration only the content of the blocks and, at higher levels in the hierarchy, it can provide a good distinction between blocks that belong to different topics. Therefore, this technique can help the indirect method and can correct some of the similarity scores computed. However at lower levels the direct method does not have a big impact due to the decreased number of words. When combining the direct with the indirect method, more importance can be given to one of the two methods. For example if the similarity score between 2 blocks, computed using the direct method, is higher than 0.5, those blocks can be considered similar even if the similarity score obtained using the indirect method is very low. Therefore, blocks that share more vocabulary are favored. We have performed such experiments, but still the best improvements for the gradient computation were obtained when the two methods had the same importance in its computation.

5.6 Validation and evaluation

As mentioned in the subsections above, the model proposed in this work consists in integrating a vectorization technique with a modified version of the TextTiling algorithm. To obtain a hierarchical topic segmentation the algorithm is applied recursively. Before running experiments on the model proposed, an oracle was considered for its validation. The motivation for doing this is to prove that the model could provide an accurate topic segmentation. In addition, it justifies the design choices considered for selecting the PDs. In the following part of the report the validation and the experiments performed with the corresponding results are presented, but first of all several characteristics of the data used are given.

5.6.1 Experimental data

The experiments are performed on a French TV broadcast corpus, which contains 7 Envoyé Spécial reports, from 2008 and 2009. Each report is 2 hours long. This corpus was preferred as opposed to other TV reports, like Sept à Huit, because it has a more important hierarchical structure. The reference segmentations were obtained by manually dividing the TV reports into themes and sub-themes. These segmentations have 3 levels of hierarchy: the first contains 26 frontiers, the second one 246 and the third 722. The segments of the first level in the hierarchy can be characterized as long and relatively stable in size, the ones at lower levels are short and have few word repetitions. In the table below, a comparison between the first two levels is provided.

Another characteristic of these TV reports is favoring outdoor investigations, which lead to a high word error rate (WER) in the transcripts obtained from the ASR system.

	number of frontiers	average duration of segments	average number of repeated words per segments	number of words in each segment
first level in the hierarchy	26	32 min max:55 min, min:22 min	140	1639
second level in the hierarchy	246	3.4 min max:20 min, min:7 sec	15	173

Table 3: Comparison of different levels of granularity from Envoyé Spécial corpus [15]

5.6.2 Evaluation

For evaluating the hierarchical topic segmentation a basic method is considered and consists in producing the ground truth for each level in the hierarchy and compare the hypothesized segmentations with the reference ones for each level. The hypothesized segmentation was obtained after plotting and smoothing the gradient obtained for all the potential frontiers. This step involved using a prediction window for searching local minimum in the plot obtained.

Performing an evaluation at each level individually brings the drawback of not being able to take into account, as global, an error produced at a higher level in the hierarchy, which may influence the segmentation at inferior levels. There exist only one evaluation metric that was defined for hierarchical topic segmentation, which was presented in section 4. The drawback of that technique is the constraint imposed on having an equal number of hypothesized and reference frontiers. It is considered a problem since it cannot evaluate the performance of a segmentation algorithm, over and under-segmentations being important characteristics for the quality of the produced segmentation [15]. The metric chosen for evaluating the segmentation is inspired from those used in IR, respectively recall and precision. These measures are defined as follows:

$$Recall = \frac{|H \cap R|}{|R|}$$

$$Precision = \frac{|H \cap R|}{|H|}$$

where $|H|$ (resp. $|R|$) represent the number of frontiers contained in the hypothesized (resp. reference) segmentation. The recall corresponds to the proportion of reference frontiers that were detected by the method under evaluation and the precision represents the ratio of frontiers produced belonging to the reference segmentation.

These two measures are not sensitive to the variations of the segments dimensions, as *Pk* measure (described in section 3) and they do not favor segmentations with fewer number of frontiers as *WindowDiff* (also in section 3). We consider the adjustment proposed in [15], by aligning the two segmentations (the reference one

with the hypothesized one) and allowing the hypothesized frontiers to be shifted by T seconds from the references ones. In practice T was set to values from 2 seconds up to 20 seconds. The existence of elevated T can be explained by the fact that dealing with TV reports and using the spoken words for segmentation can lead to a shift between the moment the words were pronounced and the end of a report.

