Topic segmentation: application of mathematical morphology to textual data

Sébastien Lefèvre[1] and Vincent Claveau[2]

[1] VALORIA Laboratory, University of South Brittany. Address: VALORIA Lab, Campus de Tohannic, BP 573, 56017 Vannes Cedex, France
sebastien.lefevre@univ-ubs.fr
[2] IRISA-CNRS. Address: IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France
vincent.claveau@irisa.fr

Abstract. Mathematical Morphology (MM) offers a generic theoretical framework for data processing and analysis. Nevertheless, it remains essentially used in the context of image analysis and processing, and the attempts to use MM on other kinds of data are still quite rare. We believe MM can provide relevant solutions for data analysis and processing in a far broader range of application fields. To illustrate, we focus here on textual data and we show how morphological operators (here the morphological segmentation using watershed transform) may be applied on these data. We thus provide an original MM-based solution to the thematic segmentation problem, which is a typical problem in the fields of natural language processing and information retrieval (IR).
More precisely, we consider here TV broadcasts through their transcription obtained by automatic speech recognition. To perform topic segmentation, we compute the similarity between successive segments using a technique called *vectorization* which has recently been introduced in the IR field. We then apply a gradient operator to build a topographic surface to be segmented using the watershed transform. This new topic segmentation technique is evaluated on two corpora of TV broadcasts on which it outperforms other existing approaches. Despite using very common morphological operators (i.e., the standard Watershed Transform), we thus show the potential interest of MM to be applied on non-image data.

1 Introduction

Mathematical Morphology (MM) has led to many successes in image analysis and processing. But its theoretical framework is much more general, and one can wonder why MM stays almost unknown in other fields, while it is expected to provide robust and efficient data analysis tools. In this paper, we address this issue and show that MM may provide very relevant solutions to problems encountered with non-image data. More precisely, we focus on topic segmentation which is a common problem of (Multimedia) Information Retrieval (IR) and Natural Language Processing.

Topic segmentation is of high interest in Multimedia IR. Indeed, it is needed to perform automatic structuring of TV streams, a keystone for every processing of such streams, which is still done manually in national archive agencies like the French INA. A way to obtain this structuration is to first transcribe the audio tracks of the TV streams into textual data, and then perform the topic segmentation from textual data to split the streams into semantic units (e.g., reports).

In this paper we address the problem of topic segmentation of textual data in this applicative framework using mathematical morphology. To do so, we show that topic segmentation and image segmentation have common characteristics (Sec. 2). From this observation we build a topic segmentation method based on the watershed transform. Moreover, we suggest to build the topographic surface from which the watershed lines are identified using a gradient computation method adapted to the problem under consideration, and thus consider here a vectorization-based gradient approach inspired from recent advances in IR (Sec. 3). Experiments performed on two TV broadcast corpora are presented and discussed (Sec. 4). Finally, Sec. 5 concludes this work and provides future research directions.

2 From image to text: links between morphological and topic segmentation

2.1 Morphological segmentation

Mathematical morphology is both a rich theoretical framework and a complete toolbox of efficient and robust tools for solving image analysis and processing problems. Among these problems, image segmentation aims to split the input image into a set of uniform regions given a predefined uniformity criterion (intensity or colour, texture, etc.) and is a preprocessing step required by many computer vision tasks. The most famous morphological method for image segmentation is certainly the watershed transform, even if connective segmentation has gathered great interest recently.

We recall very briefly the principle of watershed-based segmentation [10]. The image I to be segmented is first represented as a topographic surface. Watershed lines identified on this surface are then associated to region frontiers resulting from the segmentation process. This relatively simple principle led to various paradigms, and we consider here the flooding approach. It simulates the progressive flooding of the surface starting from its local minima, and builds some dams to avoid merging water from two different catchment basins. At the end of the process, dams correspond to the watershed lines or, in other words, to the region frontiers.

