TITLE ASSIGNMENT FOR AUTOMATIC TOPIC SEGMENTS IN TV BROADCAST NEWS

Abdessalam Bouchekif [1,2] *Géraldine Damnati* [1] *Delphine Charlet* [1] *Nathalie Camelin* [2] *Yannick Estève* [2]

[1] Orange Labs, 2 avenue Pierre Marzin 22300, Lannion, France.
[2] Laboratoire d'Informatique de l'Universite du Maine, LIUM - France.

ABSTRACT

This paper addresses the task of assigning a title to topic segments automatically extracted from TV Broadcast News video recordings. We propose to associate a topic segment with the title of a newspaper article collected on the web at the same date. The task implies pairing newspaper articles and topic segments by maximising a given similarity measure. This approach raises several issues, such as the selection of candidate newspaper articles, the vectorial representation of both the segment and the articles, the choice of a suitable similarity measure, and the robustness to automatic segmentation errors. Experiments were conducted on various French TV Broadcast News shows recorded during one week, in conjunction with text articles collected through the Google News homepage at the same period. We introduce a full evaluation framework allowing the measurement of the quality of topic segment retrieval, topic title assignment and also joint retrieval and titling. The approach yields good titling performance and reveals to be robust to automatic segmentation.

***Index Terms*—** Topic segmentation, title assignation, term weighting, Okapi, similarity measures.

1. INTRODUCTION

For the last decade, web technologies have made available to the public huge amount of information and making relevant information fastly and easily accessible is a crucial issue. This is particularly true for very recent information coming from news companies, TV channels, newspaper websites. Some web companies automatically collect such news from many sources, and process them in order to present information in a structured way around topic [1]. TV Broadcast News (TVBN) videos accessible on the web usually contain the entire show, addressing several subjects. Being able to segment a show into topically coherent segments offers the possibility to navigate more easily through the show. Additionally, providing a title to each segment offers the possibility for a user to have a rapid access to a subject that he is interested in.

Topic segmentation (TS) [2, 3, 4] is the task which consists in splitting a video into thematically homogeneous fragments, each fragment addressing only one subject. In this study, we address the complementary task that consists in giving a title, when it is possible, to each topic segment extracted from TVBN shows. The video itself can contain relevant information for this titling task such as oral presentation of the main topics at the very beginning of the show, title captions that could be extracted by video OCR techniques. However, the former is not exhaustive as only the main topics are introduced and the latter heavily depends on editorial rules of each channels, as text captions are not always present. In order to have a generic approach, independent to any editorial policies, and to produce well-formed titles, we propose to use external data. Namely, we propose to compute pairings between topic segments extracted from daily videos and newspaper articles collected on the web at the same date. The title attributed to a topic segment is the title of its paired newspaper article, if any. The difficulty is that a fine-grained pairing stategy must be developped in order to have a precise title (it is not acceptable to propose the title of any article generally speaking about football if the topic segment is about a particular game of a particular team). However, in addition to giving a precise "well-written" title to video topic segments, this approach has the advantage of offering an interesting possibility to directly point relevant newspaper articles related to a segment. Eventually, this approach can also yield to News aggregation services where newspaper articles can be aggregated with excerpts of TVBN shows.

In the literature, research works on TVBN titling are very rare, while such works are more frequent for textual data. When associated to TS the labeling task is usually considered as a supervised classification task, where titles are searched from a predefined set of labels [5].There are more studies in the domain of topic models labeling where titles are searched for on the basis of a list of Top N words selected from the topic modeling process. [6] proposes a method to select specific words from Wikipedia article titles that match these Top N words in order to label topic models estimated on text documents, like the LDC Gigaword corpus. Apart from working on topic models rather than on topic segments, another difference with our approach is related to the evaluation protocol: they propose a list of ranked candidates that are *a posteriori* validated by annotators. In [7], terminology and keyphrases are automatically extracted from text segmentation in order to browse collections of documents, while in [8] topic keyphrases are extracted from Twitter. Recently, [9] presented an approach to segment and label asynchronous human-human text conversations such as emails and blog comments. In their case, target labels are short descriptions of the conversation and they explore generative methods to automatically derive them.

Our TS algorithm is briefly presented in section 2. Section 3 presents the topic titling approach. In section 4, we introduce new metrics for the segmentation and titling tasks as well as for the joint process. Finally, section 5 reports the details of our databases and experimental results.

