INTRA-CONTENT TERM WEIGHTING FOR TOPIC SEGMENTATION

Abdessalam Bouchekif [1,2] *Géraldine Damnati* [1] *Delphine Charlet* [1]

[1] Orange Labs, Multimedia contents Analysis technologieS, Lannion, France.
[2] Laboratoire d'Informatique de l'Universite du Maine, LIUM - France.

ABSTRACT

Term weighting is an important task in many applications, such as information retrieval, extraction of significant words or automatic summarization. It translates the capacity of a term to discriminate a document within a collection, or a part of a document within a whole document. This paper deals with term weighting strategies in the context of lexical cohesion based topic segmentation. The aim is to propose a term weighting method which does not require any external information data. Weights are estimated from the content itself which is considered as a collection of mono-thematic documents. Two approaches are proposed and significant improvements are observed on a rich corpus covering various formats of Broadcast News shows from 8 French TV channels.

Index Terms— Topic segmentation, lexical cohesion, term weighting, TF-IDF, $Okapi$.

1. INTRODUCTION

Topic segmentation (TS) consists in splitting a document (text, audio or video) into thematically homogeneous fragments. Several systems for TS of TV Broadcast News (TVBN) are described in the literature. Three categories of cues or features have been explored in the task of TS : *lexical*, *acoustic* and *visual* cues. **Lexical cues** borrowed from traditional text segmentation. Main lexical approaches include the notions of lexical cohesion introduced in [1] and lexical chaining [2], [3], [4]. [5] adapted statistical models designed for text segmentation in [6] to TVBN. **Acoustic cues** can be *pause duration*, *speech type*, *pitch*, etc... **Visual cues** (such as a *news title caption*, *logo detection*, *shot detection*, etc...), can also reveal topic shifts but they heavily rely on editorial rules [7]. Higher level cues based on role analysis can be used. Anchor detection based either on lexical, acoustical [8], [9], [10] or visual cues (anchor face [11]) can be strongly correlated with story boundaries, but it is also very dependent on editorial rules. Dumont et al. [12] showed that audio cues are more important than visual cues. In this paper, we rely on the audio, not exploiting any information from the video. We give more importance to lexical cues because they reveal topic boundaries via semantic variations across the transcription.

Various methods for TS based on lexical cohesion have been proposed relying on similarity computation between two vectors of word counts. For example the TextTiling approach [1] measures similarity between pairs of blocs using sliding windows along the show. Local minima are detected as topic boundaries. In $C99$ approach [13], the similarity between each pair of sentences is computed. The algorithm uses a local ranking of the sentence similarity matrix and a clustering strategy. For the MinCut approach [14] authors see the segmentation task as the partition of graph which is none other than a representation of the similarity matrix.

Malioutov et al. [14] insisted on the fact that weighting plays an important role in segmentation performance. Different weight scores such as TF-IDF and $Okapi$-$BM25$ are widely used in Information Retrieval (IR) to translate the capacity of a term *t* to discriminate a document *d* relatively to a collection of documents. For TS, TF-IDF coefficients are usually estimated on large corpora (e.g [15] make use of a keyword extraction tool *kiwi* [16]), this solution is very dependent of language models (training data period). Originaly [14] proposed to split the content into uniform chunks, simulating documents in the IR terminology. These chunks are used to compute intra-document TF-IDF weights. Oracle experiments of TS where we compute weights from chunks corresponding to reference topic segments of the content revealed a significant gap in performances. Following this observation, we propose several approaches attempting to improve partioning into chunks and to bridge the gap towards Oracle performances. Section 2 presents our lexical cohesion based algorithm. Section 3 proposes two novel weighting computation approaches. Experiments are presented and discussed in section 4.

2. TOPIC SEGMENTATION ALGORITHM

As in the original TEXTTILING approach, similarity is computed between each pair of adjacent blocs (one bloc equals a set of sentences, ...). In automatic transcription, sentence boundary detection is not a trivial task, there are neither ponctuations nor capital letters, but it contains breath groups (BG) which are sequences of words between two pauses in a speech turn (the minimum and maximum size of the pauses are 0.03 and 0.37 respectively). Pauses are automatically detected by the Automatic Speech Recognition engine. Hence similarity

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

is computed between weighted term vectors of adjacent blocs of K BGs, with a sliding analysis window. The similarity values constitute a cohesion curve, each point of the curve being the value associated to a potential boundary. A selection algorithm is applied in order to detect disruptions on the curve.