Most often, this approach is not directly applied on the image I to be segmented, since it would then seek for frontiers of high intensity (watershed lines) separating areas of lower intensity (catchment basins). Before applying the segmentation, an image transform is rather performed as a preprocessing in order

to highlight values of edge pixels and to lower pixel values in homogeneous areas. Among the transforms which may be involved, an image gradient (noted ∇I hereafter) is usually computed to enhance transition areas (which generally correspond to object frontiers). Various gradient computation methods exist; the most famous ones rely on a convolution with a weighted local mask to measure the dissimilarity within the neighbourhood of each pixel. The choice of the transform to be applied before the segmentation is of high importance, since it will directly influence the segmentation result produced by the watershed method. Indeed, gradient computation methods are often very sensitive to noise, and they often produce many local minima. Since each local minimum is associated to a new catchment basin, and thus to a new region in the image, watershed segmentation most often faces the problem of oversegmentation.

In order to reduce oversegmentation, several strategies may be considered: defining a robust gradient, including some oversegmentation reduction steps in the process (e.g., by merging basins or regions), or setting the predefined number of regions with markers which define the initial catchment basins [7]. These strategies may be supervised or not, respectively leading to user/knowledge-based and automatic methods. A supervised approach for morphological segmentation has been recently introduced in [3]. It relies on a fuzzy classification of the input multipsectral image, and suggests among other contributions to compute the gradient on class membership values associated to each pixel rather than original multispectral pixel values in the input image. Regions built by this method are thus composed of neighbouring pixels which share the same similarities to user-predefined classes, but not necessarily the same values in the input image. This method has been shown to reduce the oversegmentation phenomenon by considering a more robust data representation space. Later in this paper, we will inspire from this principle to transpose the morphological segmentation to the problem of topic segmentation.

2.2 From image to text

The analogy between image and text segmentation can be drawn very simply. The pixel is the base element in the image and is described by its greylevel or color/multispectral values. Its equivalent in texts is the sentence (or sometimes the paragraph) which is described by the words it contains.

In our framework of video segmentation, our texts are obtained from automatic transcription. These transcriptions are not composed of sentences but of utterances (sequence of words spoken between two breaths or silences) identified by a timestamp. These utterances are the minimal units of the text (i.e., they are equivalent to image pixels) and topic breaks will be sought between them.

Besides, our texts are flows of utterances. They are then represented as 1-D signal, while images are most often 2- or 3- dimensional. However, nothing prevents the watershed technique to be applied on a single dimension. Thus our approach relies on a gradient computed on the sequence of utterances, and topic breaks are identified using the watershed transform. Gradient computation, which is a key step of the segmentation process, is detailled in Sec. 3. The

watershed technique used here is the standard one described previously. We have only included a gradient smoothing step to remove irrelevant local minima.

3 Gradient computation using vectorization

3.1 Vectorization principles

Vectorization is an embedding technique which aims to project any similarity computation between two documents (or one document and one request in the context of IR) in a vectorial space. It has been introduced and experimented in a standard IR scenario [2] where it has shown to provide both a low complexity and accurate results. We recall here its main principles.

Its principle is relatively simple. For each document of the considered collection, it consists in computing with an initial similarity measure (eg. standard similarity measure used in IR), whatever it is, some proximity scores with m pivot-documents. These m scores are then gathered into a m-dimensional vector representing the document (*cf.* Fig. 1).

Comparing two documents (or a document and a request) can then be performed in a very standard way in this vectorial space (e.g., using a L_2 distance). Many algorithms are available to compute or approximate very efficiently such distances.

