Exploiting speech for automatic TV delinearization: From streams to cross-media semantic navigation

Guillaume Gravier[a], Camille Guinaudeau[b],
Gwénolé Lecorvé[a], Pascale Sébillot[a]

(a) IRISA UMR 6074 – CNRS & INSA Rennes
(b) INRIA Rennes – Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France

{*guillaume.gravier,camille.guinaudeau,gwenole.lecorve,pascale.sebillot*}*@irisa.fr*

Abstract

The gradual migration of television from broadcast diffusion to Internet diffusion offers countless possibilities for the generation of rich navigable contents. However, it also raises numerous scientific issues regarding delinearization of TV streams and content enrichment. In this paper, we study how speech can be used at different levels of the delinearization process, using automatic speech transcription and natural language processing (NLP) for the segmentation and characterization of TV programs and for the generation of semantic hyperlinks in videos. Transcript-based video delinearization requires natural language processing techniques robust to transcription peculiarities, such as transcription errors, and to domain and genre differences. We therefore propose to modify classical NLP techniques, initially designed for regular texts, to improve their robustness in the context of TV delinearization. We demonstrate that the modified NLP techniques can efficiently handle various types of TV material and be exploited for program description, for topic segmentation, and for the generation of semantic hyperlinks between multimedia contents. We illustrate the concept of cross-media semantic navigation with a description of our news navigation demonstrator presented during the NEM Summit 2009.

1 Introduction

Television is currently undergoing a deep mutation, gradually shifting from broadcast diffusion to Internet diffusion. This so-called TV-Internet convergence raises several issues with respect to future services and authoring tools, due to fundamental differences between the two diffusion modes. The most crucial difference lies in the fact that, by nature, broadcast diffusion is eminently linear while Internet diffusion is not, thus permitting features such as navigation, search and personalization.

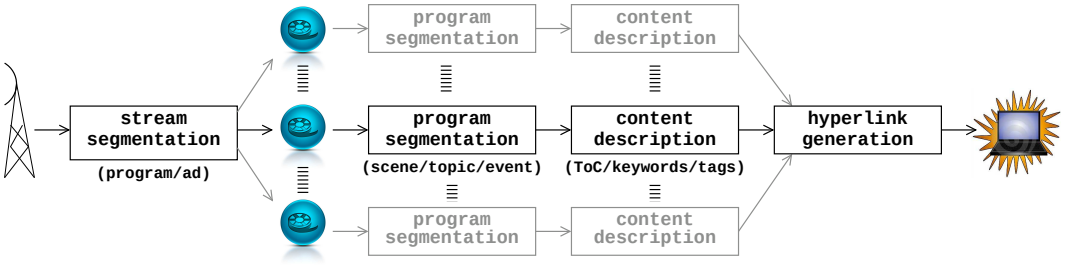


Figure 1: Schematic view of a typical delinearization process. The input stream is divided into programs which are in turn decomposed into segments (typically scenes, events, or topics). Each segment is further enriched with a description before creating links in between segments or between segments and external resources.

In particular, navigation by means of links between videos or, in a more general manner, between multimedia contents, is a crucial issue of Internet TV diffusion.

One of the challenges behind changing the diffusion mode is that of the *delinearization* of contents generated for stream diffusion. Delinearization consists in breaking a continuous video stream into basic elements—such as programs and scenes—for which a description is made available together with links to related contents. The various stages of a typical delinearization chain are illustrated in Fig. 1. Clearly delinearization of TV streams is already a fast growing trend with the increasing number of catch-up TV sites and video on demand portals. Even if one can anticipate that Internet diffusion will predominate in a near future, we firmly believe that the two diffusion modes will still coexist for long as they correspond to very different consumption habits. Linear or *streaming* diffusion, in which a continuous TV stream is accessible, is passive while a "search and browse" enabled diffusion mode requires action from viewers. Such cohabitation is already witnessed with all major channels providing catch-up videos on the Internet for their key programs.

The growth and the impact of non-linear Internet TV diffusion however remain limited for several reasons. Apart from strategical and political reasons[1], several technical reasons prevail. Firstly, the amount of content available is limited as repurposing TV contents for Internet diffusion is a costly process. The delinearization overhead is particularly cumbersome for services which require a semantic description of contents. Secondly, most Internet diffusion sites offer poor search features and lack browsing capabilities enabling users to navigate between contents. Indeed, browsing is often limited to a suggestion of videos sharing some tags which poorly describe the content. The main reason for this fact is, again, that obtaining an exploitable semantic description of some content is a difficult task. In brief, Internet browser-enabled diffusion of TV contents is mostly limited by the lack of a detailed semantic description of TV contents for enhanced search and navigation capabilities. There is therefore a strong need for automatic delinearization tools that

[1]Currently, commercial policies of broadcasters imply that navigation is almost always limited to contents within a broadcaster's site to prevent users from browsing away. However, we believe that in the future, these limitations will vanish with the emergence of video portals independent from the broadcasting companies.

break streams into their constituents (programs, events, topics, etc.) and generate indexes and links for all constituents. Breaking video streams into programs have been addressed on several occasions [1, 2, 3, 4] but does not account for a semantic interpretation. Breaking programs into their constituents have received a lot of attention for specific genres such as sports [5, 6] and broadcast news [7, 8, 9]. Most methods are nonetheless either highly domain and genre specific or limited in their semantic content description. Moreover, regardless of the segmentation step, automatically enriching video contents with semantic links, eventually across modalities, have seldom been attempted [10, 11].

Spoken material embedded in videos, accessible by means of automatic speech recognition (ASR), is a key feature to semantic description of video contents. However, spoken language is seldom exploited in the delinearization process, in particular for TV streams containing various types of programs. The main reason for this fact is that, apart for specific genres such as broadcast news [8, 12, 9, 13], natural language processing (NLP) and information retrieval (IR) techniques originally designed for regular texts[2] are not robust enough and fail to perform sufficiently well on automatic transcripts—mostly because of transcription errors and because of the lack of sentence boundary markers—and/or are highly dependent on a particular domain. Indeed, depending on many factors such as recording conditions or speaking style, automatic speech recognition performance can drop drastically on some programs. Hence, the need for genre and domain independent spoken content analysis techniques robust to ASR peculiarities for TV stream delinearization

In this paper, we propose to adapt existing NLP and IR techniques to ASR transcripts, exploiting confidence measures and external knowledge such as semantic relations, to develop robust spoken content processing techniques at various stages of the delinearization chain. We show that this strategy is efficient in robustifying processing of noisy ASR transcripts and permits speech-based automatic delinearization of TV streams. In particular, the proposed robust spoken document processing techniques are used for content description across a wide variety of program genres and for efficient topic segmentation of news, reports and documentaries. We also propose an original method to create semantic hyperlinks across modalities, thus enabling navigational features in news videos. We finally illustrate how those techniques were used at the core of a news navigation system, delinearizing news shows for semantic cross-media navigation, demonstrated during the NEM Summit 2009.

The paper is organized as follows. We first present the speech transcription system used in this study, highlighting peculiarities of automatic transcripts. In Section 3, a bag-of-words description of TV programs is presented to automatically link programs with their synopses. In Section 4, a novel measure of lexical cohesion for topic segmentation is proposed and validated on news, reports and documentaries. Section 5 is dedicated to an original method for the automatic generation of links across modalities, using transcription as a pivot modality. The NEM Summit 2009 news navigation demonstration is described in Section 6. Section 7 concludes the

[2]By *regular* texts, we designate texts originally designed in their written form, for which structural elements such as casing, sentence boundary markers and eventually paragraphs are explicitly defined.

paper, providing future research directions towards better automatic delinearization technologies.

2 Transcription of spoken TV contents

The first step to speech-based processing of TV contents is their transcription by an automatic speech recognition engine. We recall here the general principles of speech recognition to highlight the peculiarities of automatic transcripts with respect to regular texts and the impact on NLP and IR techniques. In the second part, details on the ASR system used in this work are given.

2.1 Transcription principles

Most automatic speech recognition systems rely on statistical models of speech and language to find out the best transcription hypothesis, i.e., word sequence, given a (representation of the) signal y, according to

$$\widehat{w} = \arg\max_{w} p(y|w)\ P[w]\ \ . \tag{1}$$

Language models (LM), i.e., probability distributions over sequences of N words (N-gram models), are used to get the prior probability $P[w]$ of a word sequence w. Acoustic models, typically continuous density hidden Markov models (HMM) representing phones, are used to compute the probability of the acoustic material for a given word sequence, $p(y|w)$. The relation between words and acoustic models of phone-like units is provided by a pronunciation dictionary which lists the words recognizable by the ASR system, along with their corresponding pronunciations. Hence, ASR systems operate on a closed vocabulary whose typical size is between 60,000 and 100,000 words. Words out of the vocabulary (OOV) cannot be recognized as is and are therefore one cause of recognition errors, resulting in the correct word being replaced by one or several similarly sounding erroneous words. The vocabulary is usually chosen by selecting the most frequent words, eventually adding domain-specific words when necessary. However, named entities (proper names, locations, etc.) are often missing from a closed vocabulary, in particular in the case of domain independent applications such as ours.

