6-2008

Measuring novelty and redundancy with multiple modalities in cross-lingual broadcast news

Xiao WU

Alexander G. HAUPTMANN

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Citation

1

Available online at www.sciencedirect.com

ScienceDirect

Computer Vision and Image Understanding 110 (2008) 418–431

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

Measuring novelty and redundancy with multiple modalities in cross-lingual broadcast news

Xiao Wu [a,b,*], Alexander G. Hauptmann [a], Chong-Wah Ngo [b]

[a] *School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA*
[b] *Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong*

Received 1 November 2006; accepted 5 September 2007
Available online 15 December 2007

Abstract

News videos from different channels, languages are broadcast everyday, which provide abundant information for users. To effectively search, retrieve, browse and track news stories, news story similarity plays a critical role in assessing the novelty and redundancy among news stories. In this paper, we explore different measures of novelty and redundancy detection for cross-lingual news stories. A news story is represented by multimodal features which include a sequence of keyframes in the visual track, and a set of words and named entities extracted from speech transcript in the audio track. Vector space models and language models on individual features (text, named entities and keyframes) are constructed to compare the similarity among stories. Furthermore, multiple modalities are further fused to improve the performance. Experiments on the TRECVID-2005 cross-lingual news video corpus showed that modalities and measures demonstrate variant performance for novelty and redundancy detection. Language models on text are appropriate for detecting completely redundant stories, while Cosine Distance on keyframes is suitable for detecting somewhat redundant stories. The performance on mono-lingual topics is better than multilingual topics. Textual features and visual features complement each other, and fusion of text, named entities and keyframes substantially improves the performance, which outperforms approaches with just individual features.

Keywords: Similarity measure; Novelty; Redundancy detection; Multimodality; Cross-lingual; Near-duplicate keyframe; News videos

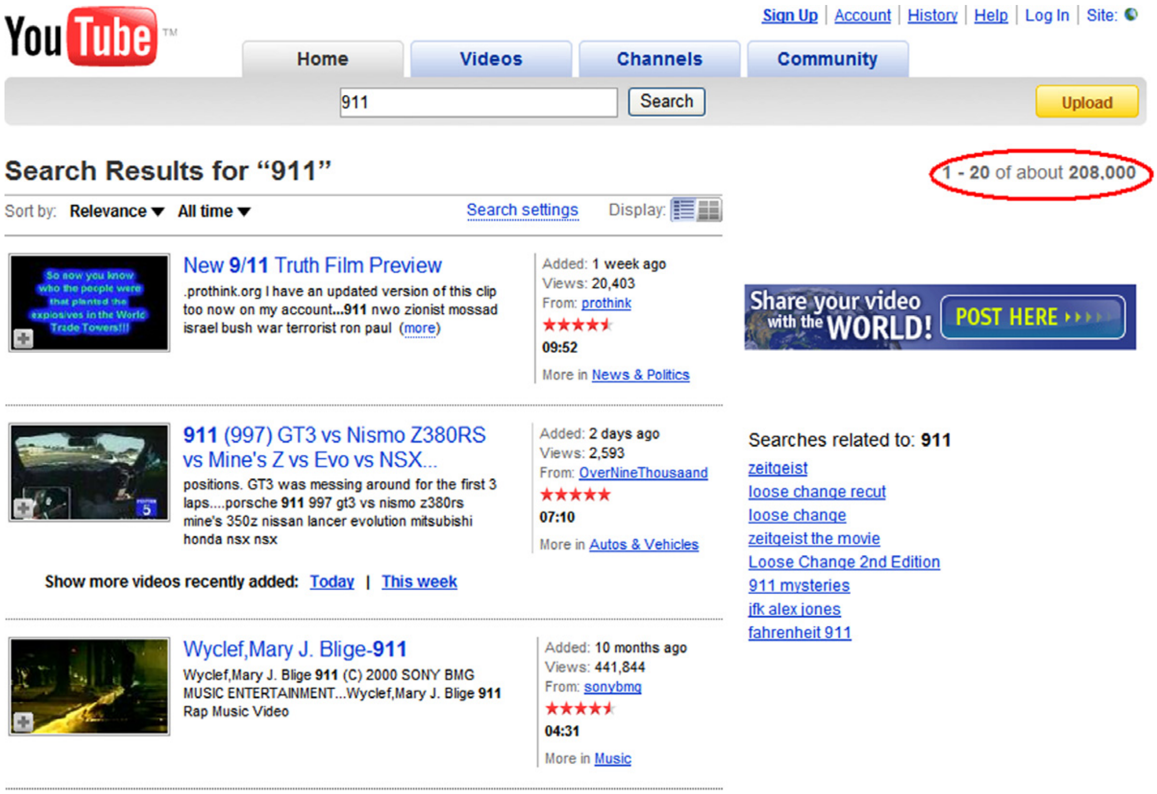
1. Introduction

With the overwhelming volume of news videos from different sources, languages and nations available nowadays, it becomes important to manage these news stories in an automatic and efficient way so that it satisfies user's information needs with minimum effort. Especially with the explosion of Internet, videos are easily accessible and have grown exponentially. For instance, video website YouTube has 50 million users from different countries, and users post about 60,000 videos each day. When searching "9/11" at YouTube, there are 208,000 videos returned (see Fig. 1). Among these cross-lingual news stories, some reports convey the exactly duplicate contents with different languages, while others may carry fresh information, and even totally different viewpoints toward the same event. It is difficult for ordinary users to view all of them, and find out the interest parts among the huge volume of videos. To facilitate effective search, retrieval, browsing and tracking of news stories, news story similarity plays a fundamental role in measuring the novelty and redundancy among news stories. For news stories, novelty and redundancy detection is a special case of similarity measure. However, cross-lingual news story similarity assessment remains a challenging problem that has rarely been explored. In this paper, we discuss different similarity measures on multiple modalities to detect the novelty and redundancy for cross-lingual news stories.

[*] Corresponding author. Address: Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong.

E-mail addresses: wuxiao@cs.cityu.edu.hk (X. Wu), alex@cs.cmu.edu (A.G. Hauptmann), cwngo@cs.cityu.edu.hk (C.-W. Ngo).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2007.09.015


YouTube
Sign Up | Account | History | Help | Log In | Site:
Home
Videos
Channels
Community
911
Search
Upload
Sort by: Relevance ▾ All time ▾
Search settings
Display:
So now you know who the people were that planted the explosives in the World Trade Towers!!!
Share your video with the WORLD! POST HERE

Search Results for "911"

1 - 20 of about 208,000

New 9/11 Truth Film Preview
.prothink.org I have an updated version of this clip too now on my account...911 nwo zionist mossad israel bush war terrorist ron paul (more)
Added: 1 week ago
Views: 20,403
From: prothink
★★★★½
09:52
More in News & Politics

911 (997) GT3 vs Nismo Z380RS vs Mine's Z vs Evo vs NSX...
positions. GT3 was messing around for the first 3 laps....porsche 911 997 gt3 vs nismo z380rs mine's 350z nissan lancer evolution mitsubishi honda nsx nsx
Added: 2 days ago
Views: 2,593
From: OverNineThousaand
★★★★★
07:10
More in Autos & Vehicles

Show more videos recently added: Today | This week

Wyclef,Mary J. Blige-911
Wyclef,Mary J. Blige 911 (C) 2000 SONY BMG MUSIC ENTERTAINMENT...Wyclef,Mary J. Blige 911 Rap Music Video
Added: 10 months ago
Views: 441,844
From: sonybmg
★★★★½
04:31
More in Music

Searches related to: 911
zeitgeist
loose change recut
loose change
zeitgeist the movie
Loose Change 2nd Edition
911 mysteries
jfk alex jones
fahrenheit 911

Fig. 1. There are 7470 videos on "9/11" at YouTube website.

A news story is defined as a segment of a news broadcast with a coherent news focus containing at least two independent and declarative clauses [1], which is a meaningful and semantic unit about one topic. Different from traditional text-based news articles, news videos include both audio and visual channels. In addition to the text transcript, news videos carry useful visual contents. In broadcast videos, stories under the same topic are often accompanied by a couple of shots that tend to be used repeatedly due to the lack of fresh materials for footage or as a reminder. For example, the scene of planes hitting the World Trade Center is repeatedly broadcast by stories on "9/11 terrorist attack". These shots can be reused with some modifications either as a reminder of the story or due to a lack of video material for the current footage, which provides useful cues to measure the similarity of news stories. Previous novelty and redundancy detection mostly focuses on textual news articles. However, the employment of either textual or visual information may not be enough since either content can appear differently over time. Therefore, we will exploit the textual and visual information to measure the news story similarity.