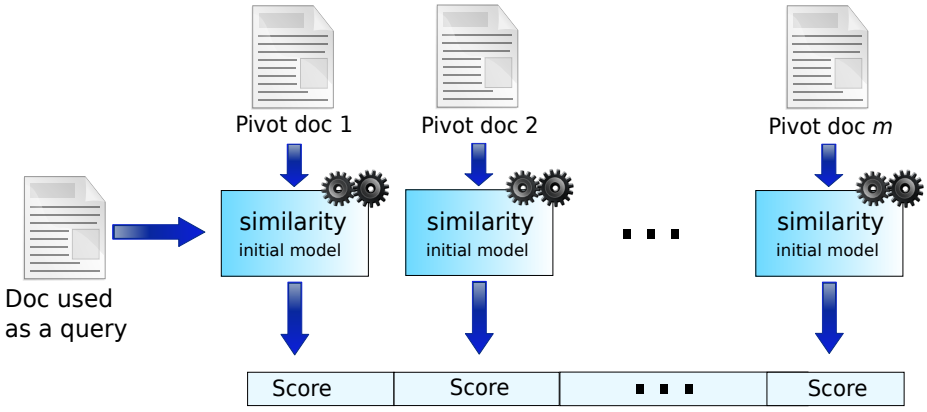

Fig. 1. Vector design from pivot-documents

More formally, we note $\text{Vect}(D, \mathcal{P}, Sim)$ the vector representing the document D built from the initial similarity measure Sim on pivot-documents $\mathcal{P}$. For instance, $\text{Vect}(D, [P_1, P_2, P_3], \text{TF.IDF/cosine})$ is a 3-dimensional vector; its first component is the similarity score between the document D and the pivot-document P_1 returned by a system using TF.IDF representation associated to

the cosine distance measure (which corresponds to a very standard way to compute similarities in the IR field; TF and IDF respectively stand for Term Frequency and Inverse Document Frequency [8, for details]), and so on for the next components.

3.2 Properties

It is important to notice that vectorization results in a change of the representation space, contrary to existing works consisting rather of a dimension reduction or a distance approximation (e.g., [1]). This space transform offers several nice properties which will be discussed here.

The first interest of this embedding is to reduce complexity when the initial similarity computation may be computationally expensive (e.g., some graph comparison computations used in complex IR systems). In a IR context, vectors associated with each document may be built offline, and when a request has to be processed, we only need to compute its similarity with the m pivot-documents rather than to do it with all documents in the collection. This property is nevertheless not useful in the context of a segmentation task.

The second nice property comes from the fact that two documents will be considered as similar if they are similar to the same pivot-documents. This indirect comparison, or second-order affinity, let us compare two textual documents which do not share any common word. This property will be helpful in our segmentation task. Indeed, it will solve the problem brought by the lack of repetition between utterances. This problem is particularly noticeable when the segments to be compared are of short duration (i.e., they will contain less words, and thus will share only a few words in common in the best case, and no common words in the worst case).

3.3 Usage

A gradient is computed between each utterance. In other words, we compute the similarity using the vectorization principle between previous and next utterances. Let us note that we do not compare only the previous to the next utterance, but we also consider the n previous ones vs. the n next ones (similarly to TEXTTILING, a common approach for topic segmentation).

In experiments described in the following section, the initial similarity measure used in the vectorization process is a L_2 distance associated with a weighting of utterances by $\sqrt{TF}$. It means that we first represent each breath group by a sparse vector in which each dimension represent a word; the value for this dimension is the square root of the number of occurrences of the word in the breath group. The same is done for the pivot document. The distance between the breath group vector and the pivot vector is computed with a L_2 distance; the resulting value forms one of the dimension of the new vector.

Similarly to some image gradient computation methods (e.g., Sobel), we give more importance to close utterances and less importance to utterances which are far to the candidate edge. This is ensured through a simple convolution with

a kernel (e.g., Gaussian kernel). Let us notice that the way the convolution is applied depends on the way the documents are represented in the initial model of similarity computation. With the vectorial representation used in our experiments, this convolution is simply taken into account: when computing $\sqrt{TF}$, the occurrence of a word counts for one in the breath group which is the closest from the candidate edge, but counts for less when considering an occurrence from a breath group further of the candidate edge. In practice, a linear penalty is applied. From now we will write $\mathcal{C}_{prev}(i)$ (respectively $\mathcal{C}_{next}(i)$) the result of the convolution operator applied on utterance i and those which are preceding (respectively following) it.