5.6.3 Oracle

The vectorization technique consists in creating PDs that can help the segmentation by alleviating the problems caused by the lack of common words between blocks of breath groups. As mentioned above the PDs considered for the segmentation task were build in two different ways: from an external corpus, represented by newspaper articles, and from the TV reports corpus. In order to justify the idea that randomly selecting m PDs from the available ones can improve the segmentation process, an oracle was considered. The basic principle is that knowing the thematic segments before segmentation (i.e., knowing the thematic frontiers), m PDs can be found, good enough to provide an accurate segmentation. This means that there actually exist m PDs that can help distinguish between the topics in the reports.

First of all, a statistic was done to observe the amount of vocabulary shared between the PDs and the TV reports (Table 1 and 2). Each PD has different topic with respect to the other PDs and choosing the PDs that share more vocabulary with the TV reports can lead to a better differentiation between topics. In addition longer articles provide a better coverage of the topic and have higher probability of matching a certain block from the reports. The basic idea of the algorithm consists in finding the m PDs that cover most of the vocabulary of each topic (knowing the topics in the report) while eliminating the PDs that are common to more segments in the report. A threshold t is defined to select at least t PDs for each topic. Below a more detailed description of the algorithm is provided (in the following algorithm description a TV report is denoted by ES)

1: $select\ m\ random\ PDs$
2: $part-of-speech\ tag\ the\ PDs$
3: $foreach\ pair\ (PD_i, TopicES_j)$
3: $statistic-analysis(PD_i, TopicES_j)$
  $if\ \#words-in-common(PD_i, TopicES_j)\ \times 100\ /\ \#words-in-PD_i\ > 30\%$
    $if\ !found(PD_i, queue)$
      $enqueue PD_i$
      $stamp(PD_i) = 0$
      $countTopicES_j++$
    $else$
      $if\ stamp(PD_i) == 0$
        $stamp(PD_i) = 1$
        $countTopicES_j--$
4: $foreach\ PD\ in\ queue$
  $if\ stamp(PD_i) == 1$

$dequeue(PD_i)$
5:$foreach\ countTopicES$
$if\ countTopicES_j\ < t\ goto\ 1.$
6: $if\ size-of-queue\ <\ m\ goto\ 1.$

The results obtained for the selection of the "correct" PDs for segmenting the Envoyé Spécial report broadcasted in 15/09/2009 are presented in Table 4. For the segmentation, a window of size 500 was used. In the table, the results provided were obtained by first performing just the modified version of TextTiling (justTT), then by integrating the vectorization using the PDs found with the oracle (justPD) and the last results are obtained by combining the previous 2 methods. The values corresponding to R(10) and P(10) are the recall and precision with a threshold of 10 seconds, (meaning that a frontier is considered correct if it is at most 10 seconds away from the considered hypothesized frontier), R(15) and P(15) are computed with a threshold of 15 seconds and R(20) and P(20) with a threshold of 20 seconds. Recall and precision are computed as described in subsection 5.6.2. The number of reference segments is denoted by *nref* and the hypothesized ones by *nhyp*

	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
justTT	4	4	33.3	33.3	66.7	66.7	66.7	66.7
justPD	4	4	100	100	100	100	100	100
combined	4	4	33.3	33.3	100	100	100	100

Table 4: Results at the first level in the hierarchy, for topic segmentation on the Envoyé Spécial report from 15/09/2009, using the oracle to select PDs

As it can be observed the use of vectorization can bring more accurate results than TextTiling. The decrease in recall and precision obtained when combining the two approaches is justified by the fact that the direct method influences more the values of the potential frontiers. This happens because the blocks considered when applying TextTiling are large (500 breath groups) and the number of PDs used is small (36). However the results presented above correspond to the first level in the hierarchy. When going further in the hierarchy the previously used PDs need to be replaced, in order to be able to distinguish topics at lower levels. PDs that worked for a superior level will not necessarily work for the inferior one, especially if they are chosen using an oracle (choosing them specifically for topics at superior level).

For the same report as above we have performed the segmentation for the second level in the hierarchy, on one of the topics from the first level. The size of the window used for TextTiling was smaller (30 and 100 breath groups). The results obtained are presented in Table 5.