The scenario is more complicated when considering news stories from multilingual sources such as Chinese, English and Arabic. Due to errors and lack of proper processing tools for the speech recognition and machine translation, the transcripts are usually noisy and even unavailable when facing multilingual videos. The performance of previous text-based approaches applied to news videos is still not clear.

We define "*Redundant*" to mean that most of the textual and visual information in a news story is covered by previously delivered news stories. The definition of redundancy includes "duplicate" and "near-duplicate" news stories as well as news stories that are redundant in content but very different in presentation. For those news stories that have similar visual concepts but carry new information, or their textual contents are same while video shots are totally different, they are somewhat redundant. Novel stories bring fresh textual and visual contents and introduce new emergence of themes.

An important application of similarity measure for news stories is topic detection and tracking (TDT) [2]. When a specific event breaks out (e.g., "9/11 terrorist attack"), news stories are broadcast immediately to report this event from multiple channels in different languages. Among them, there exist a large number of repeated reports. With the evolving of the event, new contents emerge over time while some information changes slowly or even remains unchanged throughout the topic. Users of news channels do not want to view every piece of video over and over again. They are primarily interested in learning how the topic is evolved or what the highlight of this topic is. It is difficult for ordinary users to manually filter redundant reports and identify fresh development among the huge volume of information within limited time. Thus, identifying and reorganizing news stories by novelty/redundancy promises to be useful. Avoiding redundancy and overlap can help minimizing the overhead associated with searching and tracking news stories.

In this paper, we explore the novelty and redundancy detection for cross-lingual news stories. Measures on vector space model and language model are applied to text, named entities and keyframes, respectively, to evaluate the performance. Moreover, fusion of text, named entities and keyframes is also explored. The rest of this paper is organized as follows. In Section 2, we give a brief description of related work. Proposed framework is introduced in Section 3. The multimodality representation for news story and the similarity measures for novelty and redundancy detection are discussed in Sections 4 and 5, respectively. Section 6 presents the experiments. Finally, we conclude the paper with a summary.

2. Related work

2.1. Textual information retrieval

Text-based similarity measures have been extensively studied previously for estimating the similarity of text passage. Several techniques have been developed from low-level syntactic similarity (e.g., document fingerprinting technique [3]) to high-level semantic similarity (novelty and redundancy detection [4]). The high-level similarity measure, i.e., novelty and redundancy detection, is more appropriate for news story similarity measure due to the diversity of vocabulary used by different channels. Novelty detection has been mainly explored at three different levels: the event level, document level and sentence level. Novelty detection at the event level originated from the work of new event detection (NED) or first story detection (FSD) [5]. The most related work to our research in novelty/redundancy detection at the document and sentence level are [4,6,7]. Zhang et al. [4] addressed the novelty and redundancy detection of relevant documents. Allan et al. [6] discussed the novelty detection at the sentence level. Among these techniques, new words appearing in documents or sentences contribute to the score used to measure the novelty. Named entities were exploited to do the new event detection and novelty detection in [8–10]. In [8], Newsjunkie employed novelty analysis algorithms that represent articles as words and named entities to provide personalized newsfeeds. Furthermore, cross-language tracking was explored in [11]. However, these approaches were based on the textual concepts. The performance applied to noisy news videos scenario is still uncertain.

Topic detection and tracking (TDT) [2] is a research program that investigates several aspects for the automatic organization of news stories in textual area. TDT encompasses several tasks, such as topic detection, topic tracking, first story detection and so on. The FSD task is to detect the first story that discusses a previously unknown event. Some novelty/redundancy measures are motivated by work for FSD. A common solution to FSD is comparing news stories to clusters of stories previously known events. An incoming story is either marked as old if the similarity between the story and the closest cluster is above a certain similarity threshold, or marked as the first story if it is novel enough.

2.2. Multimedia information retrieval

Previously clip and video copy detection (e.g., [12,13]) were investigated by using image similarity measure with low-level global features, for instance, color histograms. Fast signature-based methods (e.g., [12]) were proposed to identify similar clips, which are global statistics of low-level features in clips. Global signatures are suitable for matching clips with almost identical content, and incompetent for changes due to compression, formatting, or minor editing in spatial or temporal domain. Clip similarity ranking [13] was built on top of the shot similarity and combined temporal order, granularity and so on. Bipartite graph based algorithms were proposed in [14] to compare the similarity of two clips. However, previous clip or video based similarity approaches ignored the interrelationship between audio contents and visual information. Furthermore, shot similarity detections built on global features are not robust enough for news story similarity measure due to the complicated variations of keyframes.

In news videos, news stories are often accompanied by short video shots that tend to be repeated during the course of the topic. Traditionally, a shot is usually represented by a keyframe. There are a significant number of near-duplicate keyframes (shots) existed in news stories. *Near-duplicate keyframes* (NDK) are close to the exact duplicate of each other, but different in the capturing conditions (camera, camera parameter, view angle, etc.), acquisition times, rendering conditions or editing operations [15]. Detection and retrieval of near-duplicate keyframes is very critical in novelty/redundancy detection and topic threading. The methods to detect the near-duplicate keyframes were proposed in [15–18]. Duygulu et al. [16] presented the technique to detect and track the repeated sequence of shots. Different from global features, interest point based local feature detection approaches [17,18] avoided the shortcoming of global features and achieved a good and robust detection result. Hsu et al. [19] tracked four topics with visual duplicates and semantic concepts, and found that near-duplicates significantly boosted the tracking performance. Recently, near-duplicate keyframes were also exploited in [20] to boost the performance of interactive video search.

News videos provide richer information than text streams. For novelty/redundancy detection and topic threading in news videos, either pure textual or visual method may overlook the interactions between textual and visual information. Both textual and visual concepts should be incorporated to generate a more robust and reasonable result. Zhu et al. [21] presented a hierarchical video content description and summarization strategy supported by a joint textual and visual similarity. Zhai et al. [22] linked news stories by combining keyframe matching and textual correlation. Cosine Distance measure combined the text and human labeled NDK was applied to thread the news topics in our early work [23]. However, current approaches do not fully consider the textual and visual concepts in news stories to detect the novelty and redundancy, and the performance on cross-lingual environment is not sure, which motivates our research to systematically discuss the cross-lingual news story similarity measure.

3. Framework

The framework of cross-lingual news story similarity measure is shown in Fig. 2, in which the novelty and redundancy detection module is the core part in the whole framework. News videos from multiple sources with different languages constitute the news database. Topics are formed by clustering algorithms or classifiers. News stories related to a specific topic are presented in a chronological order based on their publication time and each one must be evaluated before the next is seen. For simplicity, we assume all news stories in the topics are relevant, and the earliest new story that appeared in this topic is the first story of the topic. News stories are meaningful units which include multimodal features (audio and visual contents) related to events. A news story consists of a sequence of shots in the visual track, and a set of words and named entities extracted from speech transcript in the audio track. If a news story from other languages except English is appeared, its text transcripts and named entities are translated into English counterparts by machine translation. For visual track, a news story can be regarded as a group of shots depicting an event. Usually, a representative keyframe is extracted to represent each shot, and thus a story can be viewed as a list of shots represented by keyframes in the visual track. Near-duplicate keyframe detection is performed to check the visual similarity among keyframes. If keyframes are also appeared in other places of the whole corpus, they will be regarded as near-duplicate keyframes (NDK). Otherwise, they are non-near-duplicate (non-NDK). Finally, these multimodal features extracted from textual and visual information (text, named entities, NDK and non-NDK) are send to the novelty and redundancy detection module to evaluate its redundancy by comparing it with previously broadcast news stories. In this paper, we mainly focus on the novelty and redundancy detection.

4. Multimodality representation for similarity measure

To compare two news stories, two fundamental issues need to be addressed: (1) news story representation; (2) news story similarity measure.

Previous novelty and redundancy detections are based on text. But for news videos, visual shots provide critical and useful information. Textual contents and visual contents complement each other. Fig. 3 shows two stories on "Arkansas school shooting" discussed different themes but with similar textual concepts. Most of the textual contents of the first story are covered by the second story. The similar keywords are bolded in Fig. 3. However, from the visual track, the shots of two stories are totally different. In essence, the themes of these two stories are not similar. The first story describes "Students are back in class", while the second introduces "The funerals for two students killed in the shooting". Since both stories have different themes, the second story should be treated as novel story. However, text-based redundancy detection regards the second story as somewhat redundant or even completely redundant story depending on different threshold settings, owing to the fact that important concepts and named entities such as "funeral" and "Paige Herring" are concealed by a list of other keywords. Visual contents, on the other hand, provide a new set of keyframes that allow the discrimination of themes in both stories.