Evaluating Eq. (1) over all possible word sequences of unknown length is costly in spite of efficient approximate beam search strategies [14, 15] and is usually performed over short utterances of 10 s to 30 s. Hence, prior to transcription, the stream is partitioned into short sentence-like segments which are processed independently one of another by the ASR system. Regions containing speech are first detected and each region is further broken into short utterances based on the detection of silences and breath intakes.

Clearly, ASR transcripts significantly differ from regular texts. First, recognition errors can strongly impact the grammatical structure and semantic meaning of the transcript. In particular, poor recording conditions, environmental noises—such as laughter and applause—, and spontaneity of speech are all factors that might

occur in TV contents and which drastically increase recognition errors. Second, unlike most texts, transcripts are unstructured, lacking sentence boundary markers and paragraphs. In some cases, transcripts are also case insensitive so as to limit the number of OOV words. These oddities might be detrimental to NLP where casing and punctuation marks are often considered as critical cues. However, ASR transcripts are more than just degraded texts. In particular, word hypotheses are accompanied by confidence measures indicating for each word an estimation of its correctness by the ASR system [16]. Using confidence measures for NLP and IR can help avoiding error-prone hard decisions from the ASR system and partially compensate for recognition errors. But this requires that standard NLP and IR algorithms be modified, as we propose in this paper.

2.2 The IRENE transcription system

In this paper, all TV programs were transcribed using our IRENE ASR system, originally developed for broadcast news transcription. IRENE implements a multiple pass strategy, progressively narrowing the set of candidate transcriptions—the search space—in order to use more complex models. In the final steps, a 4-gram LM over a vocabulary of 65,000 words is used with context-dependent phone models to generate a list of 1,000 transcription hypotheses. Morphosyntactic tagging, using a tagger specifically designed for ASR transcripts, is used in a post-processing stage to generate a final transcription with word posterior based confidence measures, combining the acoustic, language model and morphosyntactic scores [17]. Finally, part-of-speech tags are used for lemmatization and, unless otherwise specified, lemmas[3] are considered instead of words in this work.

The language model probabilities were estimated on 500 million words from French newspapers and interpolated with LM probabilities estimated over 2 million words corresponding to reference transcription of radio broadcast news shows. The system exhibits a word error rate (WER) of 16 % on the non accented news programs of the ESTER 2 evaluation campaign [18]. As far as TV contents are concerned, we estimated word error rates ranging from 15 % on news programs to more than 70 % on talk shows or movies.

3 Using speech as a program descriptor

The first step in TV delinearization is the stream segmentation step which usually consists in splitting the stream into programs and inter-programs (commercials, trailers/teasers, sponsorships or channel jingles). Several methods have been proposed to this end, exploiting information from an electronic program guide (EPG) to segment the video stream and label each of the resulting segments with the corresponding program name [2, 3, 19, 4]. Note that stream segmentation exploiting inter-program detection, as in [4], results in segments corresponding to a TV program, to a fraction of a program or, on some rare occasions, to several programs.

[3] A lemma is an arbitrary canonical form grouping all inflexions of a word in a grammatical category, e.g., the infinitive form for verbs, the masculine singular form for adjectives, etc.

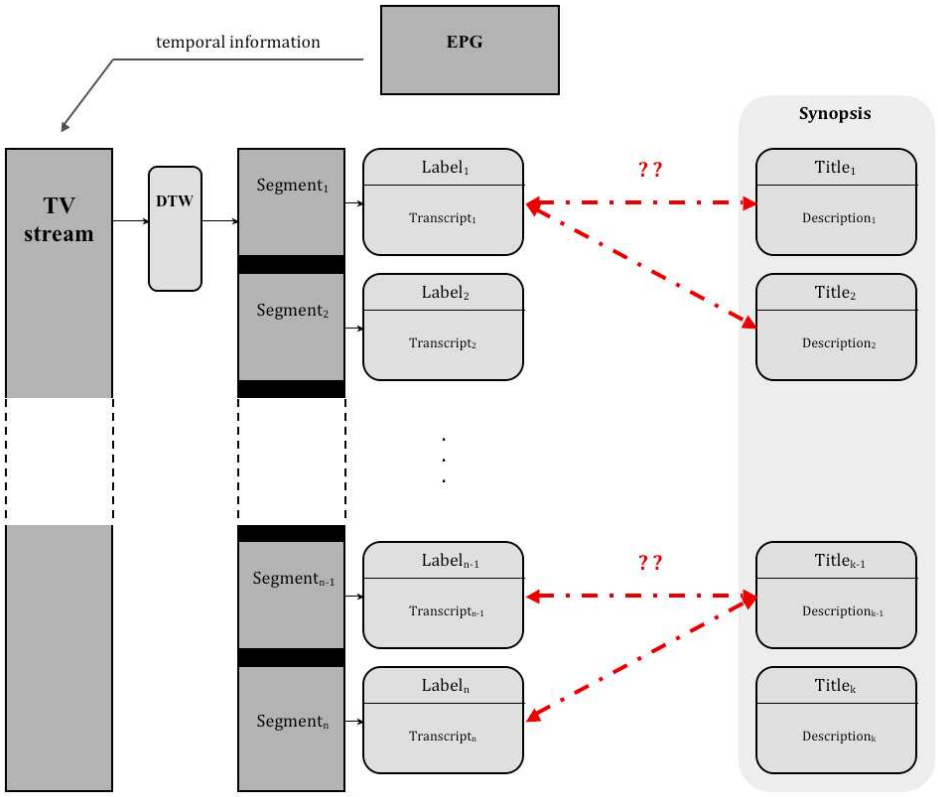


Figure 2: Principle of the speech-based validation of labels obtained from EPG alignment.

In all cases, aligning the video signal with the EPG relies on low-level audio and visual features along with time information and does not consider speech indexing and understanding to match program descriptions with video segments.

We investigate the capability of error-prone speech transcripts as a semantic description of arbitrary TV contents. We propose to adapt information retrieval techniques to associate each of the segments resulting from the stream segmentation step with short textual synopses describing programs. In the delinearization framework considered, the relations established between a segment's transcript and the synopses are used to validate, and eventually correct, labels resulting from the EPG-based stream segmentation. The entire process is depicted in Fig. 2.

We first describe how traditional information retrieval approaches are modified to create associations between synopses and a video segment based on its transcription. The use of these associations to label segments is rapidly discussed, the reader being referred to [20] for more details.

3.1 Pairwise comparison of transcripts and synopses

The entire process of associating synopses and transcripts relies on pairwise comparisons between a synopsis and a segment's transcript. We propose a technique for such a pairwise comparison, inspired from word-based textual information retrieval techniques. In order to deal with transcription errors and OOV words, some modifications of the traditional vector space model (VSM) indexing framework [21] are proposed: confidence measures are taken into account in the index term weights, and a phonetic-based document retrieval technique, enabling to retrieve in a transcript the proper nouns contained in a synopsis, is also considered.

3.1.1 Modified *tf-idf* criterion

In the vector space model, a document d—in our case, a transcript or a synopsis—is represented by a vector containing a score for each possible index term of a given vocabulary. In our case, the set of index terms is the set of lemmas corresponding to the vocabulary of the ASR system. The popular normalized *tf-idf* weight is often used as the score. Formally, term frequency (tf) for a lemma l is defined as

$$tf(l,d) = \frac{f(l,d)}{\max\limits_{x \in d} f(x,d)} \tag{2}$$

where $f(x,d)$ denotes the frequency of occurrence of x in d. Inverse document frequency (idf), estimated over a collection $\mathcal{C}$, is given by

$$idf(l,\mathcal{C}) = -\log \frac{|\{c \in \mathcal{C} : l \in c\}|}{|\mathcal{C}|} \tag{3}$$

where $|.|$ denotes the cardinality operator. The final *tf-idf* weight of l in d is then defined as

$$S_d(l) = \frac{tf(l,d) \times idf(l,\mathcal{C})}{\max\limits_{x \in d} tf(x,d) \times idf(x,\mathcal{C})} \ . \tag{4}$$

Following the same philosophy as in [22], the weights $S_d(l)$ are modified in the case of automatic transcripts so as to account for confidence measures, thus indirectly compensating for transcription errors. Confidence measures are used to bias the *tf-idf* weights according to

$$S'_d(l) = \left[\theta + (1-\theta)\, c_l\right] \times S_d(l) \ , \tag{5}$$

where c_l denotes the average word-level confidence over all occurrences of l in d. Eq. 5 simply states that words for which a low confidence is estimated by the ASR system will contribute less to the *tf-idf* weight than words with a high confidence measure. The parameter θ is used to smooth the impact of confidence measures. Indeed, confidence measures, which correspond to a self-estimation of the correctness of each word hypothesis by the ASR system, are not fully reliable. Therefore, θ, experimentally set to 0.25 in this work, prevents from fully discarding a word based on its sole confidence measure.