The threshold t was set to 10; therefore at least 10 PDs should be found to characterize each topic. The reason for which the oracle does not perform well for finding articles characteristic for the second level is because the method randomly

	nref	nhyp	R(2)	P(2)	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
w=30										
justTT	9	10	0	0	12.5	11	25	22.2	25	22.2
justPD	9	9	12.5	12.5	12.5	12.5	25	25	25	25
combined	9	8	0	0	0	0	0	0	12.5	14.3
w=100										
justTT	9	8	12.5	14.3	25	28.6	25	28.6	25	28.6
justPD	9	6	12.5	20	25	40	25	40	25	40
combined	9	7	0	0	25	33.3	25	33.3	25	33.3

Table 5: Results at the second level in the hierarchy, for topic segmentation on the Envoyé Spécial report from 15/09/2009, using the oracle to select the PDs, and 2 different window sizes

selects the PDs and for lower levels in the hierarchy the algorithm described is time-consuming, it takes more time to find PDs that are specific to the sub-topics. However the results using just the PDs are better than TextTiling alone, meaning that if some PDs more specific to the sub-topics can be selected, the segmentation can be improved. Therefore using the PDs for the segmentation is effective.

5.7 Results

In this subsection the results obtained from the experiments done are presented. The tests were performed on each TV report, but also on the TV reports concatenated as one, for a global evaluation. The experiments were done in the following order: using the modified version of TextTiling (subsection 5.7.1), then integrating the vectorization technique (subsection 5.7.2) and in the end combining the direct with the indirect method (subsection 5.7.3). When performing the tests, the parameters used were varied on a large range, so that conclusions can be drawn with respect to the appropriate values for them.

5.7.1 Topic segmentation with TextTiling

The details regarding how the data is prepared for the segmentation are provided in the data pre-pocessing subsection 5.2. As detailed there, only several words are kept and they are lemmatised. However at the beginning, tests were done also with stems[8] to see which approach gives better results. In Table 6 the results obtained when using either lemmas or stems are presented. As it can be observed, the use of lemmas provides better results in terms of precision (P) and recall (R). This happens because using lemmas a better differentiation between blocks can be done. Using stems we would have more terms repetitions but we will end up connecting more

[8]Stem - the main part of an noninflected word to which affixes may be added to form inflections of the word. http://en.wiktionary.org/wiki/stem

LEMMA	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
w=300	26	24	32	34.8	56	60.9	60	65.2
w=200	26	23	24	27.3	52	59.1	56	63.6
w=100	26	25	16	16.7	52	33.3	36	37.5
STEM	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
w=300	26	23	32	36.4	52	59.1	56	63.6
w=200	26	23	24	27.3	48	54.5	52	59.1
w=100	26	23	12	13.6	24	27.3	28	31.8

Table 6: Results at the first level in the hierarchy, for topic segmentation on the Envoyé Spécial reports concatenated using lemmas and stems

blocks that have different topics. For this reason the following experiments were done using only lemmas.

The results previously presented were on all the TV reports concatenated. The results obtained on each of the 7 TV reports independently are given in Table 7. The window size considered is 300. For several TV reports the results obtained are

	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
ES_01_15_2009	4	4	33.3	33.3	100	100	100	100
ES_01_22_2009	4	4	33.3	33.3	66.7	66.7	66.7	66.7
ES_09_26_2008	3	3	100	100	100	100	100	100
ES_10_10_2008	4	4	33.3	33.3	66.7	66.7	66.7	66.7
ES_11_06_2008	4	5	33.3	25	33.3	25	33.3	25
ES_11_20_2008	3	4	50	33.3	100	66.7	100	66.7
ES_11_29_2008	4	4	0	0	33.3	33.3	66.7	66.7

Table 7: Results at the first level in the hierarchy, for topic segmentation on each Envoyé Spécial with a window of size 300

very good and the reason is that the topics are very different from one another. For example in ES_09_26_2008 the three topics approached are: "Ecole, la violence entre les lignes", "Avoir 20 à Lhasa" and "Magasins, vols à tous les étages".