Formally, the gradient is thus defined by:

$$\nabla(i) = L_2(\text{Vect}(\mathcal{C}_{prev}(i-1), \mathcal{P}, \sqrt{TF}/L_2), \text{Vect}(\mathcal{C}_{next}(i), \mathcal{P}, \sqrt{TF}/L_2))$$

Pivot-documents we are using are simply sequences of utterances built from random splits of the considered broadcast. While each utterance is associated to a vectorial representation, we can observe than the gradient in each utterance results in a scalar value. Thus there is no need to use multivariate morphology and to adapt the watershed algorithm.

Fig. 2 shows an example of a vectorization-based gradient computed on a sample of one of our corpora (see below). We can observe that the signal contains local minima and thus needs a smoothing step. As indicated previously, we involve a smoothing step to remove such minima and help the watershed segmentation process. Resulting segmentation is provided in Fig. 3. The considered sample contains 4 segments (ground truth is shown in green full line; detected edges by our system using watershed transform are given in dotted lines). Utterances are represented by their starting time. For a given time index, the higher the smoothed gradient is, the more important the dissimilarity between previous and next groups is. In other words, significant local maxima of gradient values indicate a topic break. Nevertheless, local maxima are not sufficient to identify topic breaks. Indeed, we cannot make any assumption regarding the segment length. Extraction of local maxima by signal analysis with a sliding window is then inappropriate. Moreover, such an approach would have lack of robustness.

4 Experiments

4.1 Experimental data

Our experiments are performed on two French TV broadcast corpora for which the topic segmentation is of high interest. The first corpus is a set of 60 TV news of the France 2 channel (called *News* further). Each of these sample has been broadcasted in the beginning of 2007 and is 40 minutes long. The second corpus is made from TV reports: 12 samples of *Envoyé spécial* (2008, 2 hours long each), and 16 *Sept à huit* (2008, 1 hour long each). This corpus is called *Reports* in the following experiments.