Given the vector of *tf-idf* weights for a synopsis and the vector of modified *tf-idf* weights for a segment's transcript, the pairwise distance between the two is given by the cosine measure between the two description vectors.

3.1.2 Phonetic association

Named entities, in particular proper names, require particular attention in the context of TV content description. Indeed, proper names are frequent in this context (e.g., characters' names in movies and series) and are often included in the synopses. However, proper names are likely to be OOV words that will therefore not appear in ASR transcripts. As a consequence, proper names are likely to jeopardize or, at least, to not contribute to the distance between a transcript and a synopsis when using the *tf-idf* weighted vector space model.

To skirt such problems, a phonetic measure of similarity is defined to phonetically search a transcript for proper names appearing in a synopsis. Each proper name in the synopsis is automatically converted into a string of phonemes. A segmental variant of the dynamic alignment algorithm is used to find in the phonetic output of an ASR transcript the sub-string of phonemes that best matches the proper name's phonetization. The normalized edit distance between the proper name's phoneme string and the best matching sub-string defines the similarity between the ASR transcript and the proper name in the synopsis. The final distance between the synopsis and the transcript is given by summing over all proper names occurring in the synopsis.

3.2 Validating the segmentation

We demonstrate on a practical task that the comparison techniques of Section 3.1 enable the use of ASR transcripts for genre independent characterization of TV segments in spite of potentially numerous transcription errors. The word and phonetic level pairwise distances are used to validate, and eventually modify, the label (i.e., the program name) attached to each segment as a result of the alignment of the stream with an EPG. This validation step is performed by associating a unique synopsis with each segment before checking whether the synopsis corresponds to the program name obtained from the EPG or not, as illustrated in Fig. 2. In case of mismatch, a decision is made to maintain, to change or to invalidate the segment's label, based on the scheduled and broadcasted start times. Associating a unique synopsis with each segment relies on shortlists of candidate segments for each synopsis. For a given synopsis, two shortlists of candidate segments are established, one based on the word-level distance as given using the modified *tf-idf* criterion, one based on the phonetic distance. Details on shortlist generation can be found in [20]. The synopsis associated with a given segment is the one with the highest association score among those synopses for which the shortlists contain the segment.

Results are reported on a subset of the 650 segments resulting from an automatic alignment of a continuous TV stream of 10 days with an EPG [3][4]. Coming from a

[4]In [3], a label error rate of about 65 % is reported, considering all segments.

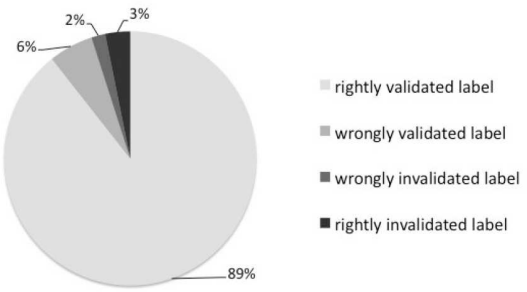


Figure 3: Results of the validation of the labels provided by the alignment of the stream with the EPG.

long continuous stream, segments include all genres of TV programs and transcripts exhibit word error rate ranging from 15 % to 70 % on the most difficult programs. Transcripts vary in length from 7 to 27,150 words, with an average of 2,643. A subset consisting of the 326 segments containing more than 600 words is considered in this work[5]. Around 250 synopses corresponding to the time period of the stream considered were taken from an on-line program guide, with descriptions varying both in length (average: 50 words) and precision (from a title to a precise description of the content). Finally, 63% of the program descriptions contain at least one proper noun with an average of 7.5 per description. The association of a synopsis with each segment exhibits a recall of 86 % with a precision of 63 %. Results on the validation of the labels from the EPG alignment step are reported in Fig. 3. The EPG labels are validated in 89 % of the cases. Corrections based on the synopsis' titles decrease the labeling error rate by 0.2 %, the number of correct changes made almost being equal to the number of erroneous ones. Erroneous corrections are always related to segmentation errors where the starting time of the segment does not correspond to any description.

In spite of the very limited gain incurred by the synopsis based label correction process, these results clearly demonstrate that the proposed lexical and phonetic pairwise distances enable to efficiently use automatic speech transcripts as a description of TV segments, for a wide range of program genres. However, the word-level description considered is a "bag of words" representation which conveys only limited semantics, probably partially explaining the robustness of the description to transcription errors. For programs with reasonable error rates between 15 % and 30 %, such as news, documentaries and reports, speech can be used for finer semantic analysis, provided adequate techniques are proposed to compensate for the peculiarities of automatic transcripts. In the following sections, we propose robust techniques for topic segmentation and link generation respectively, limiting ourselves to news, documentaries and reports.

[5]The reason for ignoring short segments is that they often have neither description in the EPG, nor related synopsis, thus making it impossible to evaluate the segment/synopsis association. Indeed, short segments mostly correspond to fillers inserted by broadcasters to adjust broadcasting schedules and to weather forecast programs.

4 Topic segmentation of TV programs

Segmentation of programs into topics is a crucial step to allow users to directly access parts of a show dealing with their topics of interest, or to navigate between the different parts of a show. Within the framework of TV delinearization, topic segmentation aims at splitting shows for which the notion of topic is relevant (e.g., broadcast news, reports and documentaries in this work) into segments dealing with a single topic, for example to further enrich those segments with hyperlinks. Note that such segments usually include the introduction and eventually the conclusion by the anchor speaker in addition to the development[6] itself and are therefore hardly detectable from low level audio and visual descriptions. Moreover, contrary to the TDT framework [23], no prior knowledge on topics of interest is provided so as to not depend on any particular domain and, in the context of arbitrary TV contents, segments can exhibit very different lengths.

Topic segmentation has been studied for years by the NLP community which developed methods dedicated to textual documents. Most methods rely on the notion of lexical cohesion, corresponding to lexical relations that exist within a text, and mainly enforced by word repetitions. Topic segmentation methods using this principle are based on an analysis of the distribution of words within the text: a topic change is detected when the vocabulary changes significantly [24, 25]. As an alternative to lexical cohesion, discourse markers, obtained from a preliminary learning process or provided by a human expert, can also be used to identify topic boundaries [26, 27]. But discourse markers are domain and genre dependent and sensitive to transcription errors while lexical cohesion does not depend on specific knowledge. However, lexical cohesion is also sensitive to transcription errors. We therefore propose to improve the lexical cohesion measure at the core of one of the best text segmentation method [28] to accomodate for confidence measures and to account for semantic relations other than the mere word repetitions (e.g., the semantic proximity between the words "car" and "drive") to compensate for the limited number of repetitions in certain genres. As we will argue in Section 4.2, the use of semantic relations serves a double purpose: better semantic description and increased robustness to transcription errors. However, such relations are often domain dependent and their use should not be detrimental to the segmentation of out of domain transcripts.

We rapidly describe the topic segmentation method of Utiyama and Isahara [28] which serves as a baseline in this work, emphasizing the probabilistic lexical cohesion measure on which this method is based. We extend this measure to successively account for confidence measures and semantic relations. Finally, experimental results on TV programs are presented in Section 4.3.

[6] By development, we refer to the actual report on the topic of interest. A typical situation is that of news programs where the anchor introduces the subject before handing over to a live report, the latter being eventually followed by a conclusion and/or a transition to the next news item. All these elements should be kept as a single topic segment.

4.1 Topic segmentation based on lexical cohesion

The topic segmentation method introduced by Utiyama and Isahara [28] for textual documents was chosen in the context of transcript-based TV program segmentation for two main reasons. It is currently one of the best performing method that makes no assumption on a particular domain (no discourse markers, no topic models, etc.). Moreover, contrary to many methods based on local measures of the lexical cohesion, the global criterion used in [28] makes it possible to account for the high variability in segment lengths.

The idea behind the topic segmentation algorithm is to search among all possible segmentations for the one that globally results in the most consistent segments with respect to the lexical cohesion criterion. This optimization problem is expressed in a probabilistic framework as finding out the most probable segmentation of a sequence of l basic units (lemmas or lemmatized sentences) $W = W_1^l$ according to

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}}\, P[W|S]\; P[S] \;\; . \tag{6}$$

In practice, assuming that $P[S_1^m] = n^{-m}$, with n the number of words in the text and m the number of segments, and relying on the traditional hypothesis of conditional independence of the observations[7], the search problem is given by

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} \Big(\ln(P[W_{a_i}^{b_i}|S_i]) - \alpha \ln(n) \Big) \;\; , \tag{7}$$

where α allows for a control of the average segment size and where $P[W_{a_i}^{b_i}|S_i]$ denotes the probability of the sequence of basic units corresponding to S_i.

In Eq. (7), lexical cohesion is considered by means of the probability terms $P[W_{a_i}^{b_i}|S_i]$ computed independently for each segment. As no prior model of the distribution of words for a given segment is available, generalized probabilities are used. Lexical cohesion is therefore measured as the ability of a unigram language model Δ_i, whose parameters are estimated from the words in S_i, to predict the words in S_i.