For performing the hierarchical segmentation, TextTiling is applied recursively. The size of the window, as going deeper in the hierarchy, needs to decrease in order to capture the topic changes that are more subtle. In Table 8, the results obtained for the first 2 levels in the hierarchy, when using a window of size 100, are presented. As it can be observed the results for some of the TV reports at the first level are worse than those obtained with a window of bigger size, it is the case for example of the first TV report from 01/15/2009 (ES1). However for the report from 11/20/2008 (ES6) the results at the first level are better than those obtained with a bigger window. At the second level the results vary from report to report, still the P and R are small.

	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
ES1_level1	4	4	33.3	33.3	66.7	66.7	66.7	66.7
ES1_level2	38	20	5.4	10.5	18.9	36.8	24.3	47.4
ES2_level1	4	6	0	0	66.7	40	66.7	40
ES2_level2	45	18	6.8	17.6	11.4	29.4	13.6	35.3
ES3_level1	3	4	50	33.3	50	33.3	50	33.3
ES3_level2	42	22	14.6	28.6	17.1	33.3	19.5	38.1
ES4_level1	4	4	33.3	33.3	33.3	33.3	33.3	33.3
ES4_level2	36	27	20	26.9	34.3	46.2	42.9	57.7
ES5_level1	4	6	33.3	20	33.3	20	33.3	20
ES5_level2	31	14	10	23.1	10	23.1	20	46.2
ES6_level1	3	3	50	50	100	100	100	100
ES6_level2	20	20	10.5	10.5	31.6	31.6	36.8	36.8
ES7_level1	4	4	0	0	0	0	33.3	33.3
ES7_level2	34	23	18.2	27.3	27.3	40.9	36.4	54.5

Table 8: Results at the first and second level in the hierarchy, for topic segmentation on each Envoyé Spécial report with a window of 100

5.7.2 TextTiling with vectorization

Before performing the segmentation, the m PDs need to be generated. For this task the m was varied from 30 to 400. Since the PDs are randomly selected, the same m was consider several times.

As it can be observed in Table 9 the results obtained using PDs created from

	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
PD75	26	18	8	11.8	20	29.4	24	35.3
PD100	26	24	8	8.7	24	26.1	24	26.1
PD150	26	15	4	7.1	4	7.1	4	7.1
PD150	26	33	4	3.1	12	9.4	20	15.6
PD300	26	29	12	10.7	32	28.6	36	32.1
PD400	26	24	8	8.7	16	17.4	20	21.7
PDES	26	25	12	12.5	24	25	28	29.2

Table 9: Results at the first level in the hierarchy, for topic segmentation with vectorization on the Envoyé Spécial reports concatenated

the TV reports do not differ a lot with respect to those obtained when using PDs generated from the newspaper articles (PDES). Another interesting observation is related to the results obtained when the same number of PDs was generated several times. In the table there are 2 different results when using 2 different generations of 150 PDs. This means that randomly selecting m PDs can sometimes give poor results while reelecting m PDs better results can be obtained. In order to take care of this, a large number of repeated selections of the same number of PDs can

be made and an average should be computed over the results obtained with each selection of PDs.

The results in Table 9 were computed considering all the reports concatenated. In what follows we present the results in detail for each report (Table 10). The size of the window is 300. The values obtained demonstrate that at least for the first level

	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
PD300								
ES1	4	5	66.7	50	66.7	50	66.7	50
ES2	4	6	66.7	40	66.7	40	66.7	40
ES3	3	2	50	100	50	100	50	100
ES4	4	5	0	0	33.3	25	66.7	50
ES5	4	2	0	0	0	0	0	0
ES6	3	4	0	0	100	66.7	100	66.7
ES7	4	4	0	0	66.7	66.7	66.7	66.7
PDES								
ES1	4	3	33.3	50	33.3	50	33.3	50
ES2	4	5	0	0	33.3	25	33.3	25
ES3	3	4	50	33	50	33.3	50	33.3
ES4	4	4	0	0	33.3	33.3	66.7	66.7
ES5	4	6	33.3	20	33.3	20	33.3	20
ES6	3	5	0	0	50	25	50	25
ES7	4	5	33.3	25	66.7	50	66.7	50