2. TOPIC SEGMENTATION

Our topic segmentation (TS) algorithm is based on a cohesion curve computed between adjacent blocks, using a sliding window along the show. The unitary element is the breath group (BG), *i.e.* sequences of words between two pauses. Each BG boundary can be a potential topic segment boundary. A bloc then corresponds to a window of a given number of BGs. Each BG has a vectorial representation formalizing a *bag-of-words* approach. Cohesion is computed thanks

This work was partially funded by the European Commission through the EUMSSI project, under the contract number 611057, in the framework of the FP7-ICT-2013-10 call.

to the cosine similarity metric between vectorial representations of blocks. Then a recursive process detects local minima on the cohesion curve: low cohesion values mean that there are few terms in common between adjacent blocks, and the boundary between these blocks is likely to be a topic boundary. Then a final validation step is applied that confirms or rejects an hypothesized boundary by computing the cohesion between the corresponding adjacent hypothesized segments.

This algorithm is derived from the original *TextTiling* algorithm [10] principle but beyond the sliding-window mechanism, we have introduced in previous work several original modifications in the algorithm itself and more specifically in the vectorial representation of data. Hence, even though the grounding approach is simpler than other TS approaches, our algorithm yields good performances on TVBN data (see *eg.* [4] for a more detailed state of the art on TS).

The main original properties of our algorithm are related to the chosen space for vectorial representation: In [4] we have introduced the notion of *speech cohesion*, extending the lexical vectorial representation to speaker information available from speaker diarization. A potential boundary is effective if the joint distribution of terms and speakers is different enough from one side of the boundary to the other. Additionally, in [11] we have proposed to increase the relevance of lexical cohesion thanks to semantic relations between terms rather than simply considering term repetition. The semantic relation between two terms is given by the Cosine similarity between their word embeddings. Embeddings are estimated with the *word2vec* toolkit [12] on a diachronic corpus extracted from Google News on the same day of the TVBN show to be segmented.

3. TOPIC TITLING

The underlying idea of our approach is to compute a similarity measure between a topic segment and newspaper articles in order to assign to the segment the title of the article that maximizes a similarity measure, provided that this measure is above a given threshold, or to decide that the segment cannot be assigned any title otherwise. Actually, it can happen that some subjects addressed in TVBNs may not be addressed in newspaper articles: TVBN editors sometimes propose reports on subjects which are not necessarily relayed by printed and digital media. From an applicative point of view, it is important that our system doesn't propose any title for such segments (other approaches should be explored such as extractive or generative titling). This raises several issues, such as the selection of candidate articles, consistent vectorial representation for segments and articles and the choice of a suitable similarity measure.

3.1. Candidates selection

Candidate titles are those appearing on the *Google News* homepage during the day when the segment is broadcasted. This web page is constantly updated and gathers many articles from various press website. Articles are clustered and presented by News topic. For each topic cluster, one article is highlighted as the main article of the topic and other ones are shown as related. In order not to discard too many articles, we have chosen to download the web page every hour. Only the article content itself is relevant while the remaining (*e.g.* navigation menu, reader comments, pictures, *etc.*) is uninformative. We use the Boilerpipe [13] library that provides accurate algorithms to detect and remove all the surplus of a web page.

3.2. Vectorial representation

Similarity is computed between spoken segments and written texts. Due to their nature, the language style may vary between these documents and automatic transcriptions may contain errors. Hence it is important to choose a robust representation with efficient preprocessing for both written and spoken documents. Topic segments and articles are represented by a list of words, sorted by their order of relevance. In order to extract the list of words, a pre-processing step is applied thanks to the lia_tagg software [14]. Only nouns, adjectives and non-auxiliary verbs are retained, in their lemmatized form. As for TVBN topic segments, words that have a low confidence measure are also discarded.

For relevance estimation, we use $\text{TF} - \text{IDF}_{BM25}$ weighting scheme (also named $Okapi$[15]). $Okapi$ computation is performed by considering each topic segment of a given TVBN show as a document. Conversely, the relevance score of a term in an article is computed relatively to all articles collected on the same day. Terms with a negative $Okapi$ score (in particular for terms occurring in more than half of the documents) are discarded from the list. Positive scores can be greater than one: in order to alleviate scaling issues, we systematically normalize positive scores by the maximum value in each document. Additionally, in order to focus on relevant terms, terms with a normalized score under 0.25 are discarded. As a result, each topic segment and each article is represented by a list of terms, with a relevance score between 0.25 and 1.0.