Intuitively, named entities would be informative for differentiating topics. News stories can be represented by four

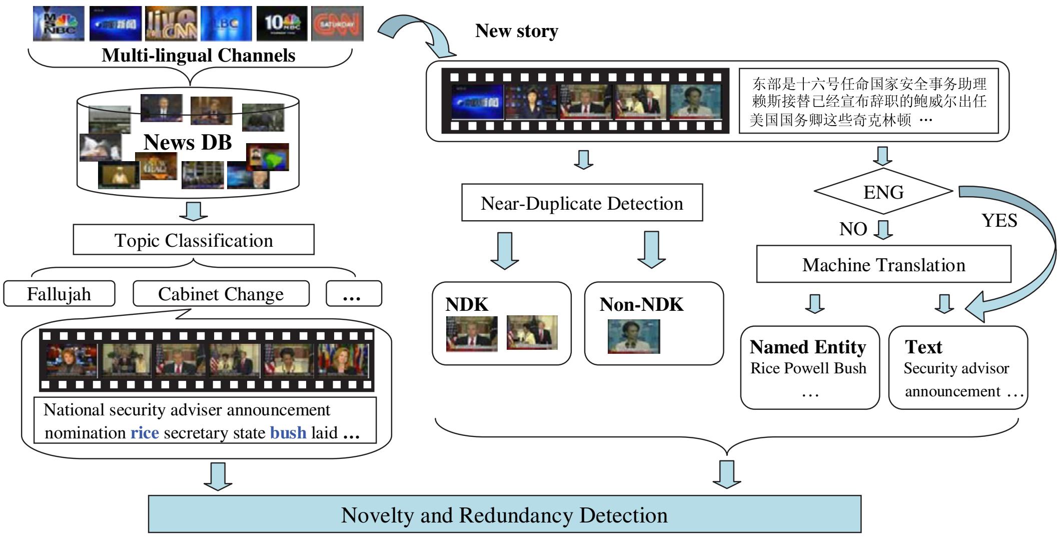
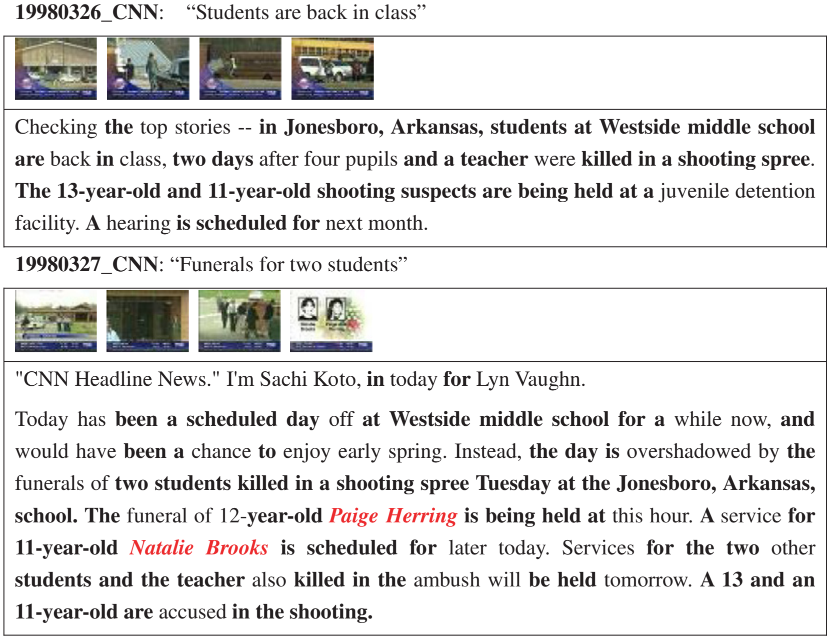

Multi-lingual Channels
News DB
Topic Classification
Fallujah
Cabinet Change
...
National security adviser announcement nomination rice secretary state bush laid ...
New story
东部是十六号任命国家安全事务助理赖斯接替已经宣布辞职的鲍威尔出任美国国务卿这些奇克林顿 ...
Near-Duplicate Detection
NDK
Non-NDK
ENG
NO
YES
Machine Translation
Named Entity
Rice Powell Bush
...
Text
Security advisor announcement ...
Novelty and Redundancy Detection

Fig. 2. Similarity measure in multilingual multimedia environment.



19980326_CNN: “Students are back in class”

Checking **the** top stories -- **in Jonesboro, Arkansas, students at Westside middle school are** back **in** class, **two days** after four pupils **and a teacher** were **killed in a shooting spree**. **The 13-year-old and 11-year-old shooting suspects are being held at a** juvenile detention facility. **A** hearing **is scheduled for** next month.

19980327_CNN: “Funerals for two students”

"CNN Headline News." I'm Sachi Koto, **in** today **for** Lyn Vaughn.

Today has **been a scheduled day** off **at Westside middle school for a** while now, **and** would have **been a** chance **to** enjoy early spring. Instead, **the day is** overshadowed by **the** funerals of **two students killed in a shooting spree Tuesday at the Jonesboro, Arkansas, school. The** funeral of 12-**year-old** ***Paige Herring*** **is being held at** this hour. **A** service **for** **11-year-old** ***Natalie Brooks*** **is scheduled for** later today. Services **for the two** other **students and the teacher** also **killed in the** ambush will **be held** tomorrow. **A 13 and an 11-year-old are** accused **in the shooting.**

Fig. 3. Two stories under the topic “Arkansas school shooting” discuss different themes but with similar textual contents.

kinds of information: who (persons), when (time), where (locations) and what (keywords). Novel information is often conveyed through the introduction of new named entities, such as the names of people, organizations and places, e.g., “Paige Herring”, “Natalie Brooks” in Fig. 3. Their effects for news story novelty/redundancy have not been examined.

Visual information, in particular near-duplicate keyframes (NDK), provides reliable measure to compare the similarity among stories across sources, time and languages. It is especially useful when the text transcript is very noisy or the transcripts are not available when appropriate speech recognition and machine translation tools are absent. NDK pairs, intuitively, provide strong similarity measure for news stories with similar visual contents. A statistic in [15] indicated that there are approximately 10–20% of NDK in the news stories. Three sets of near-duplicate keyframes appeared in three news stories are shown in Fig. 4 with different color borders. Among them, the first two stories are from English channels and the third one is from Chinese channel.

We regard news story as a bag of features. At here features means words, named entities or keyframes. A news story is represented as a vector of features or as a smoothed probability distribution over all the features.

Similar to documents which are treated as a bag of words, we can regard news story is composed of a bag of keyframes in the visual track. And the keyframes are further classified as near-duplicate keyframes (NDK) appeared multiple times within or across other news stories and non-near-duplicate keyframes (non-NDK) that appeared once in the whole corpus. Keyframes can be

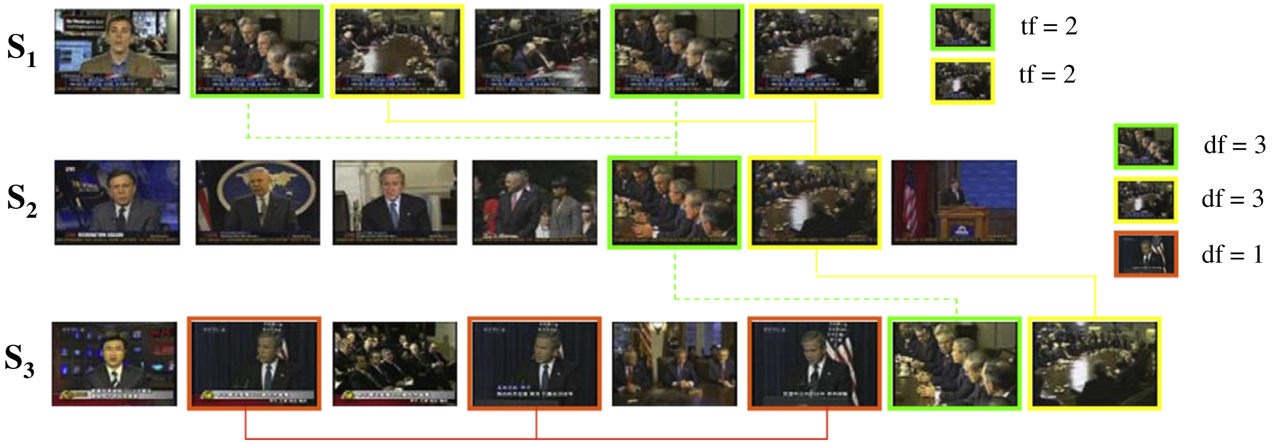


Fig. 4. Near-duplicate keyframes appeared in three news stories of different channels.

seen as a special kind of words. Therefore, the techniques used in the text area can be applied to the keyframes. We define the frequency of NDK in one news story as term frequency (*tf*), and the frequency that NDK appeared in different news stories as the document frequency (*df*). For example, the term frequency (*tf*) of NDK with red border in story S_3 (i.e., the keyframes that Bush is addressing) is 3 because it appeared three times in story S_3. Its document frequency (*df*) is one since these keyframes do not appear in other news stories. For non-NDK, the term frequency and document frequency are both 1.