2.1. Intra-document weighting

The principle is to split the show into N chunks, each of them represents the notion of document in the classical IR. A term t in a breath group x will be associated to a weight $w(c(x), t)$ depending on the chunk $c(x)$ in which it occurs.

$$w_{TfIdf\ (c(x),t)} = TF_{(c(x),t)} \times IDF_{(t)} \quad (1)$$

where

$TF_{(c(x),t)}$ is the frequency of term t in chunk $c(x)$

$IDF_{(t)} = \log(\frac{N}{n_t})$

n_t is the number of chunks containing term t.

Okapi is similar to TF-IDF but takes better into account the length of the blocks.

$$w_{BM25\ (c(x),t)} = TF_{BM25\ (c(x),t)} \times IDF_{BM25(t)} \quad (2)$$

where

$TF_{BM25(t)} = \frac{TF_{(c(x),t)*}\,(k+1)}{TF_{(c(x),t)+}\,k\,(1-b+b*dl_{(c(x))}\,/dl_{avg})}$

$IDF_{BM25(t)} = \log\left(\frac{N-n_t+0.5}{n_t+0.5}\right)$

$dl(c(x))$ is the length of chunk $c(x)$ (number of words).

dl_{avg} is the average chunk length.

Following [17], b and k are classicaly set to 0.75 and 2 respectively. The benefit of this intra-content approach is that it does not require *a priori* information unlike other approaches. In [14], the authors split the show into uniform chunks, in section 3 we propose alternative partitioning approaches.

2.2. Similarity Computation

The widely used cosine similarity allows to measure the proximity between representation vectors V_j and V_{j+1} of two adjacent blocs b_j and b_{j+1}. The vector coordinate of term t in bloc b is a weighted value $v(b,t)$. In our representation, there is not one unique weight for a term t in bloc b as breath groups of the bloc may not all belong to the same chunk. Hence, the bloc level weighted value $v(b,t)$ is obtained by summing up weighted values for each BG x contained in the bloc.

$$v(b,t) = \sum_{x \in b} \left(f_{x,t} \times w\left(c\left(x\right),t\right)\right) \quad (3)$$

where: $f_{x,t}$ is the frequency of term t in BG x.

For a given potential boundary between blocs b_j and b_{j+1}, the similarity is $cohesion(j) = cosine(V_j, V_{j+1})$.

2.3. Boundary selection

Several strategies have been explored to select boundaries from the lexical cohesion curve plotted along all the BGs of a show. The classical approach [1] consists in detecting valleys as the lowest point between two peaks (local maxima), and selecting x_j as a boundary if the valley depth $d(j)$ is above a given threshold. The depth is computed as the sum of the differences between the lowest point value and the left and right peaks values. Note that $d(j)$ is equal to 0 when there is no valley at the j^{th} position. Directly working on valley depth is not always optimal as some topic segments may not contain enough term repetitions to yield high lexical cohesion values. Similarly, directly searching for low lexical similarity values is not optimal (some topic shifts between closely related topics can result into average lexical similarity values). We use a combination of those measures through linear interpolation. For a potential boundary j the score is given by:

$$score(j) = \lambda(1 - cohesion(j)) + (1-\lambda)depth(j) \quad (4)$$

We fixed λ to 0.75. This score emphasizes low values of cohesion which are also local minima.

Rather than simply applying a threshold to determine high values of this score, we propose an iterative splitting algorithm. Claveau et al. In ([17]) applied an alternative algorithm called "Watershed" derived from the mathematical morphology to achieve partitioning of the shows from the curve. Our iterative splitting algorithm is also designed to alleviate the local maxima phenomenon:

1. For initialisation, the entire show is considered as one coherent segment.
2. Each segment is split into two, the split point is the maximum value of score (if above a fixed threshold).
3. BGs within a time interval around the selected split point are discarded for the next iterations.
4. The resulting segments are submitted to step 2.
5. Stop if there is no possible partition.

Step 3 guarantees that local phenomena with several very close high values of the score will not lead to several consecutive topic boundary selections.

3. IMPROVED TERM WEIGHTING

Splitting the show into N uniform chunks does not really reflect the importance of terms in the topic. Ideally the partition into chunks should be as close as possible to the topic boundaries. In this section, we introduce two approaches for chunk definition. The first one consists in using an iterative weighting scheme, in the second one we use structural information.