Table 10: Results at the first level in the hierarchy, for topic segmentation using vectorization on each Envoyé Spécial using 300 PDs formed from newspaper articles and PDS formed from the TV reports

in the hierarchy the choice of using PDs created from the TV reports does not bring additional advantage. We have mentioned in subsection 5.2 that for selecting PDs from the TV reports several combinations are considered. In Table 11, we detail the results obtained for the second level using the various combination, and a window of size 70. The splits on the TV reports to create the PDs were in the range of [30,500]. The results obtained for the second level in the hierarchy using the PDs created from the TV reports are better that those obtained using the newspaper articles, which can be seen in Table 12. It can also be observed that using the same report to build the PDs as for segmentation dose not give better results than using the entire corpus of TV reports to build the PDs. However, when we perform over-segmentation at the second level using PDs from the newspaper articles, the results are better than those obtained by over-segmentation on PDs from TV reports. The reason is that using PDs from the articles give more accurate results in terms of the time difference between the frontiers predicted and the hypothesized ones.

ES_15_01_2009	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
PDES	38	32	18.9	22.6	18.9	22.6	24.3	29
PDES_exceptES1	38	38	16.2	16.2	24.3	24.3	29.7	29.7
PDES1	38	35	13.5	14.7	16.2	17.6	21.6	23.5
ES_22_01_2009	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
PDES	45	30	18.2	27.6	20.5	31	20.5	31
PDES2_ES3	45	41	18.2	20	27.3	30	27.3	30
PDES2	45	41	11.4	12.5	20.5	22.5	25	27.5

Table 11: Results at the second level in the hierarchy, for topic segmentation on 2 Envoyé Spécial reports using different combinations of PDs from TV reports

ES_15_01_2009	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
PD75	38	38	8.1	8.1	18.9	18.9	24.3	24.3
PD100	38	36	8.1	8.6	13.5	14.3	16.2	17.1
PD150	38	38	10.8	10.8	18.9	18.9	21.6	21.6
PD300	38	39	13.5	13.2	21.6	21.1	27	26.3
PD400	38	33	8.1	9.4	13.5	15.6	18.9	21.9
ES_22_01_2009	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
PD75	45	41	11.4	12.5	15.9	17.5	27.3	30
PD100	45	43	6.8	7.1	9.1	9.5	15.9	16.7
PD150	45	40	6.8	7.7	9.1	10.3	11.4	12.8
PD300	45	45	13.6	13.6	18.2	18.2	22.7	22.7
PD400	45	43	11.4	11.9	13.6	14.3	18.2	19

Table 12: Results at the second level in the hierarchy, for topic segmentation on 2 Envoyé Spécial reports using different combinations of PDs from TV reports and a window of size 70

5.7.3 Direct with indirect method

The first results obtained for the combination of the direct with the indirect method for computing the similarity scores did not ameliorate nor made worse the results for the first level in the hierarchy. For this reason we have performed a statistic analysis on the PDs and the TV reports. The TV reports were split with windows of varying sizes into blocks of breath groups. For each block of a certain size the median, average and variation coefficient were calculated considering the number of words contained. For the PDs, the size in terms of number of words contained in each PD was computed and the same statistic was done. In Table 13, only the values obtained using a window of 100 for each TV report are given.

Envoyé Spécial	median	mean	CV
ES_01_15_2009 w=100	128	128.378	2.22
ES_01_22_2009 w=100	120	124.6	3.53
ES_09_26_2008 w=100	113	113.004	4.2
ES_10_10_2008 w=100	125	126.31	3.06
ES_11_06_2008 w=100	128	130.51	5.25
ES_11_20_2008 w=100	121	123.012	2.67
ES_11_29_2008 w=100	128	127.87	3.9
ES_ALL w=300	314	319.47	5.1
PDs			
PDES	92	194.87	174.09
PD75	87	115.9	80.59
PD150	126	122.72	63.14
PD150	121	134.75	80.21
PD500	114	121.46	65.1
PD100	106	125.21	86.45
PD300	121	125.56	71.78

Table 13: Statistics on the vocabulary contained in TV reports and PDs

The coefficient of variation (CV) has high values for the PDs which means that the size of PDs is highly varying. The TV reports instead have low CV, meaning that most block have the size of the median (or the average). In order to find PDs that share more vocabulary with the reports, the size of the PDs needs to be bigger. Therefore to perform the segmentation at the first level, we considered selecting the PDs that have the number of words higher than the average number of words in the PDs. The results obtained after this are presented in the table below.