5. Similarity measures

Different people have different definitions of redundancy and different redundancy thresholds. In order to alleviate such inconsistency, similar to [4], we classify news stories into three classes in this paper: *completely redundant*, *somewhat redundant* and *novel*. If a news story contained no new information, this news story is regarded as *completely redundant*. A news story which is a review or a near-duplicate news story with very minor adjustment of previously delivered news story is completely redundant news story. Only when both visual (shot) and textual contents of a news story are totally covered by previous stories, it is treated as completely redundant. Those news stories that have some new information and contain many redundant contents are marked as *somewhat redundant* news stories. Somewhat redundant stories usually convey the gradual development of a theme. For news stories where most of their contents are new, they are marked as *novel* stories which indicate the emergence of new themes.

In this section, we discuss different measures on vector space model and language model to detect the redundancy of news stories. *Set Difference* is a set-oriented measure. *Cosine Distance* is based on the vector space, while *Language Model* with the KL divergence is designed to measure the feature distribution. These measures are applied to different features (text, named entities and keyframes) and finally the fusion of these features is explored. In these approaches, it is based on a pairwise similarity measure between the current story and each previously broadcast story in the topic.

The redundancy of news story S_i is computed through a pairwise comparison between S_i and every previously seen news story S_j. The previously seen news story that is most similar to S_i determines the redundancy measure of S_i:

$$R(S_i \mid S_1, \ldots, S_{i-1}) = \max_{1 \leqslant j \leqslant i-1} R(S_i \mid S_j)$$

We treat novelty and redundancy as opposite ends of a continuous scale. Therefore, ranking the news stories by increasing redundancy score is equivalent to ranking them by decreasing novelty score [4].

5.1. Cosine Distance

The *Cosine Distance* metric is a popular similarity measure in information retrieval. The cosine of the angle between a news story vector and each previously delivered news story vector determines the redundancy score for that news story. It is a pairwise measure, defined by:

$$R_{\mathrm{cd}}(S_i \mid S_j) = \frac{\sum_{k=1}^{m} f_k(S_i) f_k(S_j)}{\sqrt{\sum_{k=1}^{m} f_k(S_i)^2 \sum_{k=1}^{m} f_k(S_j)^2}}$$

where $f_k(S_i)$ is the weight for feature f_k in story S_i. The weighting function used in our experiments is a tf-*i*df function specified by the following formula:

$$f_k(S_i) = \frac{\mathrm{tf}(f_k, S_i)}{\mathrm{tf}(f_k, S_i) + 0.5 + \left(1.5 * \frac{\mathrm{len}(S_i)}{asl}\right)} * \frac{\log \frac{n+0.5}{\mathrm{sf}_k}}{\log(n+1.0)}$$

$\mathrm{tf}(f_k S_i)$ is the term frequency of feature f_k in story S_i, *asl* is the average number of features in a story for the topic, sf_k is the document frequency of feature, $\mathrm{len}(S_i)$ is the number of features in the story and n is the number of stories for the topic. In our experiments, n and sf_k are computed incrementally based on the stories already in the stream.

5.2. Set Difference

The *Set Difference* redundancy of a news story is measured by the following formulation:

$$\begin{aligned} R_{\mathrm{sd}}(S_i \mid S_1, \ldots, S_{i-1}) &= \max_{1 \leqslant j \leqslant i-1} R_{\mathrm{sd}}(S_i \mid S_j) \\ &= \max_{1 \leqslant j \leqslant i-1} \left(\frac{\mid F_{S_i} \cap F_{S_j} \mid}{\mid F_{S_i} \mid} \right) \end{aligned}$$

The redundancy of a story is measured by the proportion of features in new story also appeared in the previously delivered story. The intuition is that two stories are likely to be similar to some extent if they have many features in common. Normalization by story length is essential because the measure will tend to rise with length without normalization. The longer the story is, the higher the chance it contains more features.

5.3. Language Models

In the *Language Model* approach, a document is represented by a unigram feature distribution θ. A distributional similarity measure, *Kullback–Leibler divergence*, is commonly used to measure the similarity between two stories:

$$R(S_i \mid S_j) = -\mathrm{KL}(\theta_i, \theta_j) = -\sum_{f_k} p(f_k \mid \theta_i) \log\left(\frac{p(f_k \mid \theta_i)}{p(f_k \mid \theta_j)}\right)$$

where θ_i is the language model for story S_i, which is a multinomial distribution.

The simplest way to estimate $p(f_k|S_i)$ is the *maximum likelihood estimation* (*MLE*), simply given by relative counts:

$$p(f_k \mid S_i) = \frac{\mathrm{tf}(f_k, S_i)}{\sum_{f_k} \mathrm{tf}(f_k, S_i)}$$

However, the problem of maximum likelihood estimation is that it will generate a zero probability if a feature never occurs in the story S_i, which will cause $\mathrm{KL}(\theta_i, \theta_j) = \infty$.

Smoothing techniques are used to assign a non-zero probability of the unseen features and improve the accuracy of feature probability estimation. Prior research [2,4,24] showed that different smoothing methods highly affect the performance. For language model, we mainly use *Bayesian* smoothing with *Dirichlet* priors and *Shrinkage*. Furthermore, *mixture model* is also tried.

5.3.1. Dirichlet smoothing

This smoothing technique uses the conjugate prior for multinomial distribution, which is the Dirichlet distribution. It automatically adjusts the amount of reliance on the features according to the total number of the features. For a Dirichlet distribution with parameters:

$$(\mu p(f_1 \mid C), \mu p(f_2 \mid C), \ldots, \mu p(f_n \mid C))$$

the posterior distribution using Bayesian analysis is:

$$p_\mu(f_k \mid S_i) = \frac{\mathrm{tf}(f_k, S_i) + \mu p(f_k \mid C)}{\sum_{f_k} \mathrm{tf}(f_k, S_i) + \mu}$$

$p(f_k|C)$ is the collection language model and μ is a parameter learned from experiments.

5.3.2. Shrinkage smoothing

Shrinkage smoothing is a special case of Jelinek–Mercer smoothing method, which involves a linear interpolation of the maximum likelihood model with n-gram model [24]. Based on the assumption that a story is generated by sampling from three different language models: a story model, a topic model, and a model for the collection, the language model of a story is determined by:

$$p(f_k \mid \theta_\mathrm{S}) = \lambda_\mathrm{S} p(f_k \mid \theta_{\mathrm{ML_S}}) + \lambda_\mathrm{T} p(f_k \mid \theta_{\mathrm{ML_T}}) + \lambda_\mathrm{C} p(f_k \mid \theta_{\mathrm{ML_C}})$$

using coefficients λ_S, λ_T and λ_C to control the influence of each model, where $\lambda_\mathrm{S} + \lambda_\mathrm{T} + \lambda_\mathrm{C} = 1$.

$\theta_{\mathrm{ML_T}}$ is the maximum likelihood language model of the topic and $\theta_{\mathrm{ML_C}}$ is the maximum likelihood language model of the collection. In our experiments, the topic model is built on all presumed relevant stories for the topic and collection model is built on all the stories in the corpus.