3.1. Iterative intra-content weighting scheme

We introduce an iterative estimation using the result of our automatic thematic segmentation algorithm to determine chunks. Topic segmentation obtained at a given iteration provides a set of documents from which weights are re-estimated for the next iteration. For initialization, the show is splitted into N uniform chunks. The number of chunks

N is computed automatically for each show as a function of its overall duration and the average topic duration estimated on a held-out set of shows. The beginning of each uniform chunk is considered as the initial set of topic boundaries, and is placed in a vector hyp_0. At iteration i, hyp_{i-1} hypotheses are used to estimate $w(i)$ weights. The linear combination of lexical similarity and valley depth is re-estimated and the *splitting* algorithm is applied to determine the assumptions hyp_i. The algorithm stops when no significant change in the set of hypotheses is obtained. In order to determine an objective stopping criterion, we use the p_k [18] evaluation metric. $p_k(R, H)$ compares a reference segmentation R and an hypothesis segmentation H. The stopping criterion is reached when the value of p_k between hyp_{i-1} and hyp_i is near 1 (i.e there isn't large change between the two sets of boundary hypotheses):(1-p_k(hyp_{i-1},hyp_i))$\leq \epsilon$. Note that we haven't yet studied evidence of the algorithm convergence. The algorithm stops after 6 iterations if p_k doesn't reach the stopping criterion threshold.

3.2. Using structural information for chunk definition

In this section, we propose to make use of structural information in order to guide the definition of chunks. Among several possible structural information sources, we have selected the information provided by anchor speaker turns. In fact, anchor speakers are traditionaly in charge of introducing new topics. Eventhough this information constitutes an important cue, we have shown in previous work [10] that it is not sufficient in itself. If traditional Broadcast News shows might follow this pattern (an anchor speaker introducing a new topic, followed by a report) some channels propose a more "modern" structuration of their shows. For instance, some shows contain *readers*, where the anchor reads a (usually short) topic without any additional illustration. Several short *readers* can occur one after the other leading to several topics within a single anchor speaker turn. A topic may also not involve the anchor person at all.

In previous work [10], we have proposed to use the anchor speaker information during the boundary selection process. In this paper, our focus is on the definition of chunks for term weighting strategies and we propose to use the anchor speaker information in order to define chunks. To this purpose, the beginning of each anchor speaker turn is considered as the beginning of a new chunk. Speaker role analysis is performed following the multi-stage process described in [19]. This partition in chunks can be used in the algorithm presented in section 2 and can also be used as the initialization step in the iterative framework proposed in section 3.1.

4. EXPERIMENTS AND RESULTS

4.1. Corpus

Experiments are carried out on two sets of shows. The first set for development (Dev) is composed of 33 French TVBN shows broadcasted between October 2008 and January 2009 from 7 different channels (TF1, France2, France3, LCI, France24, Arte, M6). It contains 379 segments. Then, in order to evaluate the different approaches, a $Test$ set is used, composed of 23 shows from a new channel: D8 (during the month of October 2013). Table 1 describes the two corpora. Lexical units are lemmas obtained thanks to the lia_tagg [1] software. For a given show, the set of T lexical terms (or tokens) is the set of different lemmas obtained after discarding function words and words whose confidence score is below a given threshold. Automatic transcription is performed with the VoxSigma speech recognition engine of Vocapia Research, based on [20]. It achieves 16.1% word error rate on our Dev corpus. Manual transcriptions are not available for the Test corpus preventing ASR performance evaluation. Performances of TS are measured in terms of recall and precision by comparing time information associated to hypotheses and reference boundaries. As frequenty found in the litterature, an interval of $10s$ before and after a boundary hypothesis is tolerated in order to decide if it is correct.

	Dev	**Test**
Number of shows	33	23
Av. duration	$\sim$ 22 min	$\sim$ 13 min
Nb. of topic boundaries	379	140
Nb. of topics per show	11.5	6
Av. duration of topics	115 s	128 s

Table 1. Corpora description

4.2. Results

Weigthing	TF-IDF			Okapi-BM25		
condition	R	P	F	R	P	F
Uniform	58.3	51.7	**54.8**	51.4	60.6	**55.6**
Uniform+Iter	60.7	59.1	**59.9**	68.3	57.3	**62.3**
Anchor	69.4	60.0	**64.3**	61.2	63.9	**62.5**
Oracle	77.3	70.1	**73.5**	71.0	69.2	**70.1**

Table 2. Influence of chunk definition on Dev corpus (Recall, Precision, F-measure)

Table 2 illustrates the performances obtained with different conditions of chunks selection and for the two weighting schemes (TF-IDF and Okapi-BM25). The results show that the best performance (73.5%) is achieved with chunks derived from the reference boundaries (maximum Oracle that can be achieved with intra-content weighting). This Oracle result comforts the potential impact of the chunking strategy and reflects that significant improvements can be achieved only with a suitable weighting computation.

