FUSION OF SPEAKER AND LEXICAL INFORMATION FOR TOPIC SEGMENTATION: A CO-SEGMENTATION APPROACH

Delphine Charlet[1], Géraldine Damnati[1], Abdessalam Bouchekif[1,2], Ameur Douib[1]

[1]Orange Labs, Lannion, France
[2]LIUM, Université du Maine, Le Mans, France
firstname.lastname@orange.com

ABSTRACT

In this work, we investigate how speaker-based information and lexical-based information can be fused efficiently for topic segmentation of spoken contents. While in recent work, we have proposed an early fusion scheme, so as to jointly model speaker and lexical distribution, we propose here a co-segmentation framework, between segmentations performed in the speaker space and in the lexical space. Experiments carried out on two distinct corpora (Radio talk show and TV Broadcast News) show that, even if performances of speaker information are contrasted and closely related to the content structure, its integration with lexical information, either by early fusion or by co-segmentation, is always effective. Absolute gains of 16% (Radio corpus) and 5% (TV corpus) are observed for topic boundary detection performance.

Index Terms— Topic segmentation, speaker cohesion, lexical cohesion, co-segmentation

1. INTRODUCTION

Topic segmentation (TS) is a task which consists in finding thematically homogeneous fragments (talking about a same subject) in a given content. It can be used to quickly browse the content, or as a preliminary step for other processing such as summarization. Three categories of cues have been explored. *Lexical cues* are based on the exploitation of the automatic transcription. Main lexical approaches rely on the notions of *lexical cohesion* introduced in [1] and *lexical chaining* [2]. *Acoustic cues* can be pause duration, jingles or speech type [3][4][5]. *Visual cues* (title caption, logo, shot boundaries, etc...), can also reveal topic shifts but they heavily rely on editorial rules [6]. In this paper, we focus on the audio channel, not exploiting any information from the video. Our approach can apply to any spoken Broadcast content, from TV or radio.

Various methods for TS based on lexical cohesion have been proposed relying on similarity computation. TextTiling algorithm [1] measures lexical cohesion between adjacent pairs of blocs using a sliding window along the show. High cohesion values imply that there are terms in common between these two blocks and that they are likely to belong to the same subject. Even if other more sophisticated algorithms have been proposed in the literature (e.g. C99 [7] or MinCut [8]), we have chosen to keep the sliding window paradigm in order to deeply study the impact of data representation. More specifically, we are interested in the complementarity between the representation of the audio content in the lexical space and the speaker space.

Speaker role had been exploited in several studies. In previous work [9], we showed that detection of the *anchor* speaker can be helpful for topic segmentation. However this information is not sufficient in itself as a new topic can be introduced by another speaker, and on the opposite, the anchor speaker can occur several times during a single topic. Hence we propose to exploit speaker distribution in more details. Indeed, for some contents, the distribution of speaker turns varies from one topic to another, and a change in speaker distribution can be correlated to a topic change. Most of TS that exploit speaker information have adopted supervised framework (i.e using classifiers). For example, [6] exploit speaker segmentation via binary features (speaker change / no change). Dumont et al. [10] use the result of speaker diarization by adding the index of the speaker in each observation. In previous work [11], we have proposed an early fusion scheme to exploit speaker distribution information and have introduced the paradigm of *speech cohesion*, a new notion that generalizes lexical cohesion by jointly taking into account lexical and speaker information. In this paper, we investigate other fusion schemes between speaker and lexical information and show that their complementarity can be helpful in various types of audio contents, beyond TV Broadcast News.

The approach proposed here is inspired by the co-training paradigm [12] introduced for semi-supervised classification where several views of the same spoken language understanding problem are available. We consider the speaker space and the lexical space as two views and propose to exploit the segmentation from one view to improve the segmentation in the other view. This approach is referred to as *co-segmentation*.

Section 2 presents the topic segmentation algorithm while section 3 introduces the co-segmentation process. Experiments on two corpora are presented in section 4.