The language model Δ_i estimated on the words of S_i is a unigram language model over the set of words in the text (or transcript) to be segmented. The calculation of the language model of the segment S_i is formalized, for a Laplace smoothing, by

$$\Delta_i = \left\{ P_i(u) = \frac{C_i(u) + 1}{z_i},\; \forall\; u \in V_K \right\} \;\; , \tag{8}$$

where V_K is the vocabulary of the text, containing K different words, and where the count $C_i(u)$ denotes the number of occurrences of u in S_i. The probability distribution is smoothed by incrementing the count of each word by 1. The normalization term z_i ensures that Δ_i is a probability mass function and, in the particular case of Eq. (8), $z_i = K + n_i$ with n_i the number of word occurrences in S_i.

[7] Words within a segment are independent from words in the other segments given the segmentation.

Given the estimated language model, the lexical cohesion is measured as the generalized probability given by

$$\ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} \ln P[w_j^i; \Delta_i] \ , \tag{9}$$

where w_j^i denotes the j^{th} word in S_i. Intuitively, according to Eq. (9), lexically consistent segments exhibit higher lexical cohesion values than others as the generalized probability increases as the number of repetitions increases.

In this work, the basic units considered were utterances as given by the partitioning step of the ASR system, thus limiting possible topic boundaries to utterance boundaries. Moreover, lexical cohesion was computed on lemmas rather than words, discarding words other than nouns, adjectives and non modal verbs.

4.2 Improved lexical cohesion for spoken TV contents

As argued previously, confidence measures and semantic relations can be used as additional information to improve the generalized probability measure of lexical cohesion so as to be robust to transcription errors and to the absence of repetitions. We propose extensions of the lexical cohesion measure to account for such information.

4.2.1 Confidence measures

Confidence measures can straightforwardly be integrated to estimate the language model Δ_i by replacing the count $C_i(u)$ with the sum over all occurrences of u of their respective confidence measures, i.e.,

$$C'_i(u) = \sum_{w_j^i = u} c(w_j^i)^\lambda \ , \tag{10}$$

where $c(w_j^i)$ corresponds to the confidence measure of w_j^i. Confidence measures are raised to the power of λ in order to reduce the relative importance of words whose confidence measure value is low. Indeed, the larger λ, the smaller the impact in the total count of terms for which $c(w_j^i)$ is low.

Alternately, confidence measures can be taken into account when computing the generalized probability, by multiplying the log-probability of the occurrence of a word by the corresponding confidence measure. Formally, the generalized probability is then given by

$$\ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} c(w_j^i)^\lambda \ln P[W_j^i; \Delta_i] \ . \tag{11}$$

The idea is that word occurrences with low confidence measures contribute less to the measure of the lexical cohesion than those with high confidence measures. In this case, the language model Δ_i can be either estimated from the counts $C_i(u)$, thus limiting the use of confidence measures to the probability calculation, or from the modified counts $C'_i(u)$.

4.2.2 Semantic relations

As mentioned previously, integrating semantic relations in the measure of the lexical cohesion serves a double objective. The primary goal is obviously to ensure that two semantically related words, e.g., "car" and "drive", contribute to the lexical cohesion, thus avoiding erroneous topic boundaries between two such words. This is particularly crucial when short segments might occur as they exhibit few vocabulary repetitions. But semantic relations can also limit the impact of recognition errors. Indeed, contrary to correctly transcribed words, misrecognized words are unlikely to be semantically linked to other words in the segment. As a consequence, the weight of non properly transcribed words in the edges' weights will be less important than the one of correct words.

As for confidence measures, accounting for semantic relations can be achieved by modifying the counts in the language model estimation step. Counts, which normally reflect how many times a word appears in a segment, are extended so as to emphasize the probability of a word based on its number of occurrences as well as on the occurrences of related words. More formally, the counts C_i in Eq. (8) are amended according to

$$C_i''(u) = C_i(u) + \sum_{j=1, w_j^i \neq u}^{n_i} r(w_j^i, u) \ , \tag{12}$$

where $r(w_j^i, u) \in [0, 1]$ denotes the semantic proximity of words w_j^i and u. The semantic proximity $r(u, v)$ is close to 1 for highly related words and null for non related words. Details on the estimation of the semantic relation function r from text corpora are given in the next section.

Modified counts as defined in Eq. (12) are used to compute the language model that in turn is used to compute the generalized probability. Clearly, combining confidence measures and semantic relations is possible using confidence measures in the generalized probability computation with a language model including semantic relations and/or replacing $C_i(u)$ by $C_i'(u)$ in Eq. (12).

One of the interest of the proposed technique is that it is by nature robust to domain mismatch. Indeed, in the worst case scenario, semantic relations learnt on a specific domain will leave $C_i''(u)$ unchanged with respect to $C_i(u)$, the relations $r(u, v)$ between any two words of S_i being null. In other words, out of domain relations will have no impact on topic segmentation, a property which does not hold for approaches based on latent semantic space or model [29, 30].

4.3 Experimental results

Topic segmentation was evaluated on two distinct corpora: a *news* corpus, made up of 57 news programs ($\approx$ 30 min. each) broadcasted in February and March 2007 on the French television channel France 2, and a *reports* corpus composed of 16 reports on current affairs "Sept à Huit" ($\approx$ 1 hour each) transmitted on the French channel TF1 between September 2008 and February 2009. In the *reports* corpus,

Table 1: Comparison of the news and reports corpora in terms of word repetitions and of confidence measures

	average num. of repetitions	average confidence measure
news	1.82	0.62
reports on current affairs	2.01	0.57

longer reports and investigations can be found (around 10-15 minutes), eventually on non news topics, while the *news* corpus follows the classical scheme of rather short reports (usually 2-3 minutes). Separating the experiments in two distinct corpora enables to highlight the differences between two types of TV programs. Indeed, in addition to different program durations, the average number of topics and the average number of segments per show vary between news and reports. Moreover, the number of repetitions is less important in *news* programs than in *reports* ones, as reported in Tab. 1, while the transcription error rate is higher on the latter due to a larger amount of non professional speakers.

In each show, headlines and closing remarks were removed, these two particular parts disturbing the segmentation algorithm and being easily detectable from audiovisual clues. A reference segmentation was established by considering a topic change associated with each report, the start and end boundaries being respectively placed at the beginning of the report's introduction and at the end of the report's closing remarks. Note that in the *news* corpus, considering a topic change between each report is a choice that can be argued as, in most cases, the first reports all refer to the main news of the day and are therefore dealing with the same broad topic. A total of 1,180 topic boundaries is obtained for the *news* corpus and 86 for *reports*. Recall and precision on topic boundaries are considered for evaluation purposes after alignment between reference and hypothesized boundaries, with a tolerance on boundary locations of respectively 10 and 30 s for *news* and *reports*, while different trade-offs between precision and recall are obtained by varying α in Eq. (7).

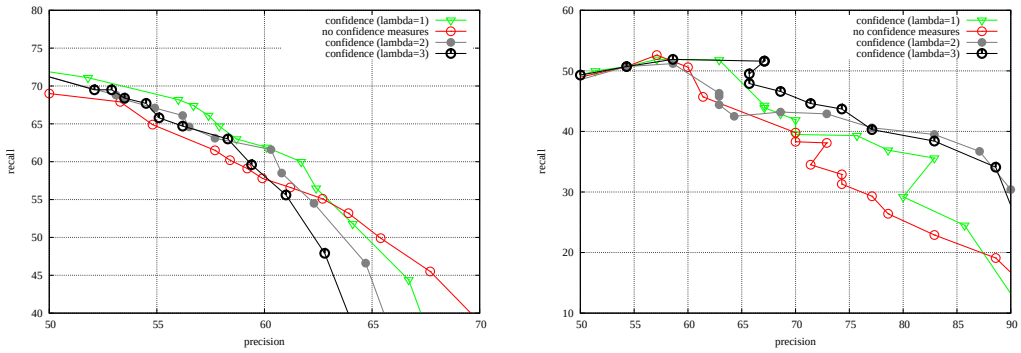
We first report results regarding confidence measures before considering semantic relations. Finally, both are taken into account simultaneously.

4.3.1 Segmentation with confidence measures

Results are reported in Fig. 4 for different values of λ, considering confidence measures simultaneously in the language model estimation and in the probability computation[8]. Using confidence measures significantly improves in all cases the quality of TV program segmentation with respect to the baseline method[9]. Moreover, con-

[8]It was experimentally observed that using confidence measures at both steps of the lexical cohesion measure leads to better results than using confidence measures solely in the language model estimation step or in the probability calculation step.

[9]A t-test was used to validate that differences in performance are statistically significant in all cases.