[1] http://pageperso.lif.univ-mrs.fr/ frederic.bechet/download.html

The iterative approach outperforms the baseline algorithm. When applied to uniform chunks with TF-IDF, the F-max raises from 54.8% to 59.9%. Using structural information also increases performances of the system. We didn't observe a considerable benefit after applying the iterative algorithm to the anchor information. All approaches have been experimented with $Okapi$ weighting. Here again, both approaches outperform the baseline Uniform approach, with a better behavior of the Uniform+Iter approach in this context. When observing in details the improvement provided by the iterative algorithm, we can see that it helps retrieving boundaries between relatively similar topics (two consecutive sport news, two consecutive reports on a same country).

In order to better observe the influence of the structural information provided by the anchor speaker turns, we have split our Dev corpus into two sets of shows. The first one is composed of *traditional* shows that follow the classical anchor/report pattern. It gathers 23 shows, corresponding to 279 topic boundaries. The second set is composed of so-called *modern* shows that contain unconventional patterns. It gathers 10 shows from Arte, M6 and France3, corresponding to 100 topic boundaries. M6 shows are composed of a succession of anchor voiceovers, without any stage scene, and with only few other speaker turns. In the middle of the Arte shows, a succession of short voice-overs narrated by a reporter is included among traditional packages. At the end of the France3 shows, series of local news package extracts are added to the national News show, each of them being narrated by their corresponding local reporter.

For the sake of comparison, if we directly select the beginning of anchor speaker turns as topic boundaries, we would get a 56.5% F-measure (58.8% recall and 54% precision) on the overall Dev corpus. This raw structural segmentation yields 59.7% F-measure on the $traditional$ sub-corpus and 47.2% F-measure on the $modern$ sub-corpus. Structural information is better suited to traditional shows. The following figures illustrate the behaviour of our different chunk selection approaches on the two sub-corpora.

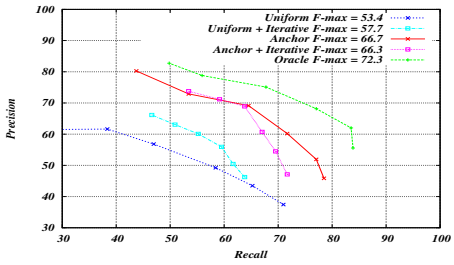

Fig. 1. Influence of chunk definition on $Dev - Traditional$

In figure 1, we can see that the Anchor approach outperforms any other approach for the $traditional$ shows. Additionaly applying the iterative algorithm over this Anchor based partition doesn't provide any improvement. On the other hand, figure 2 shows that for the more $modern$ settings,

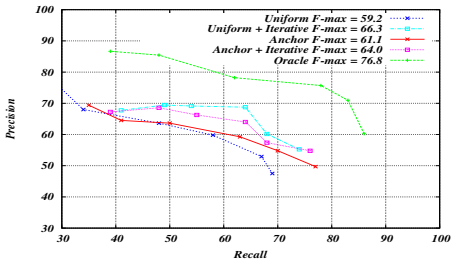

Fig. 2. Influence of chunk definition on $Dev - Modern$

the iterative approach over the Uniform partition provides better results. Anchor partition can be improved by Anchor+Iter but still doesn't reach the Uniform+Iter performances. This confirms that structural information can be very helpful when shows follow a regular setting but an approach that doesn't make use of any structural information proves to be more robust over various types of shows.

Finally, we have run a set of experiments (shown in Table 3) on a new type of shows from the D8 channel. This show follows the $traditional$ setting with an alternance of anchor speaker turns and reports. Similar overall performance are observed with the TF-IDF framework with a better recall on this corpus.

Weigthing condition	$Test$		
	R	P	F
Uniform	70.0	39.9	**50.8**
Uniform+Iter	70.7	47.6	**56.9**
Anchor	78.8	60.0	**68.0**
Anchor+Iter	75.0	53.3	**62.3**
Oracle	87.9	62.8	**73.2**

Table 3. Experimental results on $Test$

5. CONCLUSION

This paper adresses the influence of term weighting in lexical cohesion based topic segmentation. Intra-content term weighting is performed, relying on the partition of the show into chunks which simulate a collection of documents. Two approaches are proposed in order to outperform the state-of-the-art uniform chunk partition. In the first one, weights are estimated iteratively. Topic segmentation obtained at a given iteration provides a set of documents from which weights are re-estimated for the next iteration. The second approach makes use of structural information provided by anchor speaker turns detection. Both propositions yield significant improvements on a varied corpus of TVBN from 8 different channels. The benefit of the iterative approach over the anchor-based chunk partition is that it can be applied for any type of show. The approaches of weighting described in this paper can be applied on any topic segmentation algorithm which uses lexical cohesion.