3.3. Similarity measure

Similarity measures between textual documents have been widely explored and exploited for various applicative purposes. *Cosine* and *Jaccard* have proven to be efficient for several different tasks. [16] studied the impact of similarity measures for similar words clustering and introduced the *LIN* measure that was further studied in [17] for automatic thesaurus extraction and compared with a variant of extended Jaccard similarity. In this study, we compare similarity measures described in table 1, where S is the list of terms used in speech segment S and A is the list of terms used in article A, and $w^S(t)$ (resp. $w^A(t)$) is the weight of term t in segment S (resp. A).

Set_JACCARD	$\frac{\lvert S \cap A \rvert}{\lvert S \cup A \rvert}$
Extended_JACCARD	$\frac{\sum_{t \in S \cap A} w^S(t) * w^A(t)}{\lVert W^S \rVert_2^2 + \lVert W^A \rVert_2^2 - \sum_{t \in S \cap A} w^S(t) * w^A(t)}$
LIN	$\frac{\sum_{t \in S \cap A} w^S(t) + w^A(t)}{\sum_{t \in S \cup A} w^S(t) + w^A(t)}$
Cosine	$\frac{\sum_{t \in S \cap A} w^S(t) * w^A(t)}{\lVert W^S \rVert_2 * \lVert W^A \rVert_2}$

Table 1. Similarity measures

4. EVALUATION METRICS

4.1. Topic Segmentation evaluation

Several metrics have been proposed and compared in the literature. Considering topic segmentation as a topic boundary retrieval task naturally leads to the use of boundary detection precision, recall and F-measure. However beyond boundary detection, we need to use a metric that will be relevant for evaluating topic segmentation in the context of our final applicative task which is topic segmentation *and* titling; hence we need a metric that reflects the quality of segments. *pk* [18] and *windowdiff* [19] have been widely used but evaluate the overall segmentation quality along the complete show. What's more, the interpretation of the obtained numerical values is not straightforward. In [4] we have proposed to evaluate topic segmentation from the segment retrieval point of view, defining a quality measure for each reference segment.

Let S be a reference segment, the evaluation process searches for the test segment $H(S)$ which has the maximal temporal coverage of S. $H(S)$ can be seen as the system's response to the task of retrieving S. To measure the quality of the system's response for this specific reference segment S, we define the coverage $Cov_{S \to H(S)}$ as the percentage of duration of S which overlaps with $H(S)$. It can be seen as the *temporal recall* of the reference segment. Conversely, we define the coverage $Cov_{S \leftarrow H(S)}$ as the percentage of duration of the test segment which overlaps with S. It can be seen as the *temporal precision* of the system's response. Finally, we define the harmonic coverage $Cov_{S \leftrightarrow H(S)}$ as the harmonic mean of $Cov_{S \to H(S)}$ and $Cov_{S \leftarrow H(S)}$. This harmonic coverage can be seen as the *retrieval quality* for the reference segment.

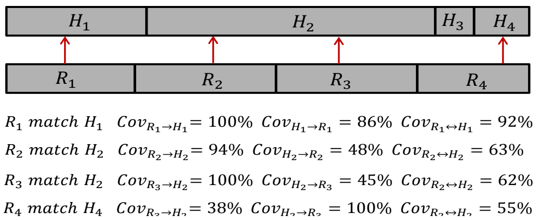

Fig. 1. Example of harmonic coverage computation

Depending on the minimum retrieval quality γ considered acceptable, a reference segment S is considered as correctly retrieved if $Cov_{S \leftrightarrow H(S)} \geq \gamma$ and incorrectly retrieved if $Cov_{S \leftrightarrow H(S)} < \gamma$. In the example presented in figure 1, if we set the required minimum harmonic coverage γ to 85%, only one segment (R_1) is correctly retrieved. Let R be the set of reference segments and $SegErr_\gamma$ the set of incorrectly retrieved segments, we can define the Segmentation Error Rate ($SegER_\gamma$), for a given level of required retrieval quality γ, as the rate of incorrectly retrieved segments:

$$SegER_\gamma = \frac{\#SegErr_\gamma}{\#R} \tag{1}$$

It is also possible to derive an equivalent metric, weighted by segment duration: $dSegErr_\gamma$ is obtained by summing up the duration of erroneous segments and dividing by the total duration of the show. It allows expressing the proportion of time that is correctly and incorrectly processed by the topic segmentation algorithm.