(a) News data (b) Reports data

Figure 4: Recall/Precision curves on the *news* and *reports* corpora for topic segmentation with confidence measures

fidence measures allow for a larger relative improvement on *reports* where the word error rate is higher. It can also be noted that results are less sensitive to variations of λ for the *news* data. This can be explained by the fact that for high values of λ, $c(w_j^i)^\lambda$ becomes negligible except for words whose confidence measure is very close to 1. As the proportion of words with a confidence measure less than 0.9 is more important in the *reports* data, the impact of the confidence measures and of λ is more perceptible on this data set. We also observed that accounting for confidence measures not only increases the number of correct boundaries detected but also improves boundary locations. Indeed, boundary locations are more precise when using confidence measures, even if this fact does not show in the recall/precision curves because of the tolerated gap on the boundary position. Finally, improving confidence measures using high-level linguistic features with a classifier [31] also benefits to boundary precision.

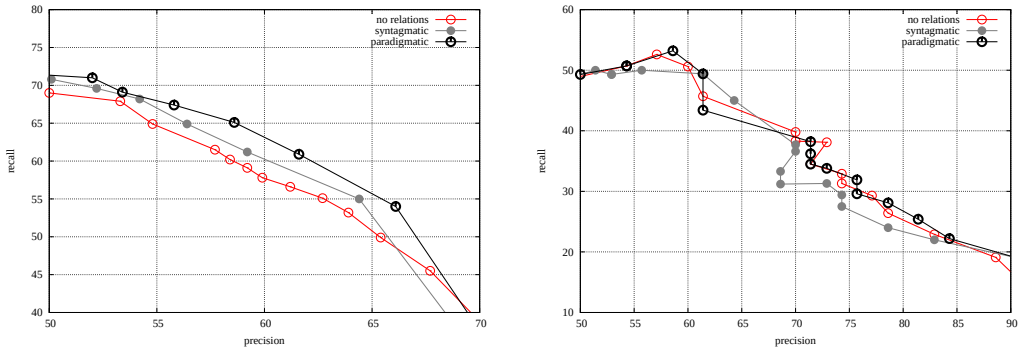
Overall, these results demonstrate not only that including confidence measures in the lexical cohesion measure improves topic segmentation of spoken TV contents, but also that the gain obtained thanks to confidence measures is larger when the transcription quality is low. This last point is a key result which clearly demonstrates that adapting text based NLP methods to the peculiarities of automatic transcripts is crucial, in particular when transcription error rates increase.

4.3.2 Segmentation with semantic relations

Two types of semantic relations, namely syntagmatic and paradigmatic ones, were automatically extracted from a corpus of articles from the French newspapers "Le Monde" and "L'Humanité" and from the reference transcript of 250 hours of radio broadcast news shows. Syntagmatic relations correspond to relations of contiguity that words maintain within a given syntactic context (sentence, chunk, fixed length window, etc.), two words being related if they often appear together. The popular mutual information cubed criterion [32] was used to acquire syntagmatic relations and was normalized in $[0, 1]$ to define the association strength $r(u, v)$. Paradigmatic relations combine two words with an important common component from a meaning

Table 2: Words with the highest association scores, in decreasing order, for the word "cigarette", automatically extracted from newspapers articles. Italicized entries correspond to cigarette brand names.

Syntagmatic relations	→	to smoke, pack, to light, smuggling, manufacturer
Paradigmatic relations	→	cigar, *Gitane*, *Gauloise*, ciggy, tobacco



(a) News data (b) Reports data

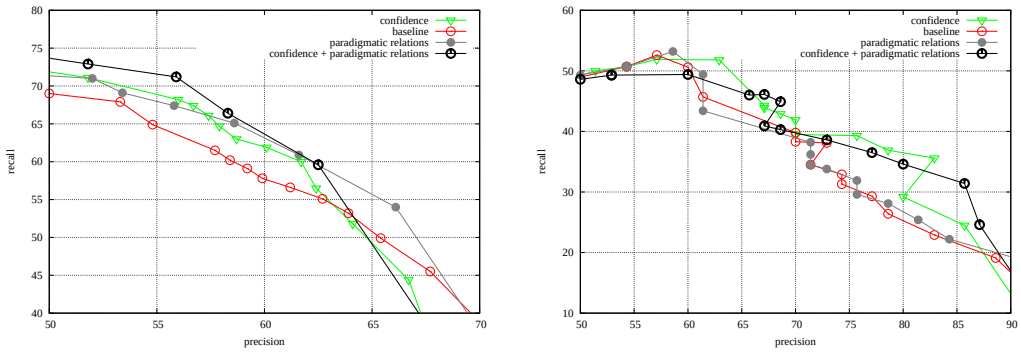
Figure 5: Recall/Precision curves for the integration of semantic relations on the *news* and *reports* data sets.

point of view. These relations, corresponding to synonyms, hyperonyms, antonyms, etc., are calculated by means of context vectors for each word, grouping together words that appear in the same contexts. The semantic proximity $r(u, v)$ is taken as the cosine distance between context vectors of u and v, normalized in the interval $[0, 1]$. Illustration of the five best syntagmatic and paradigmatic relations obtained for the word "cigarette" are given in Tab. 2[10]. Finally, various selection rules were implemented to limit the number of syntagmatic and paradigmatic relations considered, so as to keep the most relevant ones for the purpose of topic segmentation [33].

It is important to note that, contrary to many studies on the acquisition of semantic relations, both types of relations were not obtained from thematic corpora. However, they are, to a certain extent, specific to the news domain as a consequence of the data on which they have been obtained, and do not reflect the French language in general.

Results are presented in Fig. 5 on the *news* and *reports* data sets. On the *news* data, the use of semantic relations clearly improves the segmentation, paradigmatic relations yielding better performance than syntagmatic ones. This result is confirmed by observing the relations extracted: syntagmatic relations are less suited for the news domain than paradigmatic ones as they introduce more noise, connecting words and segments that should not be. Regarding the *reports* data, adding semantic relations does not improve topic segmentation, whatever the type of semantic relations considered. This result can be mainly explained by two factors. The first

[10] All examples in the article are translated from the French language.



(a) News data (b) Reports data

Figure 6: Recall/Precision curves for the combination of confidence measures and semantic relations

factor is that for the *reports* data, segments are longer and exhibit a larger number of repetitions per segment than for the *news* data, thus limiting the interest of semantic relations. The second and probably most important factor lies in the fact that semantic relations were extracted from journalistic corpora and are therefore less adapted for the *reports* corpus. As a consequence, very few words for which semantic relations were selected appear in transcripts, therefore leaving the segmentation mostly unchanged with respect to the baseline segmentation algorithm. However, it is interesting to verify that incorporating non relevant relations does not harm the segmentation process while incorporating relevant ones does help.

4.3.3 Discussion

The above series of experiments demonstrate that modifying the lexical cohesion measure to take into account confidence measures or to include semantic relations increases the robustness of topic segmentation to transcription errors as well as to genre and domain differences. In particular, we have experimentally verified that exploiting semantic knowledge from the journalistic domain does not harm (nor help) in case of out of domain data such as the ones in the *reports* data set. Final results reported in Fig. 6 also shows that the benefits of confidence measures and of semantic relations are cumulative.

As for the experiments of Section 3 in speech based program description, these results again proove that adapting NLP tools to better interface with ASR is a good answer to robustness issues. However, in spite of the proposed modifications, high transcription error rates are still believed to be detrimental to topic segmentation and progress are required towards truly genre independent topic segmentation techniques. Still, from the results presented, topic segmentation of spoken contents has reached a level where it can be used as a valuable tool for the automatic delinearization of TV data, however limiting the use of such techniques to specific program genres where reasonable error rates are achieved and where topic segmentation makes sense. This claim is supported by our experience on automatic news delinearization, as illustrated by the NEM Summit demonstration presented in Section 6 or by the

Voxalead news indexing prototype presented for the ACM Multimedia 2010 Grand Challenge [34].

5 Automatically linking contents

One of the key features of Internet TV diffusion is to enhance navigability by adding links between contents and across modalities. So far, we have considered speech as a descriptor for characterization or segmentation purposes. However, semantic analysis can also be used at the far end of the delinearization process illustrated in Fig. 1 to automatically create links between documents. In this section, we exploit a keyword based representation of spoken contents to create connections between segments resulting from the topic segmentation step or between a segment and related textual resources on the Web. Textual keywords extracted from speech transcripts are used as a pivot semantic representation upon which characterization and navigation functionalities can be automatically built. We propose adaptations of classical keyword extraction methods to account for spoken contents and describe original techniques to query the Web so as to create links.

We briefly highlight the specificities of keyword extraction from transcripts, exploiting the confidence measure weighted *tf-idf* criterion of Section 3. We then propose a robust strategy to find relations between documents, exploiting keywords and IR techniques.

5.1 Keyword characterization

We propose to use keywords to characterize spoken contents as keywords offer compact yet accurate semantic description capabilities. Moreover, keywords are commonly used to describe various multimedia contents such as images in Flickr or videos in portals such as YouTube. Hence, a keyword based description is a natural candidate for cross-modal link generation.