2. TOPIC SEGMENTATION ALGORITHM

Our algorithm is based on the principle of the Text Tiling algorithm [1]: cohesion is computed between adjacent blocs, using a sliding window along the show. The unitary element, when analyzing speech transcriptions, is the *breath group* (BG), *i.e.* sequences of words between two pauses. Hence, a bloc corresponds to a window of a given number of BGs. Low cohesion values mean that there are few terms in common between adjacent blocs, and that the boundary between these blocs is likely to be a topic boundary.

2.1. Lexical cohesion based segmentation

Previously to any cohesion computation, the automatic transcription of the show is submitted to a 3-fold pre-processing step: lemmatization; filtering of function words in order to keep only nouns, adjectives and verbs; filtering of low confidence value words. The remaining lemmatized words constitute a list of terms which is used for the vectorial representation of the show. Each breath group x has a vectorial representation where the coordinate for term t is the number of occurrences $f_{x,t}$ of t in x.

2.1.1. Intra-document term weighting

The aim of term weighting in Information Retrieval is to reflect how important a term is to a document. For topic segmentation, we perform intra-document term weighting, where the weight has to reflect how important a term is for a specific part of the show with respect to the rest of the show. In [8], TF-IDF weighting is used from a uniform partition of a show into N chunks, each chunk representing a document in the classical TF-IDF approach. In [9], we have proposed alternative chunks partitioning strategies that outperform the state-of-the-art uniform partition. Chunks can be obtained from another structural information (on TVBN, each anchor speaker turn can define the beginning of a new chunk) or in an iterative framework (topic segmentation obtained at a given iteration provides a set of segments that can be considered as chunks to reestimate TF-IDF weights for the next iteration). Whatever the chunk partitioning strategy, a term t in a breath group x will have a weight $w(c(x),t)$, depending on the chunk $c(x)$ it belongs to:

$$w(c(x),t) = TF(c(x),t).\ IDF(t)$$

$TF(c(x),t)$ is the frequency of term t in chunk $c(x)$
$IDF(t)=log(N/n_t)$ where n_t is the number of chunks containing term t and N is the total number of chunks.

2.1.2. Similarity computation and boundary detection

At the starting time of each breath group (BG) j, a similarity is computed between the bloc b_j of K BGs ending just before j, and the bloc b_{j+1} of K BGs beginning at this BG. A bloc b_j is represented by the vector V_j, whose coordinate for term t is $v(b_j,t)$ defined as:

$$v(b_j, t) = \sum_{x \in b_j} \left(f_{x,t}.\, w(c(x), t)\right)$$

The cohesion value for each BG j is the cosine similarity measure between blocs b_j and b_{j+1}:

$$cohesion(j)=Cosine(V_j,V_{j+1})$$

The cohesion values associated to each breath group can be plotted as a cohesion curve from which boundaries can be detected as low values, or values corresponding to high valley depth. We use a combination of cohesion values, and valley depth of the cohesion curve, to score each potential boundary:

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

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

Rather than simply applying a threshold to select high values of scores, a recursive process is applied to detect local maxima on the curve [9].

2.2. Speaker-cohesion based segmentation

When speaker segmentation and clustering are performed on the document, for each breath group, a label of the speaker who uttered the breath group is available. It can be seen as a "transcription" of the document in the speaker space, instead of lexical space. Thus, the previously proposed lexical-cohesion based segmentation can be translated into the speaker space, where all the terms of a given breath group are replaced with the label of the speaker s.

In the speaker space, the bloc b_j is represented by the vector V_j, whose coordinate for speaker s is:

$$v(b_j, s) = \sum_{x \in b_j} \left(f_{x,s}.\, w(c(x), s)\right)$$

In order to reflect the speaker's contribution in terms of uttered words, we define its vector coefficient as follows:

$$f_{x,s} = 1 + nb_x$$

where nb_x is the number of terms occurrences in x.

As the preprocessing step filters the uttered word sequence (function words removal and low confidence words removal) it can happen that a breath group has no remaining valid terms. Hence systematically adding 1 to the number of terms avoids non zero coefficients.

Once the vectorial representation is transposed into the speaker space, the segmentation algorithm remains unchanged (chunks partitioning for TF-IDF weighting, similarity computation, boundary selection).