4.2. Topic titling evaluation on manual segmentation

The protocol for manually annotating the ground truth for the titling task was defined as follows. Given a reference topic segment, and the set of articles' titles collected during the same day, the annotation process consists in specifying if (i) a title is *suitable* for the segment, if (ii) a title could be suitable for the segment but appears to be too specific or too generic or if (iii) it does not reflect the segment. All the titles that have been annotated as *suitable* for the segment S are gathered in the set $Accept(S)$. Titles of the second and third category are considered as non-acceptable in this study. Reference topic segments can belong to two sets. MATCH (M) contains segments that can be assigned at least one title (non empty $Accept(S)$). Conversely, NOMATCH ($\overline{M}$) contains segments that have no corresponding article in the candidate list.

The titling system's response is considered as correct when a segment S from M is assigned a title that belongs to $Accept(S)$ or when a segment S from $\overline{M}$ is not assigned any title. Errors can be of three types: substitution (Sub) when a segment S from M is assigned a title that doesn't belong to $Accept(S)$, false rejection (FR) when a segment S from M is not assigned any title and false alarms (FA) when a segment S from $\overline{M}$ category is assigned a title.

The Titling Error Rate (TER) is defined as the sum of all possible errors over the total amount of segments.

$$TER = \frac{\#Sub + \#FR + \#FA}{\#R} \tag{2}$$

4.3. Topic segmentation and titling evaluation

Our final objective is to be able to retrieve topic segments and to assign a title to automatic segments. Hence, a correct result is a segment correctly retrieved (i.e. for which $Cov_{S \leftrightarrow H(S)} \geq \gamma$) and for which the titling process provides a correct result. Indeed, titling evaluation can only be measured for automatic segment when there is no ambiguity of title reference for this segment. Hence, titling error rates are only measured on the subset of correctly retrieved segments resulting in Sub_γ, FR_γ and FA_γ. We define the overall segmentation and titling error rate ($STER_\gamma$), for a given level of required retrieval quality γ, as follows:

$$STER_\gamma = \frac{\#SegErr_\gamma + \#Sub_\gamma + \#FR_\gamma + \#FA_\gamma}{\#R} \tag{3}$$

Similarly to $dSegErr_\gamma$, we can define $dSTER_\gamma$ by weighting each error by the duration of the corresponding segment relatively to the total duration of the show.

5. EXPERIMENTS

5.1. Corpora description

From the 10^{th} to the 16^{th} of February 2014, 29 hours of BN have been recorded from 8 french TV channels (TF1, France2, France3, M6, Arte, D8, NT1, Euronews). The corpus is composed of 86 shows. Manual transcriptions of TVBN shows are not available. Automatic transcription and pause detection are performed by the *Vocapia* speech to text engine, based on the Limsi technology [20], achieving 16.1% of Word Error Rate on an equivalent Broadcast News corpus [4]. A reference topic boundary annotation has been manually performed. The definition of a topic can be subject to various interpretations. In our work, a topic subject is a piece of information that can be extracted and watched independently. If several consecutive subjects about sport are encountered, the sport section will be subdivided into several topic segments. This applicative choice makes both the segmentation and the titling task more difficult. Similarly to [21], the first and last topic segments of a show are discarded when they correspond to the titles presentation or the summary. Finally, 997 topic segments are considered in this paper. In the rest of this paper we make a distinction between short segments (duration $<$ 30s) and long segments (duration $\geq$ 30s).
During the same period, the Google News homepage has been downloaded every hour, with a focus on the main article of each News cluster. As a result, a database of $5.4k$ entries has been generated among which only $4.6k$ are unique articles (a same article can remain a main article during several hours). On average 660 articles are considered per day. As a result of the titling annotation process, 633 topic segments have been associated with a non-empty set $Accept(S)$. On average, non-empty $Accept(S)$ sets contain 10.7 articles. Table 2 provides details about the repartition of MATCH and NOMATCH segments among long and short segments. Interestingly, the proportion of NOMATCH is higher for long segments than for short segments. Hence, 86.7% of the NOMATCH segments are long segments. It is more likely that the channels who choose to present a particular topic that is not present in the Google News headlines, develop the subject in a longer segment.