The precision and recall obtained when combining the two methods justify the

	nref	nhyp	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
justTT	26	24	32	34.8	56	60.9	60	65.2
justPD	26	19	8	11.1	16	22.2	20	27.8
combined	26	22	36	42.9	64	76.2	68	81

Table 14: Results at the first level in the hierarchy, for topic segmentation on the Envoyé Spécial reports, concatenated, with PDs that have the # words>average # words on each PDs

fact that using vectorization can improve the segmentation. The results obtained only with vectorization are not as good as those obtained with TextTiling alone, but actually the frontiers that are correctly predicted are closer to the hypothesized ones. For the second level the constraint put on PDs was to share more than 50% of the words, since at lower levels in the hierarchy the segments are shorter (Table 2.). Also the size of the window used for TextTiling was smaller (30 breath groups). We have also performed the segmentation at the third level, using the same PDs as

for second level and employing the vectorization.

	nref	nhyp	R(2)	P(2)	R(10)	P(10)	R(15)	P(15)	R(20)	P(20)
level 2										
justTT	11	11	10.0	10.0	20.0	20.0	30.0	30.0	40.0	40.0
justPD	11	12	40.0	36.4	50.0	45.5	50.0	45.5	60.0	54.4
combined	11	12	40.0	36.4	50.0	45.5	50.0	45.5	60.0	54.4
level 3										
justTT	24	21	4.3	5	43.5	50	43.5	50	47.8	55
justPD	24	26	8.7	8	30.4	28	52.2	48	69.6	64
combined	24	22	17.4	19	52.2	57.1	56.5	61.9	69.6	76.2

Table 15: Results at the second and third level in the hierarchy, for topic segmentation on one topic from the first level of the Envoyé Spécial report from 15/09/2009

As it can be observed from the values in Table 15, the vectorization technique improves TextTiling and when the direct and indirect methods for computing the similarities are combined they do not have a negative impact on the final results. Moreover for the third level the segmentation is improved when the two methods are combined.

5.8 Discussion

The solution provided, based on integrating vectorization with TextTiling shows an improvement of the segmentation quality when compared to TextTiling. In what follows we will present the main advantages and disadvantages of the method proposed, together with some possible improvements.

The advantage of using a vectorization technique is that it can better discriminate between topics at lower levels in the hierarchy, if the PDs are well chosen. Still we want our method to remain generic, therefore the constraint imposed on PDs are only related to the amount of the vocabulary shared between the PDs and the TV reports. When the PDs are formed from newspaper articles and without any constraint, the results obtained are worse than when the TV reports are used for creating the PDs. The reason is that the PDs formed using the reports share more vocabulary with the data to segment. In the case of PDs created using the articles and with constraints, the results are better even without combining the direct with the indirect method. Another manner to construct PDs could be employed by combining the 2 previously considered techniques. Another advantage is that the vectorization technique could be used to improve other segmentation algorithms that perform better than TextTiling. Of course, more research should be done on how to perform the integration, since for other methods it is not as straightforward as for TextTiling.

Finally, concerning the disadvantages of the solution, it is difficult to decide the appropriate number of PDs to use, since good results were obtained both using 300 and 30 PDs. For this, more experiments should be done and, since we want a generic technique, the number may be put in relation with the size of the data to segment, or other criteria. An idea would be to combine in the end the similarity scores obtained using a wide range of PDs. Using only the PDs created from the TV reports that we want to segment, seems to create a disadvantage on the vectorization technique at lower levels in the hierarchy. Indeed at a lower level, if we divide more the data to create the PDs, those PDs will have the same problem of reduced number of words.

6 Conclusions

Our objective was to propose a method for hierarchical topic segmentation of TV reports. In order to have a generic technique we have exploited the transcripts of the speech pronounced in the TV reports. The solution proposed consists in recursively applying a method for linear segmentation, which is based on the vectorization technique, to obtain a hierarchical segmentation. The method employed was not exploited until now for hierarchical segmentation and we have showed that it can respond to this task if the positions of the words in the reports are taken into consideration and if the pivot-documents selected meet several constraints. Therefore the linear segmentation algorithm chosen, TextTiling, was modified to give more weights to words closer to potential frontiers and as for the selection of PDs, the best results were obtained when the vocabulary contained was higher than the average and when PDs shared at least a certain amount of vocabulary with the TV programs to segment.