3. FUSION OF SPEAKER AND LEXICAL INFORMATION

3.1. Fusion in the modeling space

In the fusion scheme we proposed in [11], a unified modeling space is adopted, that merges both lexical and speaker spaces, in a new representation space we name "speech". To do so, for each bloc, we concatenate the

lexical vector and the speaker vector in a single so-called speech vector. In order to have a balanced scale for coefficient $f_{x,t}$ and $f_{x,s}$ we had to normalize $f_{x,s}$ coefficients by the maximum number of terms in a breath group. Otherwise, the contribution of speaker coefficients in the overall vector space is too important.

$$f_{x,s} = \frac{1 + nb_x}{1 + max}$$

The rest of the algorithm remains unchanged and applies on the speech cohesion curve. In the experimental section, we refer to this early fusion approach as `Speech`.

3.2. Sequential fusion: co-segmentation

One key point in the baseline topic segmentation is the TF-IDF weighting, which depends on the definition of the chunks. In [9], we have made contrastive experiments to show the crucial importance of an appropriate weighting, and we showed that if the chunks used for TF-IDF weighting were perfectly chosen (through Oracle experiments, where the chunks were equal to the topic segments), the topic segmentation performances would be very significantly improved.

In the previously presented approaches, whatever the representation space (Lexical, Speaker or Speech) we adopt the iterative term weighting approach where chunk partitioning is initialized by a uniform chunk partition and the result of topic segmentation at a given iteration provides chunks for the next iteration. A stopping criterion [9] allows the process to be interrupted when there is no significant change in segmentation results with respect to the previous iteration.

Here, we propose that the segments obtained by one algorithm (either lexical-based or speaker-based) are used as chunks for the TF-IDF weighting of the other algorithm. It is expected to get better results, for each algorithm, than when staying in the same representation space for each iteration.

It can be seen as a sort of "co-training" approach, even though there is no "training" phase, strictly speaking. Thus, we propose to name it "co-segmentation". We can iterate this co-segmentation between speaker-based and lexical-based segmentations.

In this framework, only the very first segmentation (either speaker-based or lexical-based) initially uses the uniform chunk partition. Depending on which segmentation is performed first, variants are possible: we name CoSeg(A,B) the co-segmentation process between segmentation A and segmentation B, when the very first segmentation performed with uniform chunk is A. A and B can be `Lexical` (Lex), `Speaker` (Spk) or `Speech` (Speech) spaces.

Figure 1 illustrate the co-segmentation process CoSeg(Spk,Lex), where the uniform chunks hyp0 constitute the input chunk partition hypLex0 for the first segmentation step in the Speaker space.

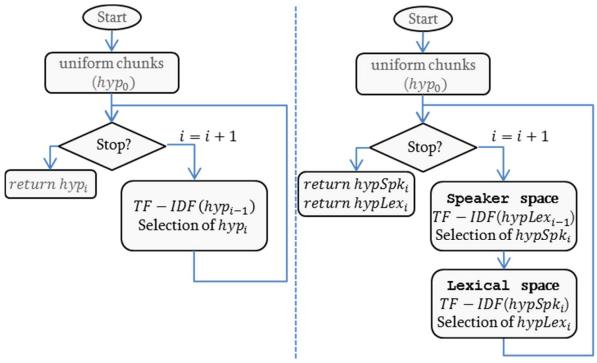

Figure 1: Iterative segmentation (left side) and iterative co-segmentation (right side, example of Coseg(Spk,Lex))

4. EXPERIMENTS

4.1. Corpora

Experiments are run on 2 distinct corpora, one from TV and one from the Radio.

4.1.1. TV Broadcast News

This set is composed of 70 different French TV Broadcast News shows, broadcasted within the same week in February 2014, from 7 different channels. The editorial policy may be very different from one news show to the other, with a varying number of anchorman (none, one or two), and different durations (from 10 minutes to about 1 hour). Manual topic segmentation was done so as to obtained reference topic segments. On the overall, the TV news set contains 856 reference topic segments, with 209 short (i.e. shorter than 30s) segments (average duration: 20.3s) and 647 long segments (average duration: 139.2s).