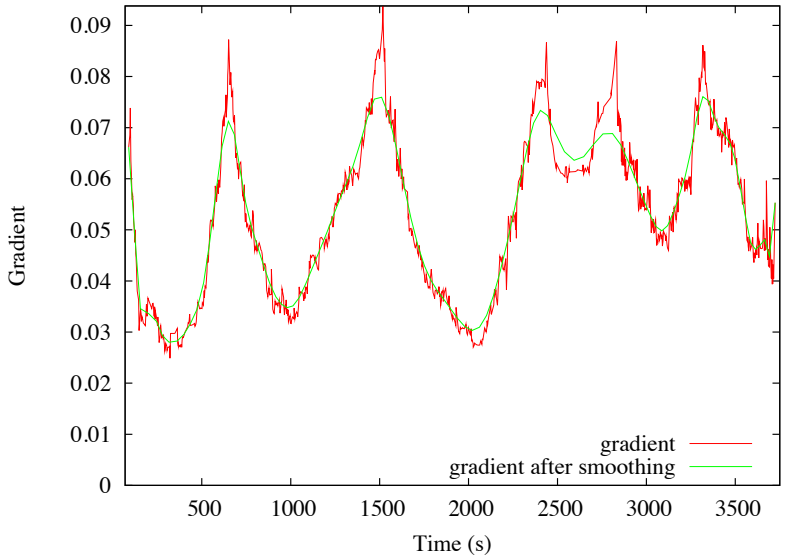

Fig. 2. Effect of the filtering step: gradient vs. starting time of utterances

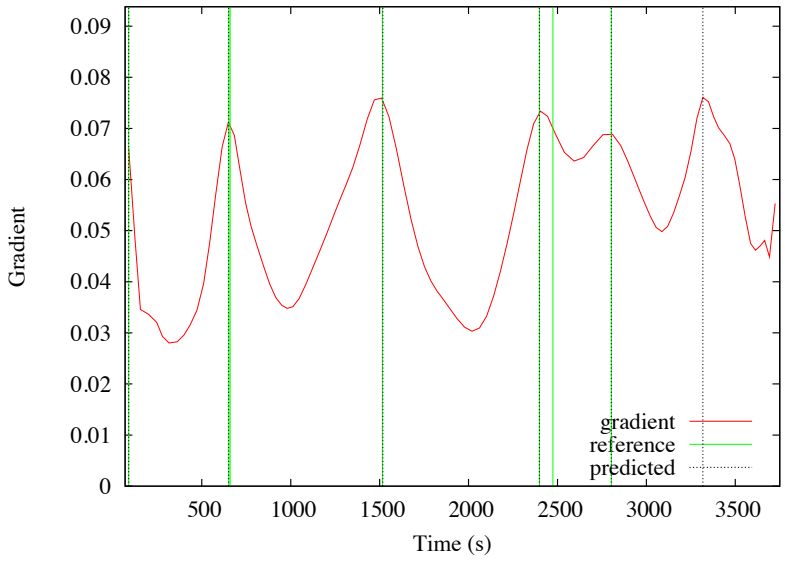

Fig. 3. Illustration of segmentation process: filtered gradient vs. starting time of utterances

These corpora [4] have different properties in terms of number and duration of topic segments. Thus, it allows us to evaluate robustness of topic segmentation methods. The *News* corpus contains 1180 segments while the *Reports* corpus only contains 140 segments.

The reference segmentation (i.e., ground truth) has been independently built by a user who was not involved in the design of a topic segmentation system. Since there is no consensus on the topic definition in the IR or NLP fields, it has been considered here that a topic change occurs for each report change. Despite this assumption being not always valid (in particular in the *News* corpus in which several successive reports may be considered as related to the same topic), it is relevant since it corresponds to an actual and well-defined applicative need.

4.2 Data preprocessing

Audio tracks of these two corpora have been automatically transcribed using the speech recognition system IRENE [6]. This system has been initially designed for transcribing radio broadcasts for which it produces a Word Error Rate of about 20%. In our context of transcribing TV broadcasts, it is very probably higher due to the more noisy environment we have to face. Transcriptions are finally part-of-speech tagged using TreeTagger[3], and only names, verbs, and adjectives are kept and stemmed.

4.3 Results

Recall, precision, and F1-measure are used as quality measures to evaluate our proposed method. We consider that a segment edge is correct as soon as it is located in the close neighbourhood (less than 10 seconds) of a reference frontier. In order to show the relevance of our contribution, we compare the results obtained by our MM-based method to those (when available) produced by several existing systems on the same corpora: the system from Utiyama and Isahara [9] relying on a Hidden Markov Model (we use the implementation from [4]), and the best results obtained from the system of [4]. We also provide results obtained by a self-implementation of TEXT-TILING [5] in which we use the same data preprocessing and the same watershed-based segmentation framework. The only difference is the way the gradient is computed (i.e., without vectorization), which can be here written:

$$\nabla(i) = cosine(\text{TF-IDF}(\mathcal{C}_{prev}(i-1)), \text{TF-IDF}(\mathcal{C}_{next}(i)))$$

The TEXT-TILING approach aims to find topic breaks where lexical coherence between previous and next text blocks is linked to a significant local minimum. That is why we have derived a watershed-based implementation by considering the inverse of the lexical coherence measure as the topographic surface to be used in the watershed process.

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

Table 1 shows results obtained on both corpora. In both cases, we can observe that our system provides better performances than existing systems. In order to better understand the interest of using MM for topic segmentation, we compare more deeply the approach introduced in this paper and our own implementation of TEXT-TILING. Both are based on the Watershed transform but they differ on the gradient computation method. We rely here on a vectorization technique rather than a cosine measure combined with a TF-IDF representation (a common approach in IR). The superiority of our approach is particularly observable on the *News* corpus, since this corpus contain very short segments, thus making the direct computation of the gradient as done in TEXT-TILING unreliable. In other words, the straight application of morphological operators may be of lower interest. It is much more relevant to adapt the morphological process to the data under consideration, e.g. here to use an appropriate gradient computation method.