The main reason for using vectorization is one of the properties it has regarding the possibility of finding similar segments that do not share much vocabulary. Using TextTiling a direct computation of the similarities between consecutive blocks is done, while using vectorization the similarity values are computed indirectly through the use of pivot-documents. Combining the two methods gave the best results in terms of precision and recall. We applied the method on a corpus containing 7 samples of Envoyé Spécial which presents a hierarchical structure of 3 levels and this lead to a precision of 81% and a recall of 68% for the first level. For the segmentation at lower levels in the hierarchy the method can be improved, by combining the two methods for selecting PDs or by imposing more constraints regarding the vocabulary contained in the PDs, while keeping the method generic. At the second level in the hierarchy the best results had a recall of 60% and a precision of 54.4%. Still this values are better than those obtained using the state of the art for hierarchical segmentation. We have also obtained encouraging results for the third level in the hierarchy (recall of 61.9% and precision of 69.6%) when the direct and indirect methods were combined. In [15], the segmentation at the third level in the hierarchy was not done because the technique employed could not overcome the lack of words at this level.

The development of hierarchical topic segmentation techniques is still at early

stages, leaving an open path for finding new solutions. The work presented in this report shows that vectorization can improve the segmentation and due to the powerful property of this technique, going deeper in the hierarchy is not a problem as for other studies done on this topic. The difficult part is selecting the PDs and we have showed in this report several ways of doing this, that provide good results for all the levels in the hierarchy. Of course the vectorization technique can be further exploited since various improvements (c.f. section 5.8) can be brought to it.

As future work we propose the integration of the vectorization technique with another linear topic segmentation algorithm [45], which has the advantage of not requiring a priori the number of expected segments. Another interesting idea consists in using [45] for the first level in the hierarchy and, in exploiting the segments identified at the first level, to improve the selection of pivot-documents for the following ones. For this, a hierarchical clustering on the pivot-documents could be employed, which may be used also as a standalone solution for hierarchical segmentation. Another alternative could be to make use of the existing techniques from other domains. We consider that a hierarchical image segmentation technique could be employed.

Another important problem regarding topic segmentation is its evaluation. For linear segmentation, several methods exist, while for the hierarchical segmentation there is only one study so far that performs an evaluation of the segmentation obtained in its hierarchical form. However this solution is considered not able to evaluate the true performance of a segmentation algorithm. Therefore we are also interested in finding solutions regarding this problem. As mentioned in [15], the method for evaluation could be inspired from the one developed for matching XML documents. The principle of such an approach consists in computing the minimum editing distance needed to transform the structure of a XML document D in D'.

References

[1] Doug Beeferman, Adam Berger, and John Lafferty. Text segmentation using exponential models. *In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 35-46, 1997.

[2] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*. MIT Press,ISBN 0-262-20152-6, 2004.

[3] Jean Bourgain. On Lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52(1), 1985.

[4] Gillian Brown and George Yule. Discourse analysis. *Cambridge University Press*, 1983.

[5] Lucien Carroll. Evaluating hierarchical discourse segmentation. *In Proceedings of the 11th International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 993-1001, 2010.

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

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

[8] Vincent Claveau. *Acquisition automatique de lexiques sémantiques pour la recherche d'information*. PhD thesis, École doctorale: MATISSE, University of Rennes 1, 2003.

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

[10] Vincent Claveau, Romain Tavenard, and Laurent Amsaleg. Vectorisation des processus dappariement document-requête. *In 7e conférence en recherche d'informations et applications, CORIA'10*, Sousse, Tunisie, 2010.

[11] Blei D., NG A., and Jordan M. Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, pages 993-1022, 2003.

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

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

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

[15] Camille Guinaudeau. *Structuration automatique de flux télévisuels*. PhD thesis, INSA de Rennes, 2011.

[16] Camille Guinaudeau and Julia Hirschberg. Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. *In 12th Annual Conference of the International Speech Communication Association, Interspeech'11*, Pages 1401-1404, 2011.

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

[18] Marti A. Hearst. Multi-paragraph segmentation of expository texts. *In Proceedings of 32nd Annual meeting of the Association for Computational Linguistics*, pages 9-16, 1994.