	total	Avg Dur.	MATCH	NOMATCH
Long Seg.	761	131.4s	467	294
Short Seg.	236	20.4s	191	45
All Seg.	997	105.1s	658	339

Table 2. Available titles for long and short segments

5.2. Experimental Results

5.2.1. Topic segmentation

From a boundary detection point of view, our segmentation algorithm yields a F-measure of 75.8% decomposing into 73.6% recall and 78.0% precision. Despite a simple sliding-window based algorithm, these results are good results compared to other state-of the art approaches on similar data. This is due particularly to our vectorial representation of breath groups that include speaker information, the use of a semantic relation matrix in the cohesion computation, and also on the final boundary validation step which increases the precision. Beyond F-measure, as mentioned in section 4.1, we evaluate topic segmentation in its capacity to retrieve segments. Table 3 presents the $SegER$ for different retrieval quality thresholds γ.

As expected, the higher the required retrieval quality, the higher

$SegER_\gamma$ (%)	$\gamma = 80\%$	$\gamma = 85\%$	$\gamma = 90\%$
Long segments (761 seg.)	21.2	23.4	30.6
Short segments (236 seg.)	68.6	75.0	80.1
All (997 seg.)	32.4	35.6	42.3
All duration ($dSegER$)	22.2	24.2	30.7

Table 3. Segment retrieval evaluation

the segmentation error rate. Considering that 85% is an acceptable minimum retrieval quality, our system has an overall $SegER$ of 35.6%, but with performance for long segments (23.4%) being far better than for short segments (75.0%), which is a common issue for topic segmentation in general. If weighted by the duration of each segment, the overall $dSegER$ expressed in the last line is equal to 24.2% meaning that 75.8% of the total duration of the shows is correctly retrieved by the segmentation algorithm.

5.2.2. Titling evaluation from reference topic segmentation

In this section the titling approach is evaluated on manually segmented topic segments. Figure 2 presents the results of our titling approach for various similarity measures. The rejection strategy for segments who shouldn't be assigned any title consists in setting a threshold on the similarity value between a segment and an article. We present the results as ROC curves (the sum of Substitution Rate and FR Rate is plotted as a function of the FA rate) where this threshold is varying. The choice of a threshold has an impact on the balance between FA and FR rates.

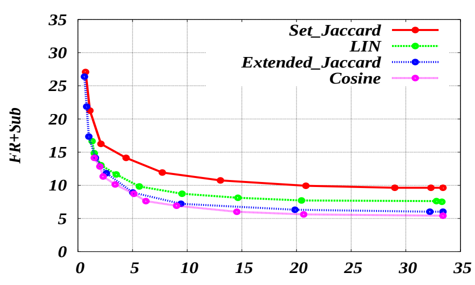

Fig. 2. TER decomposition for various similarity measures

The first observation is that using a weighted similarity metric is always better than the Set Jaccard metric. Only taking into account the number of terms in common between a segment and an article is not efficient enough. Among the metrics that involved weighted terms, Cosine similarity and the Extended Jaccard provide the best results. Cosine is considered for the rest of the paper.

Table 4 gives detailed performances for long and short segments when using the Cosine similarity metric. On the overall our titling system performs 11.8% error rate. Although the short segments have

Cosine	TER	Sub	FA	FR
Long Segments (761 seg.)	11.3	3.4	3.4	4.5
Short Segments (236 seg.)	13.6	5.5	5.1	3.0
All (997 seg.)	11.8	3.9	3.8	4.1

Table 4. Cosine for manually segmented long and short segments

a higher TER than the long ones, the difference is moderate. Hence, the proposed titling method is robust to short segments.

A deeper analysis of the remaining susbstitution errors, reveal some ambiguous cases. In particular, some events are very dynamic during a given day. For instance, during the winter Olympic Games, a newspaper article published in the morning can be outdated at midday. Some errors are also due to the fact that a similar event can be described in several areas, for instance bad weather conditions in Brittany (a French region) and in Great Britain. This is the limit of our bag of words representation for segments and articles, higher level semantic analysis should be applied to overcome these issues.