5.3.3. Mixture Model

A story is generated by the mixture of three different language models: a story-specific model θ_S, a topic model θ_T, and a model for the collection θ_C. Mixture model is based on the opposite assumption that features occurred frequently in a story than in the background should have higher probability in the story model. Therefore, the approach is to deduce the maximum likelihood story model, which is compared pairwise to each previously seen story model. Each feature in the story is generated by each of the three language models with probability λ_S, λ_T, and λ_C, respectively

$$p(f_k \mid \theta_{\mathrm{ML}_x}) = \lambda_\mathrm{S} p(f_k \mid \theta_\mathrm{S}) + \lambda_\mathrm{T} p(f_k \mid \theta_{\mathrm{ML_T}}) + \lambda_\mathrm{C} p(f_k \mid \theta_{\mathrm{ML_C}})$$

where $\lambda_\mathrm{S} + \lambda_\mathrm{T} + \lambda_\mathrm{C} = 1$.

To note, although equations of Shrinkage smoothing and mixture model look similar, the model acquired and used to calculate KL divergence is different. Shrinkage smoothing increases the probability of features that occur frequently in the topic or in the collection if they occur less frequently in story, while mixture model decreases the probability of these features [4].

Similar to [6], the language model θ_S that maximizes the likelihood of the observed story, given fixed parameters, was computed using the technique described in [25].

5.4. Multimodality fusion

In the proceed sections, we discuss different measures on individual feature to measure the similarity. However, for news stories, the employment of either textual or visual features may not be enough since either feature can appear differently over time. Textual and visual concepts complement each other. The pure textual method may overlook the interactions between textural and visual information, i.e., the visual contents determine the set of shots on which text summarization will be considered, but the textual information does not have a say about how the set of shots is selected. Robust and reasonable approaches should combine both textual and visual features to determine the degree of redundancy in news stories while exploiting the significance of these features. There are the cases that news stories have dissimilar text features but carry similar visual information, or they have similar textual content while totally novel visual cues. Fusing multiple modalities can adjust their redundancy scores. For two stories having similar visual NDK but dissimilar text features, although the visual similarity is high, the redundancy score after fusing multiple modalities will be relatively small, and vice verse.

Motivated by this fact, the following measure which integrates the textual and visual features is proposed to detect the novelty/redundancy. We use a linear weighted fusion method for combining the similarity scores from different modalities: text (T), named entities (NE) and keyframes (KF). Linear fusion model has been shown to be one of the most effective approaches to fuse textual and visual modalities in video retrieval. The similarity score is defined as:

$$R(S_i \mid S_j) = w_\mathrm{T} R_\mathrm{T}(S_i \mid S_j) + w_\mathrm{NE} R_\mathrm{NE}(S_i \mid S_j) + w_\mathrm{KF} R_\mathrm{KF}(S_i \mid S_j)$$

The measure of text, named entity and keyframe can be any method mentioned in previous sections.

The weights w_T, $\mathrm{w_{NE}}$ and w_KF are used to control the influence of each feature. The linear weights among modalities are determined empirically.

6. Experiments

6.1. Dataset

We selected all Chinese and English videos from TRECVID-2005 cross-lingual news video corpus [26] as our dataset which includes 88 Chinese and 142 English news videos from five different sources (CCTV4, NTDTV, CNN, NBC and MSNBC). Arabic news was ignored in our experiments because the assessors are not familiar with Arabic. The time span is from October 30 to December 1, 2004. Due to lack of official story boundaries, we used CMU's story boundaries [27] to segment the video streams into stories. The textual features are a list of words extracted from speech transcripts by an automatic speech recognition system (ASR) at LIMSI [28], while the visual concepts are the set of representative keyframes extracted from video corpus. Note that each shot is represented by one keyframe. The representative keyframes of shots are given. There are 42 videos having no ASR or corresponding machine translation. We supplemented them with CMU Informedia [27] speech recognition software and translated them with Google translation [29] so that all stories have English speech transcripts. Three types of named entities (NE): person, organization and location were extracted by BBN's software. We did not use specific feature selection in our experiments. In Zhai's work of language model on text [24], they did not do the feature selection, even the stop-word removal, because they believed that the effects can be achieved by language model techniques. Similar to their work, we did not do the feature selection independently. Potentially, language models and Cosine Distance with Okapi formation can achieve this objective to some extent. Unimportant and infrequent features will have little contribution to the similarity measure while popular features appeared frequently in topic or corpus will have less weight accordingly. After data preprocessing such as word stemming and stop-words removal, the dataset consists of 1334 news stories, 12,428 unique words, 3644 named entities and 19,621 keyframes.

The shots with an anchor person were automatically detected and removed from the keyframe list. To detect the near-duplicate keyframes, the local interest points of each keyframe were extracted by Lowe's DoG detector [30] and described by PCA-SIFT [17], which is a 36 dimensional vector for each interest point. One-to-one symmetric matching method [18] based on local interest points was used to detect whether two keyframes are near-duplicate. In our experiments, it will be treated as a NDK pair if there are 10 matching pairs of local interest points between two keyframes and their average similarity is above 0.9. The whole near-duplicate keyframe list was generated by transitive closure based on the information of every two keyframes. There are 8179 near-duplicate keyframes in our test set, which form 2261 groups. In each group, the near-duplicate keyframes are very similar.

Without the official annotation of topics, we built a ground truth table by manually labeling stories according to topic themes. To ensure the fairness of comparison, the topics that have less than four news stories were regarded as outliers and removed. The topics were manually labeled and annotated. There are 33 topics. The detailed information is listed in Table 1. A small number of stories may belong to multiple topics. Among them, stories of 10 topics are only reported by English channels, while six topics are only broadcast in Chinese channels.

To analyze the performance of the novelty/redundancy detection scheme, two untrained non-expert assessors were asked to watch stories and to judge one topic at a time. News stories in each topic were ordered chronologically. The assessors were requested to label the news stories with a judgment (completely redundant, somewhat redundant or novel). If the assessors believe the textual and visual contents of a news story are totally covered by previously news stories, the story is labeled as completely redundant. A

Table 1
Topic information

ID	Topics	#	CR	SR
English + Chinese				
1	APEC summit	37	5	7
2	Arafat health	155	22	42
3	Black friday	18	1	4
4	Cabinet changes	78	7	35
5	President election	184	12	20
6	War on Fallujah	203	13	64
7	AIDS	27	6	2
8	Afghan hostage	11	4	3
9	Iraq problem	158	12	21
10	Korean nukes problem	21	2	5
11	Clinton library	13	1	6
12	Iran nukes	39	0	19
13	Mideast peace	70	9	11
14	Bush second term plan	31	2	6
15	War on terrors	27	1	12
16	Thanksgiving	22	2	1
17	Ukraine crisis	66	9	20
English				
1	Intelligence reform bill	22	0	6
2	Ebersol plane crash	9	1	6
3	Bush and Blair	17	2	6
4	Bush visited Canada	10	1	4
5	CIA in turmoil	5	0	2
6	NBA brawl	32	0	21
7	Scott Peterson trial	30	0	12
8	Vioxx	10	0	3
9	Vice president's health	8	0	5
10	Veteran's day	17	2	4
Chinese				
1	Mine bombing	14	4	6
2	China and Eastern Union	24	6	3
3	Hu Jintao visited South America	35	10	5
4	WTO	11	2	2
5	Falun Gong	25	4	2
6	Yunnan air crash	7	0	0
	Total	1436	140	365

#, number of stories; CR, number of completely redundant stories; SR, number of somewhat redundant stories.

story carries new information even though parts of the story are redundant with previous stories, which is regarded as somewhat redundant. For news stories which have similar visual shots but describe new textual information, or deliver different scenes but with similar textual contents, they are somewhat redundant stories. Other stories that convey fresh development of the topic are novel.

6.2. Performance metric

To factor out the effect of the redundancy threshold, we evaluate the effectiveness of redundancy detection by comparing the average precision and recall figures for redundant stories. News stories are ranked by their redundancy scores. *Precision* and *recall* are commonly used metrics. Let $\boldsymbol{G}$ be the ground truth set of redundant stories and $\boldsymbol{D}$ be the detected one. Recall and Precision are defined as:

$$\text{Recall} = \frac{|G \cap D|}{|G|} \quad \text{Precision} = \frac{|G \cap D|}{|D|}$$

Moreover, the dataset demonstrates a considerable variance in the distribution of the topic size. The number of stories can be larger than 150 for hot topics, and smaller than 10 for unpopular topics, which is quite skewed. To avoid the results being biased by the size of topics, we also collect the average precision for topics across all recall level (0.1–1.0). The average precision is defined as:

$$\text{AP} = \left(\sum_{k=1}^{10} \text{Precision}_k\right) \Big/ 10$$

where Precision_k is the precision when the recall is equal to $k \times 0.1$.