Given a document (e.g., a segment resulting from topic segmentation), keywords are classically selected based on the modified *tf-idf* weight given by Eq. (5), keeping a few words with the highest *tf-idf* weights as keywords. However, the *tf-idf* criterion is known to emphasize words with low *idf* scores as soon as they appear in a transcript, in particular proper names (journalist or city names, trade marks, etc.). If such names are highly descriptive, they play a particular role. Some of them are related to the context (journalists' names, channel names, etc.) while some are related to the content (politicians' names, locations, etc.). We observed that keeping too many proper names as keywords often result in a poor characterization of the (broad) topic (e.g. Tennis), providing a very detailed description (e.g. Nadal vs. Federer) and, consequently, very few links can be established. Moreover, proper names are likely to be misrecognized. We therefore chose not to emphasize proper names as keywords, thus giving greater importance to broad topic characterization. Limiting the influence of proper names is implemented by applying a penalty $p \in [0, 1]$ to the

Table 3: List of the 10 keywords with the highest scores after inclusion of confidence measures within the score computation

$\sigma(\ell)$	Word class ℓ
0.992	{veil}
0.500	{secularity}
0.458	{muslim, muslims}
0.454	{adda}
0.428	{photo, photos}
0.390	{bernadette}
0.371	{prefecture}
0.328	{chador}
0.325	{carmelite}
0.321	{sarkozy}

term frequency, according to

$$tf'(l,t) = \frac{\sum_{w \in l} p_w}{|l|} \times tf(l,t) \ , \qquad \text{with } p_w = \begin{cases} 1-p & \text{if } w \text{ is a proper name} \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

where $|l|$ is the number of words whose corresponding lemma is l. This biased term frequency is used in Eq. (5) for keyword selection. In the navigation experiments presented in the next section, proper names are detected based on the part-of-speech tags and a dictionary, where nouns with no definition in the dictionary are considered as proper names.

An example of the 10 first keywords extracted from a sample segment is presented in Tab. 3, the keywords accurately defining most aspects of the topic. As it can be glimpsed, the segment relates the story of a nun who refused to take off her veil on official photos.

Beyond the help provided to users in quickly understanding the content of a segment, this characterization scheme can also be used as a support for linking segments with semantically related documents.

5.2 Hyperlink generation

In the context of delinearization and Internet TV diffusion, it is of particular interest to create links from TV segments to related resources on the Web. The problem of link generation is therefore to automatically find Web contents related to the segment's topic as characterized by the keywords. We propose a two step procedure for link generation, where candidate Web pages are first retrieved based on keywords before filtering candidate pages using the entire transcript as a description.

Table 4: Example of queries formed based on subsets of the 5 best-scored keywords. Queries in bold include at least one misrecognized word.

veil	secularity			
veil		muslims		
veil			**adda**	
veil				photo
	secularity	muslims		
	secularity		**adda**	
	secularity			photo
		muslims	**adda**	
		muslims		photo
veil	secularity	muslims		
	secularity	**muslims**	**adda**	
		muslims	**adda**	**photo**
veil	**secularity**		**adda**	
veil	secularity			photo
	secularity		**adda**	**photo**

5.2.1 Querying the Web with keywords

Given keywords, contents on the Web can be automatically retrieved using classical Web search engines[11] by deriving one or several queries from the keywords. Creating a single meaningful query from a handful of keywords is not trivial, all the more when misrecognized words are included as keywords in spite of the confidence measure weighted *tf-idf* scores. Thus, using several small queries appears to be clearly more judicious. Another issue is that queries must be precise enough to return topic-related documents without being too specific in order to retrieve at least one document. The number of keywords included in a query is a good parameter to handle these constraints. Indeed, submitting too long queries, i.e., composed of too many keywords, usually results in no or only few hits, whereas using isolated keywords as queries is frequently ineffective since the meaning of many words is ambiguous regardless of the context. Hence, we found that a good query generation strategy consists in building many queries combining subsets of 2 or 3 keywords. Furthermore, in practice, as words are more precise than lemmas when submitting a query, each lemma is replaced by its most frequent inflected form in the transcript of the segment or document considered.

An example of 15 queries derived from the sole 5 best keywords of Tab. 3 is given in Tab. 4. In this example, the keyword "adda" is a misrecognized word that has not been completely discarded by confidence measures. Nonetheless, by using several queries, this error only impacts half of the queries (in bold), which keeps still high the chance of having generated adequate meaningful queries.

[11] Yahoo! and Bing in this work.

5.2.2 Selecting relevant links

The outcome of the querying strategy is a list of documents—a.k.a hits—on the Web ordered by relevance with respect to the queries. Relevant links are established by finding among these hits the few ones that best match the topic of a segment characterized by the entire set of keywords rather than by two or three keywords. Assuming that the relevance of a Web page with respect to a query decreases with its rank in the list of hits, we solely consider the first few results of each query as candidate links. In this work, 7 documents are considered for each of the 15 queries. To select the most relevant links among the candidate ones, the vector space model with *tf-idf* weights is used. Candidate Web pages are cleaned and converted into regular texts[12] represented in the vector space model using *tf-idf* scores. Similarly, a segment's automatic transcript is represented by a vector of modified *tf-idf* scores as in Section 3. For both the Web pages and the transcript, the weight of proper names is softened as previously explained. The cosine distances between the segment considered and the candidate Web pages finally enables to keep only those candidate links with the highest similarity.

5.2.3 Discussion

We have proposed a domain independent method to automatically create link between transcripts and Web documents, using a keyword based characterization of spoken contents. Particular emphasis has been put on robustness to transcription errors, using modified *tf-idf* weights for keyword selection, designing a querying strategy able to cope with erroneous keywords and using an efficient filtering technique to select relevant links based on the characterization presented in Section 3.

Though no objective evaluation of automatic link generation has been performed, we observed in the framework of the NEM Summit demonstration described in the next section that the generated links are in most cases very relevant. However, in a different setting, these links were also used to collect data for the unsupervised adaptation of the ASR system language model [35]. Good results obtained on this LM adaptation task are also an indirect measure of the quality of the links generated. Nevertheless, the proposed hyperlink creation method could still be improved. For example, pages could be clustered based on their respective similarities. By doing so, the different topic aspects of a segment could be highlighted and characterized by specific keywords extracted from clustered pages. "Keypages" could also be returned by selecting the centroids of each cluster. The scope of the topic similarity could also be changed depending on the abstraction level desired for the segment characterization. For example, pages telling the exact event of a segment—instead of pages dealing with the same broad topic—could be returned by re-integrating proper names into keyword vectors.

Finally, let us note that, beside the retrieval of Web pages, the link generation

[12] A Web page in HTML format is cleaned by pruning the DOM tree based on typographical clues (punctuation signs and uppercase characters frequencies, length of sentences, number of non-alphanumeric characters, etc.), so as to remove irrelevant parts of the document such as menus, advertisements, abstracts or copyright notifications.

technique proposed here can also be used to infer a structure between segments of the same media (e.g., between a collection of transcribed segments as in the example below). The technique can also be extended to cross-media link generation, assuming a keyword based description and an efficient cross-modal filtering strategy are provided.

6 Illustration: automatic hypernews generation

To illustrate the use of the speech-based media processing technologies presented in this paper in the delinearization process, we describe a news navigation demonstration that was presented during the NEM Summit 2009. This demonstration automatically builds a news navigator interface, illustrated in Fig. 7, from a set of broadcast news shows. Shows are segmented and presented in a navigable fashion with relations either between two news reports at different dates or between a news report and related documents on the Internet.

Note that this preliminary demonstration, limited to the broadcast news domain, is intended as an illustration to illustrate automatic delinearization of (spoken) TV contents and to validate our work on a robust interface between ASR and NLP. Similar demonstrations on broadcast news collections have been developed in the past (see, e.g., [7, 8, 36, 10, 37]) but mostly rely on genre dependent techniques. On the contrary, we rely on robust genre and domain independent techniques, thus making it possible to extend the concept to virtually all kinds of contents. Moreover, all of the above mentioned applications lack navigation capabilities other than through a regular search engine.

We briefly describe the demonstration before discussing the quality of the links generated. For lack of objective evaluation criteria, we provide a qualitative evaluation to illustrate the remaining challenges for spoken language processing in the media context.

6.1 Overview of the hypernews showcase

The demonstration was built on top of a collection of evening news shows from the French channel France 2 recorded daily over a 1 month period[13]. After transcription, topic segmentation as described in Section 4 was applied to each show in order to find out segments corresponding to different topics (and hence events in the broadcast news context). Keyword extraction as described in Section 5.1 was applied in order to characterize each of the 553 segments obtained as a result of the segmentation step. Based on the resulting keywords, exogenous links to related Web sites were generated as explained in Section 5.2. Endogenous links between segments, within the collection, were established based on a simple keyword comparison heuristics[14].

[13] Le Journal de 20h, France 2, from Feb. 2, 2007 to Mar. 23, 2007.