4.1.2. Radio talk-show

This set is composed of 50 episodes of a French radio news talk-show (#laTAC[1]) devoted to scientific topics and broadcasted in 2013. These shows are composed of a series of discussions on different topics, with one anchorman, and several journalists. For each topic, a given journalist is the main speaker, but other journalists can interfere, and the same journalist can be the main speaker for distinct topics. Each show lasts on average 12 minutes, and contains on average 7,9 different segments with average duration 96s.

On the overall, this set contains 396 reference topic segments among which 52 segments have a duration of less than 30 seconds.

4.1.3. Pre-processing

Automatic Speech transcription and pause detection is performed by the Vocapia speech to text engine, based on the Limsi technology [13]. Manual transcriptions are not available for these corpora, but on an equivalent corpus of

[1] http://www.franceinter.fr/emission-latac

Broadcast News [11], the WER is 16.1%. Speaker Diarization is performed by [14]. It is a sequential system using firstly Bayesian Information Criterion and then Cross-likelihood Criterion. On a similar corpus of conversational broadcast [11], the Diarization Error Rate is 10.3%. Thus, for each transcribed word, the label of the speaker who uttered this word is available.

4.1.4. Evaluation Metrics

Topic segmentation is evaluated as a boundary detection task. Performances are measured in terms of precision and recall by comparing time stamps for hypothesized and reference boundaries. As frequently found in the literature, an interval of 10s is tolerated around each reference boundary. Finally, F-measure is computed as the harmonic mean of precision and recall.

4.2. Results and discussion

Table 1 presents the F-measure results obtained for speaker and lexical based segmentation, and for different fusion approaches, along with the so-called Oracle performances, which are the performances obtained when the chunks used to estimate the TF-IDF weigthing are the reference topic segments.

Segmentation space	chunks	#laTAC	TVBN
Lexical	uniform init	50.2	51.2
	Oracle	*68.5*	*65.5*
	CoSeg(Spk,Lex)	56.1	54.7
	CoSeg(Lex,Spk)	55.8	54.7
Speaker	uniform init	57.9	39.6
	Oracle	*66.4*	*42.9*
	CoSeg(Spk,Lex)	62.2	39.6
	CoSeg(Lex,Spk)	62.1	40.8
Speech	uniform init	64.2	55.3
	oracle	*82.1*	*67.5*
	CoSeg(Speech,Lex)	63.1	**56.2**
	CoSeg(Lex,Speech)	60.7	55.0
	CoSeg(Speech,Spk)	66.2	51.6
	CoSeg(Spk,Speech)	**66.3**	52.4

Table2: F-measure results for topic boundary detection, for different representation spaces and chunks partition

First, concerning the baseline lexical-based segmentation, it can be noticed that the performances are equivalent across corpora, whereas the speaker-based algorithm behaves differently from one corpus to another. Indeed, the fact that the speaker distribution may be correlated to the topics is very dependent of the structure of the show. For the radio corpus #laTAC, the speaker structure is strongly related to the topic segmentation, whereas it is not the case for TVBN. On the contrary, the lexical distribution is always related to the topics, whatever the structure of the document. These remarks are confirmed by the behavior of lexical and speaker-based segmentation with oracle chunks. Results obtained with oracle chunks also confirm the importance of an appropriate weighting TF-IDF in the segmentation.

Concerning the fusion in Speech representation space, it can be observed that this fusion is always effective, giving better results than any of the segmentations using only speaker or lexical cohesion. The performance obtained with early fusion and oracle chunks confirms the potential of this fusion scheme.

When it comes to the "co-segmentation" approaches, between `Lexical` and `Speaker` spaces, we can observe first that the assumption that one segmentation make benefits from the chunks obtained with the other segmentation is proved. The performances obtained with lexical-based segmentation in the co-segmentation frameworks are better than those obtained initially: for "#laTAC", the F-measure increases from 50.2 % up to 56.1% (depending on the co-segmentation framework), and for TVBN, the F-measure increases from 51.2% to 54.7%. Same conclusions can be drawn for the speaker-based segmentation.