6.3. Performance comparison for measures

To compare the performance of redundancy detection, we compared Cosine Distance (Cosine), Set Difference (Set_Diff), language model with Dirichlet smoothing (LM_D), with Shrinkage smoothing (LM_S) and mixture model (LM_M). All these measures are applied to text (T), named entities (NE) and keyframes (KF). Each method is represented by Measure_Feature pair, for example, LM_D_KF denotes this measure is based on language model with Dirichlet smoothing and operates on keyframes. A redundancy score is calculated for each news story by comparing it with preciously delivered stories. The Recall–Precision graphs over the set of redundant stories are demonstrated in Fig. 5 and the average precision is shown in Fig. 6. Since the number of stories in each topic is different, we also tested the performance weighted by the number of redundant stories. The overall performance is similar to the one without weight, so we do not show the results due to space limitation.

For the performance of completely redundant news stories (Figs. 5(a) and 6(a)), text-based measures have better performance than keyframe based ones, while the named entity based approaches perform the worst. The near-duplicate keyframe detection affects the performance of keyframe based methods. There is the possibility that a set of near-duplicate keyframes was wrongly detected as two or more sets. That is, they should be regarded as one concept, but were falsely treated into multiple concepts. And there are also the cases that near-duplicate keyframes were falsely detected or neglected. Therefore, the performance of keyframes based approaches is not good as text-based measures to detect the completely redundant stories. Named entities are not informative enough for comparing the similarity of two stories. Two news stories with the similar named entities are not a strong indication that they are the completely redundant stories.

For the performance of measures, Cosine Distance is still a robust measure, similar to previous work. Language models show their effectiveness, which have better or

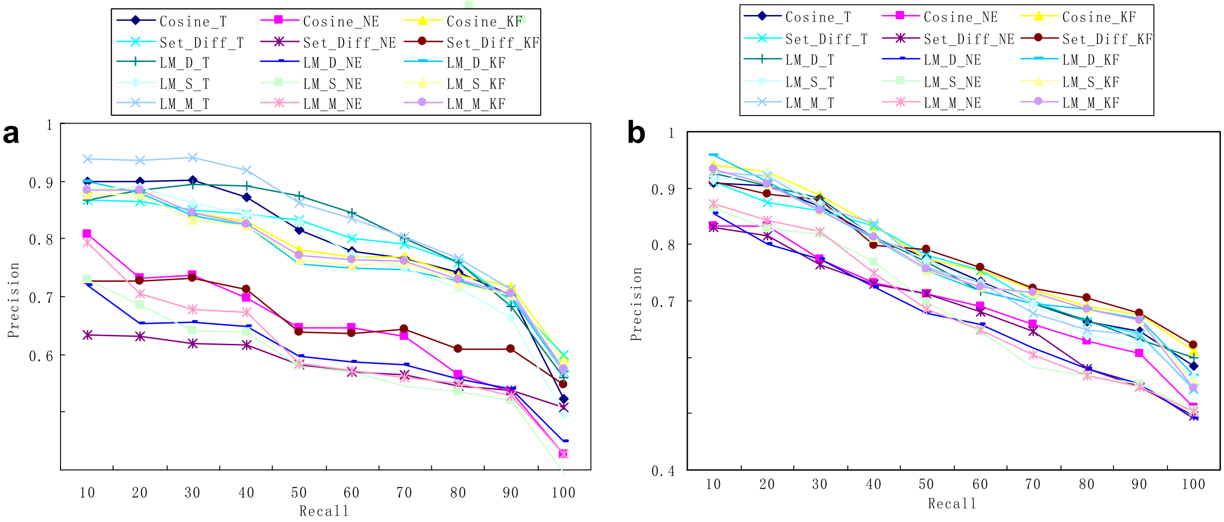
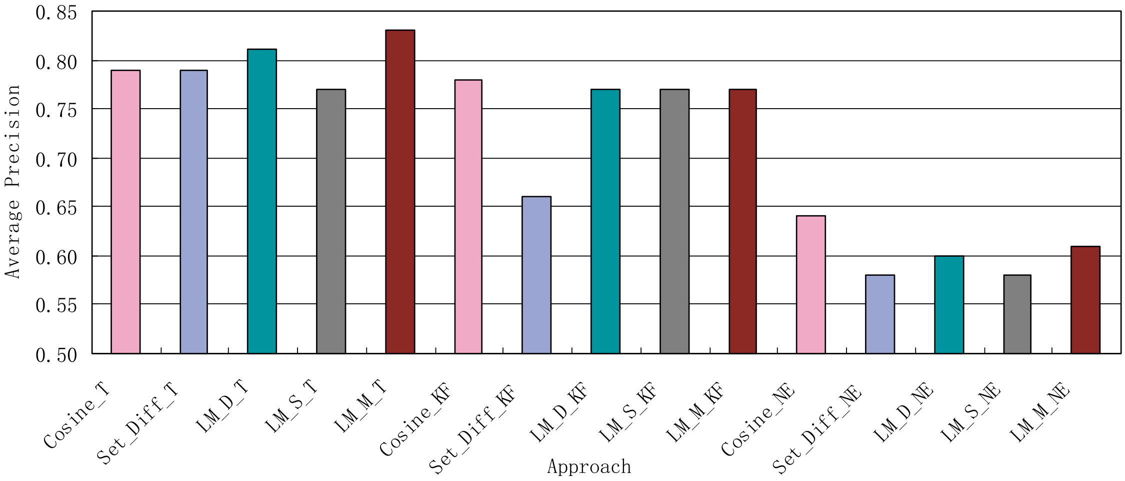
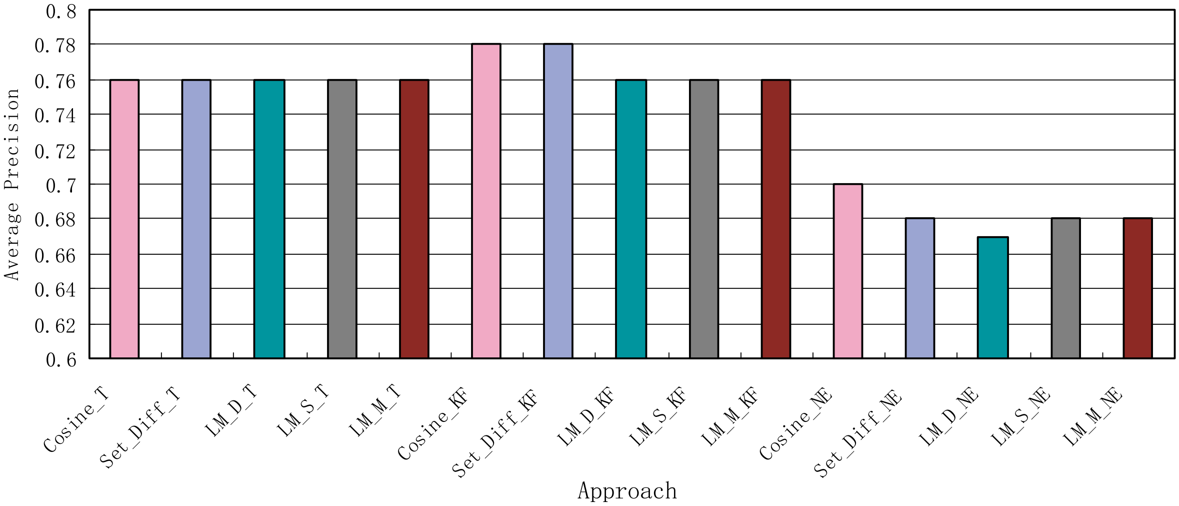


Fig. 5. Performance comparison on different redundancy measures. (a) News stories are considered redundant if assessors marked it *completely redundant*. (b) News stories are considered redundant if assessors marked it *completely redundant* or *somewhat redundant*.