[14] Note that different techniques could have been used for endogenous link generation. In particular, the same filtering technique as for exogenous links could be used. The idea behind a simple keyword comparison was, in the long term, to be able to incrementally add new segments daily, a task which requires highly efficient techniques to compare two segments.

Fig. 7(a) illustrates the segmentation step. Segments resulting from topic segmentation are presented as a table of contents for the show with links to the corresponding portions of the video and a few characteristic keywords to provide an overview of each topic addressed. Fig. 7(b) illustrates the navigation step where "See also" provides a list of links to related documents on the Web while "Related videos" offers navigation capabilities within the collection.

6.2 Qualitative analysis

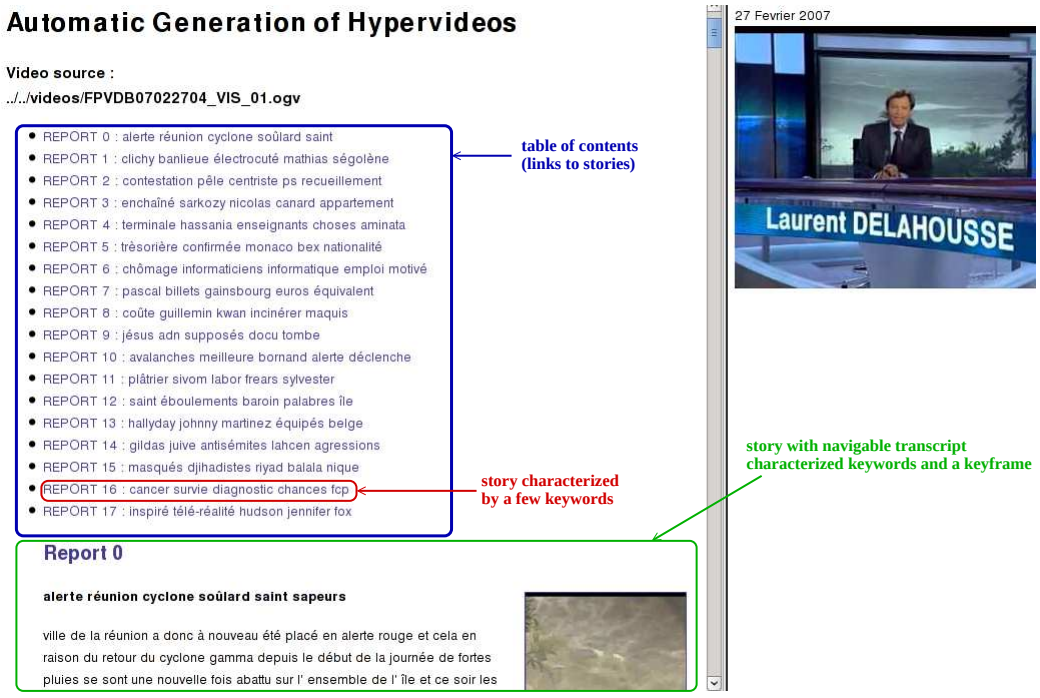
Quantitative assessment of the links automatically generated is a difficult task and we therefore limit ourselves to a qualitative discussion on the relevance of the generated links. As mentioned in the introduction, we are fully aware of the fact that a qualitative analysis, illustrated with a few selected examples, does not provide the ground for sounded scientific conclusions as a quantitative analysis would. However, this analysis gives an idea of the types of links that can be obtained and of the remaining problems.

6.2.1 External links

It was observed that links to external resources on the Web are mostly relevant and permit to access related information. As such links are primarily generated from queries made of a few general keywords that do not emphasize named entities, they point to Web pages containing additional information rather than to Web pages dealing with the same story[15]. Taking the example of the cyclone Gamède which struck the Île de la Réunion in February 2007, illustrated in Fig. 7(b), all links are relevant. Several links target sites related to cyclones in general (list of cyclones, emergency rules in case of cyclones, cyclone season, etc.) or to sites dedicated to specific cyclones, including the Wikipedia page for cyclone Gamède. Additionally, one link points to a description of the geography and climate in the Île de la Réunion while the less relevant link points to a flood in Mozambique due to a cyclone.

General information links such as those described previously present a clear interest for users and offer the great advantage of not being related to news sites whose content changes at a fast pace. Moreover, the interest of enriching contents with general purpose links is not limited to the news domain and applies to many types of programs. For example, in movies or talkshows, users might be interested in having links to documents on the Web related to the topic discussed. However, in the news domain, more precise links to the same story or to similar stories on other medias are required, a feature that is not covered by the technique proposed. We believe that accounting for the peculiar nature of named entities in the link generation process is one way of focusing links on very similar contents, yet remaining domain independent.

[15]This fact is also partially explained by the time lag between the corpus (Feb.–Mar., 2007) and the date at which the demonstration's links were established (Jun. 2009), as most news articles on the Web regarding the Feb.–Mar. 2007 period had been removed from the news sites in 2009.



(a) Navigation part of the interface illustrating segmentation into reports and characterization by a few keywords

(b) Navigation part of the interface for a particular report, illustrating links to Web documents and related segments in the collection

Figure 7: Screenshots of the automatically generated hypernews Web site

6.2.2 Internal links

Links between reports within the collection of broadcast news shows were established based on common keywords, considering their ranks. In spite of using a simplistic distance measure between reports, we observed that the first few links are most of the time relevant. Taking again the example of the Gamède cyclone illustrated in Fig. 7(b), the main title on Feb. 27 reports on the cyclone hitting the island and is related to the following reports, in order of relevance:

1. update on the island's situation [Feb. 27]
2. snow storm and avalanches in France [Feb. 27]
3. return to normal life after the cyclone and risk of epidemic [Mar. 3]
4. aftermath of the cyclone [Feb. 28]
5. damages due to the cyclone [Mar. 2]

The remaining links are mostly irrelevant, apart from two links to other natural disasters. Regardless of item 2, the first links are all related to the cyclone and, using the broadcasting dates and times for navigation, one can follow the evolution of the story across time. Note that a finer organization of the collection into clusters and threads [36, 37] is possible but the notion of threads seldom applies outside of the news domain while links generation on the base of a few keywords is domain-independent. Finally, as for external links, accounting for named entities would clearly improve relevance but possibly also prevents connections to different stories of the same nature, e.g., from the cyclone to other natural disasters.

7 Future work

We have presented research work targeting the use of speech for the automatic delinearization of TV streams. To deal with the challenges of ASR transcripts in this context, such as potentially high error rates and domain independence, we have proposed several adaptations of traditional domain independent information retrieval and natural language processing techniques so as to increase their robustness to the peculiarities of automatically transcribed TV data. In particular, in Section 3, we have proposed a modified *tf-idf* weighting scheme to exploit noisy transcripts, with error rates varying from 15 % to 70 %. We also adapted a lexical cohesion criterion and demonstrated that speech can be used for the segmentation of TV streams into topics. Experimental results show that, in spite of high error rates, a "bag of words" representation of TV contents can be used as description, for the purposes of information retrieval and navigation. All these results clearly indicate that designing techniques genuinely targeting spoken contents increase the robustness of spoken document processing to transcription errors. This in turn leads us to believe that this philosophy will pave the road towards sufficient robustness for speech to be used as a valuable source of semantic information for all genres of programs.

Clearly, not all aspects of speech based delinearization have been tackled in this paper and many work is still required in order to make the most of speech transcripts for TV stream processing. We now briefly review the main research directions that we feel are crucial.

First of all, the potential of speech transcription have been experienced solely on a very specific content type, broadcast news, and therefore still need to be validated on other types of programs such as investigation programs and debates for which transcription error rates are significantly higher. However, results presented on the validation of the EPG alignment and on the topic segmentation tasks indicates that speech is a valuable source of information to process in spite of transcription errors. In this regard, we firmly believe that a better integration of NLP and ASR—accounting for confidence measures and alternate transcription hypotheses in NLP; incorporating high level linguistic knowledge in ASR systems; accounting for phonetics in addition to a lexical transcription, etc.—is a crucial need to develop robust and generic spoken content processing techniques in the framework of TV stream delinearization. In particular, named entities such as locations, organizations or proper names play a very particular role in TV contents and should therefore receive particular attention in designing content-based descriptions. But, if acceptable named entity detection solutions exists for textual data, many factors prevent the straightforward use of such solutions for automatic transcripts from being viable. Among those factors are transcription errors and, most of all, the fact that named entities are often not in the vocabulary of the ASR system and hence not recognized (see [18] for a detailed analysis).