Correct Tiltle	Erroneous Title
Sotchi: France is waiting for its first medal	Martin Fourcade: first gold medal for France at Sotchi
Floods: three departments remain under high vigilance	Storm: 10.000 households without power in Brittany

Table 5. Examples of ambiguous titles

5.2.3. Titling evaluation from automatic topic segmentation

Table 6 illustrates the performances of the overall segmentation and titling process. The low amount of titling errors on automatic segments shows that the titling process is robust to automatic boundary detection.

$\gamma = 85\%$	$STER$	$SegErr$	Sub	FA	FR
Long (761 seg.)	34.0	23.4	3.8	1.3	5.3
Short (236 seg.)	77.5	75.0	0.8	1.3	0.4
All (997 seg.)	44.3	35.6	3.1	1.3	4.1

Table 6. Joint segmentation and titling error rate.

The overall $STER_\gamma$ is dominated by segmentation errors. Here again it is interesting to observe the performance on long and short segments. As a matter of fact the overall 44.3% $STER_{85}$ is only 34.0% when evaluated on long segments. When computed relatively to the duration of segment the overall $dSTER_{85}$ is equal to 34.2% meaning that the full system is able to correctly retrieve and title 65.8% of the duration of TVBN shows.

6. CONCLUSION

In this paper, we have proposed a fully automatic framework to segment and label topics in TVBN. After a topic segmentation step, the generated segments are assigned a title corresponding to an article collected from Google News during the same day. This matching process between a spoken document fragment and a written article relies on a similarity measure between vectorial representation of the spoken document and the article. A global evaluation framework to evaluate both segmentation and titling errors is proposed. A detailed analysis shows that the segmentation step remains the most error prone. But as long as the segmentation retrieval is good enough, the titling error rate is very low and robust to automatic TS boundary generation. As a perspective to this work, it is envisaged to combine the titling process and the segmentation process. The good titling performance observed on short segment suggest that it could be helpful to use the information of the titles that are similar to a block to improve short segments retrieval performance.

7. REFERENCES

[1] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 271–280.

[2] Lei Xie, Yulian Yang, Zhi-Qiang Liu, Wei Feng, and Zihan 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.

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

[4] Abdessalam Bouchekif, Géraldine Damnati, and Delphine Charlet, "Speech cohesion for topic segmentation of spoken contents," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1890–1894.

[5] Pei-Yun Hsueh and Johanna D Moore, "Automatic topic segmentation and labeling in multiparty dialogue," in *Spoken Language Technology Workshop, 2006. IEEE*. IEEE, 2006, pp. 98–101.

[6] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin, "Automatic labelling of topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 1536–1545.

[7] David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin, "Bayesian text segmentation for index term identification and keyphrase extraction.," in *COLING*. Citeseer, 2012, pp. 2077–2092.

[8] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*, pp. 338–349. Springer, 2011.

[9] Shafiq Rayhan Joty, Giuseppe Carenini, and Raymond T Ng, "Topic segmentation and labeling in asynchronous conversations," *Journal of Articial Intelligence Research 47 (2013)*, pp. 521–573, 2013.

[10] Marti A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, Mar. 1997.

[11] Abdessalam Bouchekif, Géraldine Damnati, Yannick Esteve, Delphine Charlet, and Nathalie Camelin, "Diachronic semantic cohesion for topic segmentation of tv broadcast news," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, September 06-10, 2015*, 2015, pp. 2932–2936.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[13] Christian Kohlschutter, "Boilerpipe," [Online]. Available: https://code.google.com/p/boilerpipe/.

[14] Frederic Bechet, "lia_tagg," [Online]. Available: http://pageperso.lif.univ-mrs.fr/f̃rederic.bechet/download.html.

[15] Karen Sparck Jones, Steve Walker, and Stephen E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 1," *Information Processing & Management*, vol. 36, no. 6, pp. 779–808, 2000.

[16] Dekang Lin, "Automatic retrieval and clustering of similar words," in *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, 1998, pp. 768–774.

[17] James R Curran and Marc Moens, "Improvements in automatic thesaurus extraction," in *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, 2002, pp. 59–66.

[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] Lev Pevzner and Marti Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Comput. Linguist.*, vol. 28, no. 1, pp. 19–36, Mar. 2002.

[20] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1, pp. 89–108, 2002.

[21] Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot, et al., "Improving ASR-based topic segmentation of tv programs with confidence measures and semantic relations," in *INTERSPEECH*, 2010, pp. 1365–1368.