(a) News stories are considered redundant if assessors marked it *completely redundant*



(b) News stories are considered redundant if assessors marked it *completely redundant* or *somewhat redundant*

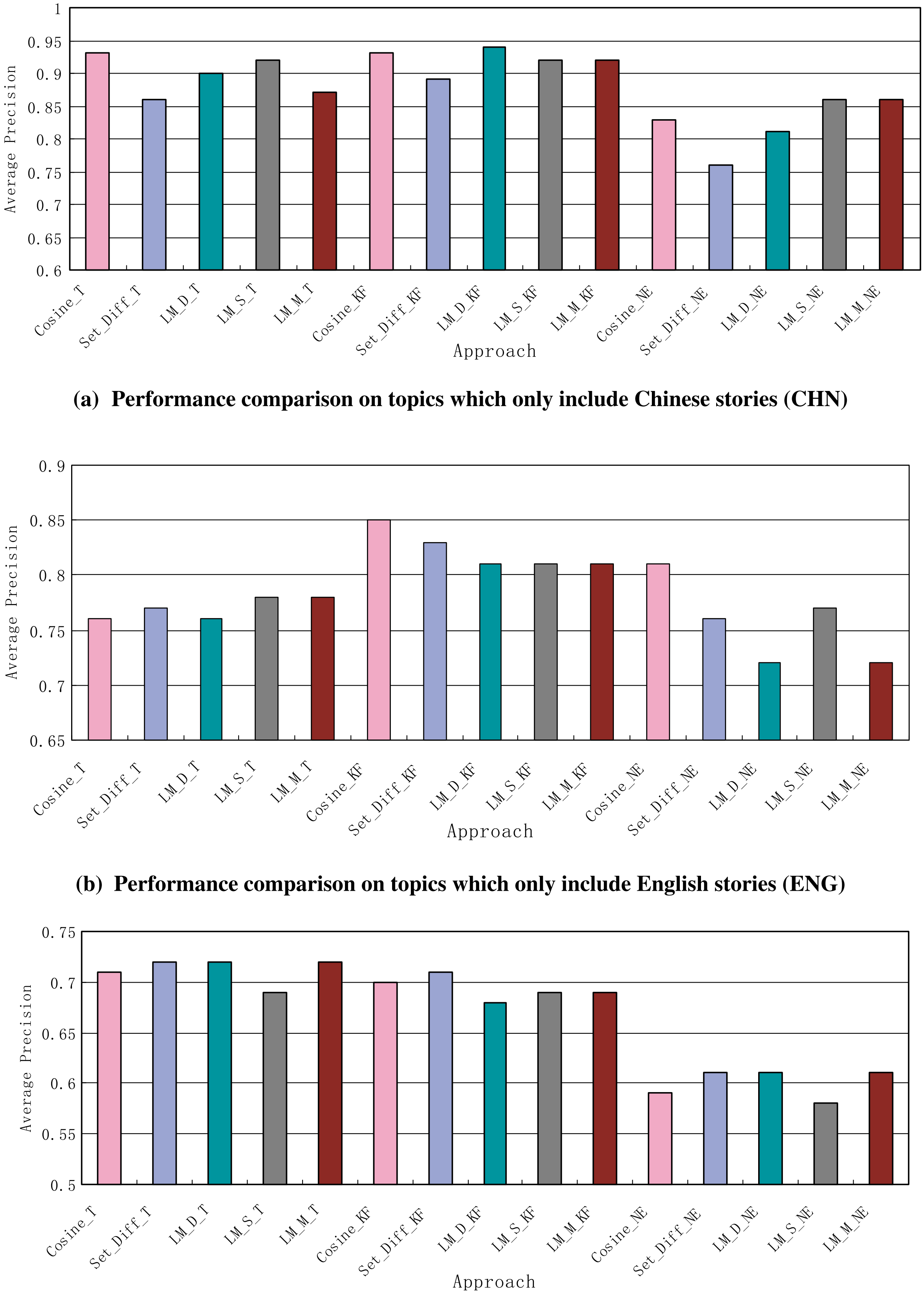
Fig. 6. Overall performance comparison on different redundancy measures (Cosine Distance, Set Difference, Dirichlet smoothing, Shrinkage smoothing, Mixture Model).

similar performance as Cosine Distance. It also demonstrates that smoothing techniques play an important role in probability estimation. For the textual measures, mixture model (LM_M_T) achieves the best performance, which shows that mixture model is more accurate than other smoothing approaches. And language model with Dirichlet smoothing (LM_D_T) is better than Cosine Distance and Set Difference.

For the performance of somewhat and completely redundant stories (Figs. 5(b) and 6(b)), the difference is not so conspicuous. Named entity is still the worst. However, different from Fig. 5(a), the performance on keyframes outperforms the text-based methods. Cosine_KF and Set_Diff_KF show their effectiveness. For completely redundant stories, they are usually come from the same channels, which are relatively easy to detect. However, for somewhat redundant stories, they are commonly across different channels, which are more challenging. The errors from speech recognition and machine translation make the text-based approaches less effective. While visual similarity between two news stories gives important clues to judge the similarity between them. Therefore, keyframe based measures perform better.

6.4. Cross-lingual performance comparison

To study the effects of different measures on cross-lingual stories, we collect the results on different topics that only include Chinese news stories (CHN), only include English news stories (ENG), and include both Chinese and English stories (CHN + ENG). The topic information is also listed in Table 1. The performance is shown in Fig. 7. The completely redundant and somewhat redundant stories are regarded as redundant stories.



(a) Performance comparison on topics which only include Chinese stories (CHN)

(b) Performance comparison on topics which only include English stories (ENG)

(c) Performance comparison on topics which include both Chinese and English stories (CHN + ENG)

Fig. 7. Performance comparison for cross-lingual measure for completely and somewhat redundant news stories: general performance on mono-lingual topics is better than multilingual topics (Cosine Distance, Set Difference, Dirichlet smoothing, Shrinkage smoothing, Mixture Model).

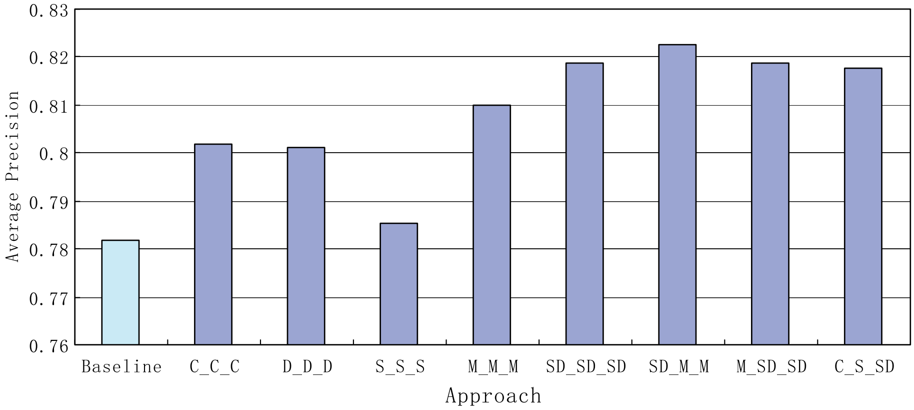


Fig. 8. Performance comparison on multimodality fusion (text, named entities and keyframes, represented by $\text{Measure}_{\text{T}}$_$\text{Measure}_{\text{NE}}$_$\text{Measure}_{\text{KF}}$ pair).

For Chinese topics (Fig. 7(a)), LM_D_KF achieves the best performance, and Cosine_T and Cosine_KF are followed. Overall, the visual information, especially the NDK, gives important indication of the redundancy. Therefore, the measures based on keyframes have better performance than measures on text. While completely redundant news stories are mainly from the same channels, they usually have similar words. So the performance based on texts is also good. For English topics (Fig. 7(b)), the number of completely redundant stories is very small. The keyframe based measures outperform text-based approaches. Cosine_KF has the best performance and Cosine_NE shows good performance.

For cross-lingual topics which include both Chinese and English stories (Fig. 7(c)), the cross-lingual performance is a little worse than the mono-lingual cases. For mono-lingual topics, they usually use the same set of words to describe the similar concepts. Even if speech recognition and machine translation falsely detected or translated a word into an unknown word or another word, they can be correctly matched when comparing the similarity. However for cross-lingual topics, different broadcast videos may use diverse sets of words to describe the topic. When facing the problem of errors caused by speech recognition and machine translation, it is not powerful enough to resolve it. Therefore, the performance of cross-lingual is worse than the mono-lingual cases.

For cross-lingual topics, the text-based approaches achieve better performance than keyframe based methods. The named entities are not robust enough. However, the difference among measures of text and keyframes is not big. Although the text information might be noisy, it can still detect the redundant stories. The editing for keyframes from different channels (e.g., logo insertion, editing style) may affect the detection of NDK.