Topic segmentation of TV programs is another point which requires additional research effort. Domain-independent topic segmentation methods such as the one presented in this paper exhibit almost acceptable performance. In fact, in the demonstration, we observed that in most cases, segmentation errors have little impact on the acceptability of the results. Indeed, in a segment where two topics are discussed, keywords will often be a mix of keywords characterizing each of the two topics. This has little impact in our demonstration since broad characterization are considered, linking segments and documents from the same broad topic. However, we expect such errors to be a strong limitation as soon as a more detailed description will be required. Unless the number of keywords is increased drastically, it will be difficult to precisely characterize a two topic segment. But significantly increasing the number of keywords will result in more noise and errors in the description. Hence, progress are still required in the topic segmentation domain. Moreover, only linear topic segmentation has been considered so far. But there is clearly a hierarchical topic structure in most programs, depending on the precision one wants on a topic. A typical example of this fact is that of the main title in news shows where several reports tackle different aspects and implications of the main title, each report eventually consisting of different points of views on a particular question. But hierarchical topic segmentation methods and hierarchical content description have been seldom studied and still require significant progress to be operational.

Finally, link generation based on automatically extracted keywords has proved quite efficient but lacks finesse in creating a semantic Web of interconnected mul-

timedia contents, even in the news domain. More elaborated domain independent techniques to automatically build threads based on speech understanding are still required in spite of the recent efforts in that direction [36, 37]. Moreover, most multimedia documents are by nature multimodal and modalities other than text (eventually resulting from automatic speech transcription) should be fully exploited. Limiting ourselves to the news domain, image comparison could for example be used to link similar contents. Evidently, modalities other than language cannot provide as detailed a semantic description as language can but we hope that, to a certain extent, they can compensate for errors in ASR and NLP and increase robustness and precision of automatically generated semantic links. However, many issues remain open in this area from the construction of threads to the use of multiple modalities for content-based comparison of documents.

From a more philosophical point of view, it is interesting to note that the key goal of topic segmentation is to shift from the notion of stream to that of *document*, the latter being the segment, in order to back off to well known information retrieval techniques which operates at the document level. For example, the very notion of *tf-idf* is closely related to that of document. So is the notion of index. Establishing links between contents also strongly relies on the notion of document as current techniques solely permit the comparison of two documents with well defined boundaries. However, one can wonder whether the notion of document still makes sense in a continuous stream or not. Going back to the cyclone example of Section 6.2, it might be interesting to link a particular portion of the report to only a portion of a related document where the latter might contain more than required. The idea of hierarchical topic segmentation is one step in that direction, enabling to choose the extent of the portion of the stream to be considered. But it might also prove interesting to revisit information retrieval techniques in the light of this reflexion and design new techniques not dependent on the notion of document.

Acknowledgments

The authors are most grateful to Sébastien Campion and Mathieu Ben for their hard work on assembling bits and pieces of research results into an integrated demonstration and for presenting the latter during the NEM Summit 2009.

This work was partially funded by OSEO in the framework of the Quaero project.

References

[1] John Zimmermann, George Marmaropoulos, and Clive van Heerden. Interface design of Video Scout: A selection, recording, and segmentation system for TVs. In *Proc. International Conference on Human-Computer Interaction*, 2001.

[2] Liuhong Liang, Hong Lu, Xiangyang Xue, and Yap-Peng Tan. Program segmentation for TV videos. In *Proc. IEEE International Symposium on Circuits and Systems*, 2005.

[3] Xavier Naturel, Guillaume Gravier, and Patrick Gros. Fast structuring of large television streams using program guides. In Stéphane Marchand-Maillet, Eric Bruno, Andreas Nürnberger, and Marcin Detyniecki, editors, *Proc. International Workshop on Adaptive Multimedia Retrieval*, volume 4398 of *Lecture Notes in Computer Science*, pages 223–232. Springer, 2006.

[4] Gaël Manson and Sid-Ahmed Berrani. Automatic TV broadcast structuring. *International Journal of Digital Multimedia Broadcasting*, 2010. doi:10.1155/2010/153160.

[5] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognition Letters*, 25(7):767–775, 2004.

[6] Ewa Kijak, Guillaume Gravier, Lionel Oisel, and Patrick Gros. Audiovisual integration for tennis broadcast structuring. *Multimedia Tools and Application*, 30(3):289–312, 2006.

[7] Andrew Merlino, Daryl Morey, and Mark T. Maybury. Broadcast news navigation using story segmentation. In *Proc. ACM International Conference on Multimedia*, 1997.

[8] Mark T. Maybury. Broadcast news navigator (BNN) demonstration. In *Proc. International Joint Conferences on Artificial Intelligence*, 2003.

[9] Katsuoshi Ohtsuki, Katsuji Bessho, Yoshihiro Matsuo, Shoichi Matsunaga, and Yoshihiko Hayashi. Automatic multimedia indexing: Combining audio, speech, and visual information to index broadcast news. *IEEE Signal Processing Magazine*, 23(2):69–78, March 2006.

[10] Mike Dowman, Valentin Tablan, Hamish Cunningham, Cristian Ursu, and Borislav Popov. Semantically enhanced television news through Web and video integration. In *Proc. Multimedia and the Semantic Web, Workshop of the 2nd European Semantic Web Conference*, 2005.

[11] Hisashi Miyamori and Katsumi Tanaka. Webified video: Media conversion from TV programs to Web content for cross-media information integration. In Kim Viborg Andersen, John K. Debenham, and Roland Wagner, editors, *Proc. International Conference on Database and Expert Systems Applications*, volume 3588 of *Lecture Notes in Computer Science*, pages 176–185. Springer, 2005.

[12] Alexander Hauptmann, Robert Baron, Ming-Yu Chen, Mike Christel, Pinar Duygulu, Chang Huang, Rong Jin, Wei-Hao Lin, Dorbin Ng, Neema Moraveji, Norman Papernick, Cees Snoek, Georgos Tzanetakis, Jie Yang, Rong Yan, and Howard Wactla. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proc. Text Retrieval Conference*, 2003.

[13] Julien Law-To, Gregory Grefenstete, and Jean-Luc Gauvain. Voxaleadnews: Robust automatic segmentation of video into browsable content. In *ACM Multimedia*, 2009.

[14] Neeraj Deshmukh, Aravind Ganapathiraju, and Joseph Picone. Hierarchical search for large-vocabulary conversational speech recognition. *IEEE Signal Processing Magazine*, pages 84–107, 1999.

[15] Hermann Ney and Stefan Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, pages 64–83, 1999.

[16] Hui Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

[17] Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. *Computer Speech and Language*, 24:663–684, 2010.

[18] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proc. Annual Conference of the International Speech Communication Association*, 2009.

[19] Xavier Naturel and Sid-Ahmed Berrani. Content-based TV stream analysis techniques toward building a catch-up TV service. In *Proc. IEEE International Symposium on Multimedia*, 2009.

[20] Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. Can automatic speech transcripts be used for large scale TV stream description and structuring? In *Proc. International Workshop on Content-Based Audio/Video Analysis for Novel TV Services*, 2009.

[21] Gerard Salton. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[22] Jonathan Mamou, David Carmel, and Ron Hoory. Spoken document retrieval from call-center conversations. In *SIGIR*, pages 51–58, 2006.

[23] James Allan, editor. *Topic detection and tracking: Event-based information organization*, volume 12 of *The Information Retrieval Series*. Kluwer Academics, 2002.

[24] Marti A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64, 1997.

[25] Paul Van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. Segmentation of automatically transcribed broadcast news text. In *Proc. DARPA Broadcast News Workshop*, 1999.

[26] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210, 1999.

[27] Heidi Christensen, Balakrishna Kolluru, Yoshihiko Gotoh, and Steve Renals. Maximum entropy segmentation of broadcast news. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

[28] Masao Utiyama and Hitoshi Isahara. A statistical model for domain-independent text segmentation. In *Proc. Annual Meeting of the Association for Computational Linguistics*, 2001.

[29] Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. Latent semantic analysis for text segmentation. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 109–117, 2001.

[30] Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappé. Text segmentation via topic modeling: an analytical study. In *Proc. Intl. Conf. on Information and Knowledge Management*, 2009.

[31] Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, and Patrick Gros. CRF-based combination of contextual features to improve a posteriori word-level confidences measures. In *Proc. Annual Intl. Speech Communication Association Conference (Interspeech)*, 2010.

[32] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. In Philip Resnik and Judith L. Klavans, editors, *The balancing act: Combining symbolic and statistical approaches to language*, pages 49–66. MIT Press, 1996.

[33] Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations. In *Proc. Annual Intl. Speech Communication Association Conference (Interspeech)*, 2010.

[34] Julien Law-To, Gregory Grefenstete, Jean-Luc Gauvain, Guillaume Gravier, Lori Lamel, and Julien Despres. VoxaleadNews: Robust automatic segmentation of video content into browsable and searchable subjects. In *ACM Multimedia*, 2010.

[35] Gwénolé Lecorvé, Guillaume Gravier, and Pascale Sébillot. An unsupervied Web-based topic language model adaptation method. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[36] Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh. Topic threading for structuring a large-scale news video archive. In *Proc. International Conference on Image and Video Retrieval*, 2004.

[37] Xiao Wu, Chong-Wah Ngo, and Qing Li. Threading and autodocumenting news videos: a promising solution to rapidly browse news topics. *IEEE Signal Processing Magazine*, 23(2):59–68, 2006.