[19] Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. *In Proceedings of the Special Interest Group on Information Retrieval*, 1993.

[20] Louis Hébert. *Tools for Text and Image Analysis: An Introduction to Applied Semiotics*. 2006. `http://www.revue-texto.net/Parutions/Livres-E/Hebert_AS/Hebert_Tools.ht%ml`.

[21] Nicolas Hernandez and Brigitte Grau. Analyse thématique du discours : segmentation, structuration, description et représentation. *In Actes du 5e Colloque International sur le Document Électronique*, pages 277-285, 2002.

[22] Julia Hirschberg and Christine H. Nakatani. Acoustic indicators of topic segmentation. *In Proceedings of the 5th International Conference on Spoken Language Processing*, pages 976-979, 1998.

[23] Yamron J., Carp I., Gillick L., Lowe S., and van Mulbregt P. A hidden Markov model approach to text segmentation and event tracking. *In Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 1998.

[24] Ewa Kijak. *Structuration multimodale des vidéos de sport par modèles stochastiques*. PhD thesis, University of Rennes 1, 2003.

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

[26] Galley M., McKeown K., Fosler-Lussier E., and Jing H. Discourse segmentation of multi-party conversation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

[27] Georgescul M., Clark A., and Armstrong S. Word distributions for thematic segmentation in a suport vector machine approach. *In Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLLX)*, pages 101-108, New York City, New York, 2006.

[28] Igor Malioutov and Regina Barzilay. Minimum cut model for spoken lecture segmentation. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25-32, 2006.

[29] David Mimno, Wei Li, and Andrew McCallum. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 633–640, New York, NY, USA, 2007. ACM.

[30] Hemant Misra and Fran cois Yvon. Modèles thématiques pour la segmentation de documents. *In Actes des 10e Journées Internationales d'Analyse Statistique des données textuelles*, pages 203-213, 2010.

[31] Marie-Francine Moens. Using patterns of thematic progression for building a table of contents of a text. *Natural Language Engineering*, 14(2):145–172, 2008.

[32] Marie-Francine Moens and Rik De Busser. Generic topic segmentation of document texts. *In Proceedings of the 24th International Conference on Research and Develoepement in Information Retrieval*, pages 418-419, 2001.

[33] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21-48, 1991.

[34] Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. Sits: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. *Association for Computational Linguistics*, 2012.

[35] Rebecca J. Passonneau and Diane J. Litman. Intention-based segmentation: Human reliability and correlation with linguistic cues. *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 148-155, 1993.

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

[37] Matthew Purver. Topic segmentation. In G. Tur and R. de Mori, editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 291–317. Wiley, 2011.

[38] François Rastier. Sémantique interprétative. *Presses universitaires de France*, 1987.

[39] Jeffrey C. Reynar. An automatic method of finding topic boundaries. *In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pages 331-333, 1994.

[40] Malcolm Slaney and Dulce Ponceleon. Hierarchical segmentation : Finding changes in a text signal. *In Proceedings of the 1st International Conference of the Society for Industrial and Applied Mathematics-Text Mining Workshop*, pages 6-13, 2001.

[41] Nicola Stokes, Joe Carthy, and Alan F. Smeaton. Segmenting broadcast news streams using lexical chains. *In Proceedings of the 1st Starting AI Researchers Symposium*, pages 145-154, 2002.

[42] Q. Sun, R. Li, D. Luo, and S. Wu. Text segmentation with LDA-based Fisher kernel. *In Proceedings of ACL-08: HLT, Short Papers*, pages 269-272, Columbus, Ohio, June 2008.

[43] Hofmann T. Topic segmentation of dialogue. *In Proceedings of the HLT-NAACL Workshop on Analyzing Conversations in Text and Speech*, New York, NY, 2006.

[44] Hofmann T. Probabilistic latent semantic indexing. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval*, pages 55-57,1999.

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

[46] Teh Y. W., Jordan M. I., Beal M. J., and Blei D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):15661581,2006.

[47] Yaakov Yaari. Segmentation of expository texts by hierarchical agglomerative clustering. *In Proceedings of the 2nd International Conference on the Recent Advances in Natural Language Processing*, 1997.