method	*News* corpus			*Reports* corpus		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
Utiyama [9]	-	-	59.44	-	-	51.09
Guinaudeau [4]	-	-	61.41	-	-	62.92
TEXT-TILING[5]	44.17	41.97	43.04	59.32	60.93	60.12
proposed	**67.47**	**61.6**	**64.4**	**77.38**	**69.65**	**73.31**

Table 1. Performances of topic segmentation systems on *News* and *Reports* corpora

5 Conclusion

In this paper, we aim to show that Mathematical Morphology can be successfully applied on non-image data. To do so, we consider the topic segmentation problem faced in the fields of information retrieval and natural language processing. We show the parallel which can be driven between topic and image segmentation. From this parallel we were able to introduce a new topic segmentation method based on morphological segmentation using the watershed transform. The results are appealing and strongly suggest that Mathematical Morphology would benefit to many fields and not only to image analysis and processing.

Moreover, we have included in this approach a vectorization-based gradient computation method. The experiments we have made lead to the expected conclusion that Mathematical Morphology should not be straightly applied to textual data. Indeed, it is more relevant to adapt some steps of the morphological data processing scheme to the data under consideration. More precisely, the vectorization technique used here is of great help to face the lack of repetitions between utterances, which is an important problem when topic segments are quite short. Applying such a gradient on the input signal, using similarity to given samples (here the pivot-documents), is not new in the field of watershed-based image segmentation. In a previous work [3], we have already suggested to

build the topographic surface through a gradient computation made on probability maps obtained from a supervised image classification. The regions produced by this method are then composed of neighbouring pixels sharing the same similarities to predefined classes rather than similarities in the initial multispectral image space. A full adaptation of this strategy to textual data will result in defining highly reliable pivot-documents (which would be more discriminative than the random ones used in this article by the vectorization technique).

As we briefly recalled in Sec. 2, there exist other techniques leading to a better segmentation result. In particular, future work will consider the intelligent morphological segmentation paradigm. In complement to the definition of a relevant image transform to build the topographic surface, the markers could also be of great interest. Besides, marker-based watershed segmentation of textual data may be a way to involve the user in the process if required. Moreover, we also consider hierarchical morphological segmentation schemes to build a multiscale topic segmentation result. This could be of great interest for the Multimedia IR community.

References

1. Abraham, I., Bartal, Y., Neiman, O.: Advances in metric embedding theory. In: Proc. of Symposium on Theory Of Computing. Seattle, USA (2006)
2. Claveau, V., Tavenard, R., Amsaleg, L.: Vectorisation des processus d'appariement document-requête. In: 7e conférence en recherche d'informations et applications, CORIA'10. pp. 313–324. Sousse, Tunisie (Mar 2010)
3. Derivaux, S., Forestier, G., Wemmert, C., Lefèvre, S.: Supervised segmentation using machine learning and evolutionnary computation. Pattern Recognition Letters 31(15), 2364–2374 (2010)
4. Guinaudeau, C., Gravier, G., Sébillot, P.: Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels. In: Actes de la conférence Traitement automatique des langues. Montréal, Canada (2010)
5. Hearst, M.: Text-tiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics 23(1), 33–64 (1997)
6. Huet, S., Gravier, G., Sébillot, P.: Morpho-syntactic post-processing with n-best lists for improved french automatic speech recognition. Computer Speech and Language 24(4), 663–684 (October 2010)
7. Rivest, J., Beucher, S., Delhomme, J.: Marker-controlled segmentation: an application to electrical borehole imaging. Journal of Electronic Imaging 1(2), 136–142 (1992)
8. Salton, G.: A Theory of Indexing. Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphy (1975)
9. Utiyama, M., Isahara, H.: A statistical model for domain-independent text segmentation. In: Proceedings of the 9th conference of the ACL (2001)
10. Vincent, L., Soille, P.: Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(6), 583–598 (1991)