6.5. Performance comparison for fusion

Because news videos convey information both at audio track and visual track, similarity measure on either textual or visual features may not be enough. A robust way is to take into account both textual and visual features while exploiting the significance of these features. In this section, we experiment various combinations of text, named entities and keyframes under five different measures (Cosine Distance, Set Difference and three language models). So totally there are 125 combinations ($5 \times 5 \times 5 = 125$). Due to the large number of combinations, we randomly compare different combinations to study the improvement by fusing text, named entities and keyframes together. The measures are represented by $\text{Measure}_{\text{T}}$_$\text{Measure}_{\text{NE}}$_$\text{Measure}_{\text{KF}}$ pair. For example, M_SD_SD denotes that this method uses mixture model to measure the text and Set Difference to measure the named entities and keyframes, and then combine the results together to get the final results. We select the measure with the best performance in Fig. 5(b) (i.e., Cosine_KF) as the baseline to compare the improvement of fusion.

From Fig. 8, we can see that fusion of text, named entities and keyframes improves the performance substantially. They achieve around 5% improvement on average over the baseline, and are much higher than the individual measure before fusion. Textual features and visual features complement each other. Especially for other stories with similar textual concepts but different visual information, or with similar visual appearance but inconsistent contents, fusing multimodalities can obtain a reasonable result. Therefore after fusion, they improve the performance. Although the overall performance of named entities is poor for individual measure, named entities have a small contribution for fusion. Due to the space limitation, detailed experiments are not shown in this paper.

7. Conclusion

News story similarity measure is a substantial task for news story search, retrieval, and tracking. In this paper, we explore different similarity measures to compare cross-lingual broadcast news videos, which include vector space models and language models on multimodalities (text, named entities, and keyframes). Experiments have been done on cross-lingual news video corpus TRECVID-2005 to study the performance of modalities and measures for novelty and redundancy detection.

We get some useful lessons from the experiments:

- In general, Cosine similarity is still a stable measure that has been demonstrated over many tasks. Language models are promising because they are built on the basis of probability estimation, but depend on the smoothing techniques.
- Approaches with texts and keyframes have different effects for the novelty and redundancy detection. However, named entities are not informative enough for comparing the similarity, which perform the worst.
- Languages models on text are appropriate for detecting completely redundant stories while Cosine Distance and Set Difference on keyframes are suitable for somewhat redundant stories.
- Detection results on mono-lingual topics have better performance than multilingual topics.
- Multiple modalities (text, keyframes and named entities) complement each other. Fusing different modalities substantially boost the performance over the approaches with single modality.

In this paper, we built language models based on keyframes, which is especially useful for environments where text transcripts are not available or speech recognition and machine learning tools are absent (e.g., videos on web). However, the number of keyframes in each story is relatively small compared to the number of words, and the accuracy of near-duplicate detection affects the performance. In the future, we will explore the so-called visual language model in depth with different threshold settings and even at the visual keyword level (local interest points). Currently, novelty and redundancy detection is relatively computationally expensive, especially for NDK detection. We will use effective strategies [31] to accelerate the process, including construction of visual vocabulary, building fast indexing structure, and the heuristic policies to reduce the number of keyframe pair comparison. Furthermore, for cross-lingual news videos, the transcripts and named entities after speech recognition and machine translation are inaccurate. It is worth investigating in the future that how to combine multiple sources to improve the performance, for example, Google online translation, SYSTRAN translation, and NIST named entity dictionary. Another way is using web to translate multilingual named entities and out-of-vocabulary words.

Acknowledgments

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 118906), and a grant from City University of Hong Kong (Project No. 7002112). We thank Wan-Lei Zhao for the work of near-duplicate keyframe detection.

References

[1] T.S. Chua, S.F. Chang, L. Chaison, W. Hsu, Story boundary detection in large broadcast news video archives – techniques, experience and trends, in: ACM MM'04, 2004, pp. 656–659.

[2] J. Allan (Ed.), Topic Detection and Tracking: Event-based Information Organization, Kluwer Academic Publishers, Boston, 2002.

[3] S. Brin, J. Davis, H.G. Molina, Copy detection mechanisms for digital documents, in: ACM SIGMOD'95, 1995, pp. 298–409.

[4] Y. Zhang, J. Callan, T. Minka, Novelty and redundancy detection in adaptive filtering, in: ACM SIGIR'02, 2002.

[5] T. Brants, F. Chen, A. Farahat, A system for new event detection, in: SIGIR'03, Canada, July 2003.

[6] J. Allan, C. Wade, A. Bolivar, Retrieval and novelty detection at the sentence level, in: ACM SIGIR'03, Canada, July 2003, pp. 314–321.

[7] D. Metzler, Y. Bernstein, W. Croft, et al., Similarity measures for tracking information flow, in: CIKM'05, Germany, October 2005.

[8] E. Gabrilovich, S. Dumais, E. Horvitz, Newsjunkie: providing personalized newsfeeds via analysis of information novelty, in: WWW'04, USA, 2004, pp. 482–490.

[9] X. Li, W.B. Croft, Novelty detection on sentence level pattern, in: CIKM'05, Germany, October 2005.

[10] Y. Yang, J. Zhang, J. Carbonell, C. Jin, Topic-conditioned novelty detection, in: SIGKDD'02, Canada, 2002.

[11] L.S. Larkey, F. Feng, M. Connell, V. Lavrenko, Language-specific models in multilingual topic tracking, in: SIGIR'04, UK, July 2003.

[12] S.C. Cheung, A. Zakhor, Efficient video similarity measurement with video signature, IEEE Trans. Circuits Syst. Video Technol. 13 (1) (2003) 59–74.

[13] A.K. Jain, A. Vailaya, W. Xiong, Query by video clip, ACM Multimedia Syst. J. 7 (1999) 369–384.

[14] Y. Peng, C.-W. Ngo, Clip-based similarity measure for query-dependent clip retrieval and video summarization, IEEE Trans. Circuits Syst. Video Technol. 16 (5) (2006) 612–627.

[15] D.-Q. Zhang, S.-F. Chang, Detecting image near-duplicate by stochastic attributed relational graph matching with learning, in: ACM MM'04, USA, October 2004.

[16] P. Duygulu, J.-Y. Pan, D.A. Forsyth, Towards auto-documentary: tracking the evolution of news stories, in: ACM MM'04, USA, October 2004, pp. 820–827.

[17] Y. Ke, R. Sukthankar, L. Huston, Efficient near-duplicate detection and sub-image retrieval, in: ACM MM'04, USA, October 2004, pp. 869–876.

[18] C.-W. Ngo, W.-L. Zhao, Y.-G. Jiang, Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation, in: ACM International Conference on Multimedia (ACM MM'06), Santa Barbara, CA, USA, October 23–27, 2006.

[19] W.H. Hsu, S.-F. Chang, Topic tracking across broadcast news videos with visual duplicates and semantic concepts, in: The International Conference on Image Processing (ICIP'06), Atlanta, GA, October 2006.

[20] S.-F. Chang, et al., Columbia University TRECVID-2005 video search and high-level feature extraction, TRECVID 2005, Washington DC, 2005.

[21] X. Zhu, J. Fan, A.K. Elmagarmid, X. Wu, Hierarchical video content description and summarization using unified semantic and visual similarity, Multimedia Syst. 9 (1) (2003) 31–53.

[22] Y. Zhai, M. Shah, Tracking news stories across different sources, in: 13th ACM Annual Conference on Multimedia (ACM MM'05), Singapore, November, 2005.

[23] X. Wu, C.-W. Ngo, Q. Li, Threading and autodocumenting news videos, IEEE Signal Process. Mag. 23 (2) (2006) 59–68.

[24] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: SIGIR'01, USA, September 2001, pp. 334–342.

[25] Y. Zhang, W. Xu, J. Callan, Exact maximum likelihood estimation for word mixtures, Text Learning Workshop at the International Conference on Machine Learning (ICML), 2002.

[26] TRECVID 2005 [online], Available from: <http://www-nlpir.nist.gov/projects/trecvid/>.

[27] CMU Informadia Project, Available from: <http://www.informedia.cs.cmu.edu/>.

[28] J.L. Gauvain, L. Lamel, G. Adda, The LIMSI broadcast news transcription system, Speech Commun. 37 (1–2) (2002) 89–108.

[29] Google translation [online], Available from: <http://translate.google.com/>.

[30] D. Lowe, Distinctive image features from scale-invariant key points, Int. J. Comput. Vis. 60 (2004) 91–110.

[31] X. Wu, W.-L. Zhao, C.-W, Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context, in: ACM International Conference on Image and Video Retrieval (ACM CIVR'07), The Netherlands, July 2007.