6. REFERENCES

[1] M. Hearst, "TextTiling: Segmenting Text Into Multi-Paragraph Subtopic Passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[2] N. Stokes, J. Carthy, and A. Smeaton, "SeLeCT: a Lexical Cohesion Based News Story Segmentation System," *AI Communications*, vol. 17, no. 1, pp. 3–12, 2004.

[3] L. Sitbon and P. Bellot, "Topic Segmentation Using Weighted Lexical Links (WLL)," in *Special Interest Group on Information Retrieval*, 2007, pp. 737–738.

[4] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing, "Discourse Segmentation of Multi-party Conversation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2003, ACL '03.

[5] C. Guinaudeau, G. Gravier, and P. Sébillot, "Enhancing Lexical Cohesion Measure With Confidence Measures, Semantic Relations and Language Model Interpolation for Multimedia Spoken Content Topic Segmentation," *Computer Speech and Language*, vol. 26, no. 2, pp. 90–104, 2012.

[6] M. Utiyama and H. Isahara, "A Statistical Model for Domain-Independent Text Segmentation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 2001, pp. 499–506.

[7] L. Xie, Y. Yang, Z. Liu, W. Feng, and Z. Liu, "Integrating Acoustic and Lexical Features in Topic Segmentation of Chinese Broadcast News Using Maximum Entropy Approach," in *International Conference on Audio, Language and Image Processing*, 2010, pp. 407–413.

[8] R. Amaral and I. Trancoso, "Exploring the Structure of Broadcast News for Topic Segmentation," in *Language Technology Conference*, 2007, pp. 1–12.

[9] A. Rosenberg and J. Hirschberg, "Story Segmentation of Broadcast News in English, Mandarin and Arabic," in *Conference of the North American Chapter of the Association for Computational Linguistics*, 2006, pp. 125–128.

[10] A. Bouchekif, G. Damnati, and D. Charlet, "Complementarity of Lexical Cohesion and Speaker Role Information for Story Segmentation of French TV Broadcast News," in *International Conference on Statistical Language and Speech Processing*, 2013, pp. 51–61.

[11] X. Wang, L. Xie, M. Lu, B. Ma, E. Chng, and H. Li, "Broadcast News Story Segmentation Using Conditional Random Fields and Multimodal Features," *IEICE Transactions*, vol. 95-D, no. 5, pp. 1206–1215, 2012.

[12] E. Dumont and Q. Georges, "Automatic Story Segmentation for TV News Video Using Multiple Modalities," *International Journal of Digital Multimedia Broadcasting*, vol. 2012, 2012.

[13] F. Y. Y. Choi, "Advances in Domain Independent Linear Text Segmentation," in *Conference of the North American Chapter of the Association for Computational Linguistics*, 2000, pp. 26–33.

[14] I. Malioutov and R. Barzilay, "Minimum Cut Model for Spoken Lecture Segmentation," in *International Conference on Computational Linguistics*, 2006, pp. 25–32.

[15] C. Guinaudeau and J. Hirschberg, "Accounting for Prosodic Information to Improve ASR-Based Topic Tracking for TV Broadcast News," in *INTERSPEECH*, 2011, pp. 1401–1404.

[16] G. Lecorvé, G. Gravier, and P. Sébillot, "An Unsupervised Web-based Topic Language Model Adaptation Method," in *International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 5081–5084.

[17] V. Claveau and S. Lefèvre, "Topic Segmentation of TV-Streams by Mathematical Morphology and Vectorization," in *INTERSPEECH*, 2011, pp. 1105–1108.

[18] D. Beeferman, A. Berger, and J. Lafferty, "Statistical Models for Text Segmentation," *Machine Learning*, vol. 34, no. 1-3, pp. 177–210, 1999.

[19] G. Damnati and D. Charlet, "Multi-View Approach for Speaker Turn Role Labeling in TV Broadcast News Shows," in *INTERSPEECH*, 2011, pp. 1285–1288.

[20] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.