None of these co-segmentation frameworks between speaker and lexical segmentation outperforms the early fusion (`Speech`) between speaker and lexical spaces. When co-segmentation is performed between `Speech` and `Lexical` or `Speaker` spaces, conclusions are different from one corpus to another. The co-segmentation between `Speech` and `Speaker` is useful for #laTAC corpus, where we obtained the best result for this corpus (66.3%), and not for TVBN corpus, whereas the co-segmentation between `Speech` and `Lexical` is harmful for #laTAC corpus and useful for TVBN, where we obtained the best results for this corpus (56.2%).

Hence, it can be observed that the early fusion in the representation space is always useful, whatever the relative behavior of speaker and lexical spaces, and the co-segmentation, using `Speech` space, is beneficial when the other space is the best one for the corpus (`Lexical` for TVBN and `Speaker` for #laTAC).

CONCLUSION

This paper explores the ability of speaker-based information to detect topic boundaries, and different fusion frameworks between speaker and lexical information for topic segmentation. Experiments on 2 distinct corpora show that whereas the lexical information provides stable performances across corpora, the efficiency of speaker information for topic segmentation is closely related to the structure of the content. However, either with the proposed co-segmentation approach, or with the previously proposed early fusion approach, significant improvement in topic boundary detection is observed when including speaker information.

REFERENCES

[1] Hearst, M., "Textiling: Segmenting text info multiparagraph subtopic passages," Computational Linguistics, vol. 23, no. 1, pp. 33–64, 1997.

[2] Galley, M., McKeown, K., Fosler-Lussier, E. and Jing, H., "Discourse Segmentation of Multi-party Conversation", Proc. ACL'03, Sapporo, Japan, 2003.

[3] Xie, L., Yang, Y., Liu, Z.Q., Feng, W., Liu, Z. "Integrating Acoustic and Lexical Features in Topic Segmentation of Chinese Broadcast News Using Maximum Entropy Approach," in Proc ICALIP, Shanghai, pp. 407- 413 ,2010.

[4] Hirschberg, Julia, and Christine H. Nakatani. "Acoustic indicators of topic segmentation." *ICSLP*. 1998.

[5] Hakkani-T, Dilek, Andreas Stolcke, and Elizabeth Shriberg. "Integrating prosodic and lexical cues for automatic topic segmentation." *Computational linguistics* 27.1 (2001): 31-57.

[6] Wang, X., Xie, L., Lu, M., Ma, B., Chng, E.S., Li, H., "Broadcast News Story Segmentation Using Conditional Random Fields and Multimodal Features", IEICE Trans. Inf & Syst, Vol E95-D, No 5, May 2012.

[7] Choi F. Y. Y., "Advances in Domain Independent Linear Text Segmentation," in *Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, USA, 2000.

[8] Malioutov, I., Barzilay, R., "Minimum cut model for spoken lecture segmentation," in Proc. ACL, Sydney, pp. 25-32, 2006.

[9] Bouchekif, A., Damnati, G. and Charlet, D., "Intra-Content Term Weighting for Topic Segmentation", *Proc. ICASSP'14*, Florence, Italy, 2014.

[10] Dumont, E., Quénot, G., "Automatic Story Segmentation for TV News Video Using Multiple Modalities" International Journal of Digital Multimedia Broadcasting, vol. 2012, 2012.

[11] Bouchekif, A., Damnati, G. and Charlet, D., "Speech Cohesion for Topic Segmentation of Spoken Contents", *Proc. Interspeech'14*, Singapore, 2014.

[12] Tur, G., Hakkani-Tür, D., & Schapire, R. E. "Combining active and semi-supervised learning for spoken language understanding". *Speech Communication*, *45*(2), 171-186, 2005.

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

[14] Charlet D., Barras, C. and Lienard J., "Impact of Overlapping Speech Detection on Speaker Diarization for Broadcast News and Debates", *Proc. ICASSP'13*, Vancouver, Canada, 2013.