VideoQA: Question Answering on News Video

Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, Tat-Seng Chua
School of Computing, National University of Singapore
Singapore 117543
{yangh, lekhacha, zhaoyl, neoshiyo, chuats}@comp.nus.edu.sg

ABSTRACT

When querying a news video archive, the users are interested in retrieving precise answers in the form of a summary that best answers the query. However, current video retrieval systems, including the search engines on the web, are designed to retrieve documents instead of precise answers. This research explores the use of question answering (QA) techniques to support personalized news video retrieval. Users interact with our system, VideoQA, using short natural language questions with implicit constraints on contents, context, duration, and genre of expected videos. VideoQA returns short precise news video summaries as answers. The main contributions of this research are: (a) the extension of QA technology to support QA in news video; and (b) the use of multi-modal features, including visual, audio, textual, and external resources, to help correct speech recognition errors and to perform precise question answering. The system has been tested on 7 days of news video and has been found to be effective.

Categories and Subject Descriptor

H.3.1 [**Content Analysis and Indexing**]: linguistic processing, thesaurus.

H.3.3 [**Information Search and Retrieval**]: query formulation, retrieval model, search process.

General Terms

Design, experimentation

Keywords

Video question answering, video retrieval, video summarization, transcript error correction

1. INTRODUCTION

Video is the most effective medium for capturing the events in the real world around us. It is also the most dramatic medium as it combines both photo-realistic images and sounds. One of the key technologies required for the efficient management of video data is to support personalized video services, especially on the more structured video sources such as news and sports. Unfortunately, the benefits of such materials are often impeded by the fundamental difficulties with information retrieval: that finding specific information on a video source can be a process that is not only time-consuming and tedious, but also frequently unreliable. Current search engines [3] on the web are designed to return relevant documents efficiently, but not precise answers. The situation is worse for video as it is digitized and stored as a continuous stream, which may be up to an hour in duration. In respond to users' questions, most current systems are able to return either a pre-defined summary, or the whole video sequence. This is unsatisfactory as it often takes a long time for the users to locate where the information is in the video stream. Ideally, the system should return a reasonably short video segment, preferably only as long as is necessary, to provide the requested information.

There are many simple factoid questions like: "*What is the score of the match last night?*" or "*What are the symptoms of atypical pneumonia?*" posed over news video collection. To unearth the concise and informative answers from a given video requires good understanding of the video semantic content. The semantic contents of video come from multiple sources including the inherent content features, accompanying speech or closed captioned text, metadata and external resources such as the web-based news articles. These sources of data may contain errors and may be inconsistent. It is thus necessary to fuse these multiple sources of often imprecise information to discover the story units in video, identify the appropriate portions of stories that answer the questions, and to generate video summary.

The realization of a video-based QA system requires the solution to at least three fundamental problems in video and text processing. The first is to segment the video sequence into story units with correct genre classification. Several works have been done on this, including [4, 13]. These approaches perform multi-modal analysis using a combination of visual, audio and textual features based on HMM or entropy techniques and reported high accuracies in story segmentation and genre classification. The analysis should also generate summary to provide concise answers.

The second problem is that the users' questions are normally short, imprecise and assume previous context. For example, to correctly answer the question: "*What is the score of the match last night?*" requires the analysis of *key terms* ("score, match"), *context* ("for example, football match between teams A & B"), *video genre* (sports), and implicit constraint like *duration* (<= 30 seconds). There are several ways to induce the precise meaning and context of a question. These include the use of query logs, user profiles, or the recent news articles available on the news

Multimedia '03, November 2-8, 2003, Berkeley, CA, USA.

web sites. The external web resources provide a general approach to unearth question semantics.

The third problem is to overcome the recognition errors in the speech accompanying the video. Text from speech (or transcript) is a major source of semantic information for news video. The conventional video retrieval based on transcript suffers a lot from the numerous speech recognition errors. These errors will affect the ability of the system to analyze and retrieve the relevant transcript at the sentence level. Most errors are of type substitution that cause many names of person, location and organization to be wrongly recognized. As names are essential to induce the semantics of news, there is a need to identify and correct such errors using the names that we already know from related news articles.

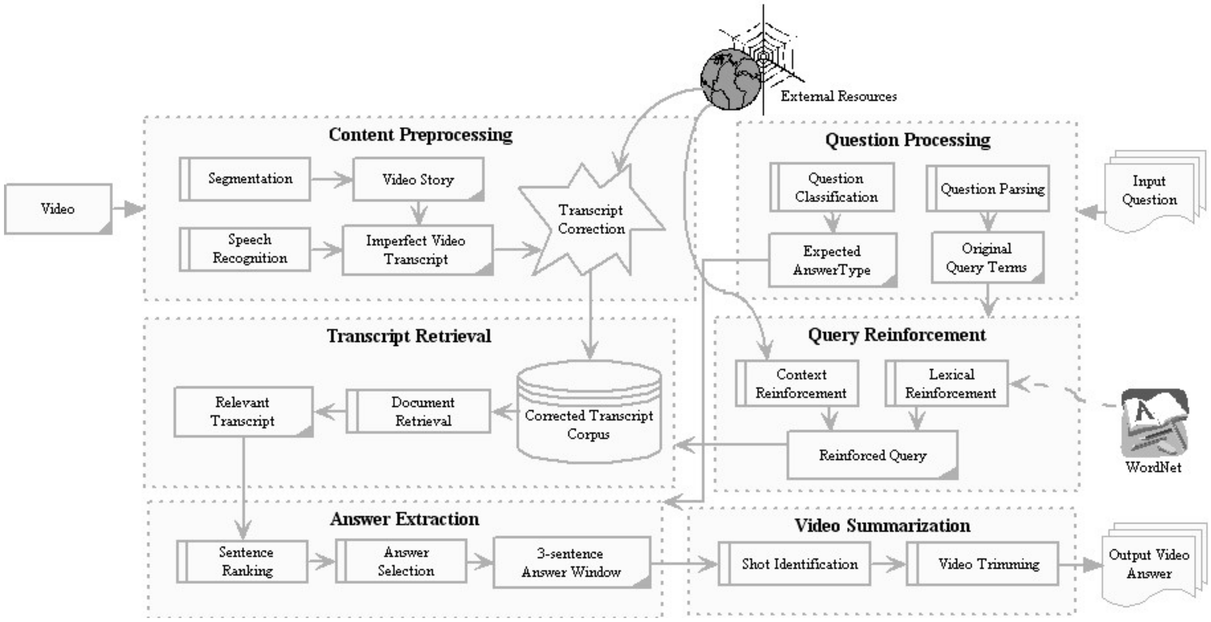

Figure 1: System architecture of VideoQA

In this paper, we discuss the design of a news video question answering system called VideoQA. Users interact with VideoQA using short natural language questions with implicit constraints on contents, context, duration, and genre of expected videos. The system returns the relevant news video fragments as the answers, supplemented by text version of latest news, summarized to the duration constraint as specified by the users. The paper discusses our research in tackling the above three problems, and the summarization of video into a single or multiple-sentence units. The main contributions of this work are: (a) the extension of question answering technology to support QA in news video; (b) the use of multi-modality analysis and external knowledge to extract story boundaries, correct speech recognition errors and to perform precise question answering.

The rest of the paper is organized as follows. Section 2 outlines our system architecture. Section 3 discusses our approach to correct news transcript errors by utilizing external web resources. Section 4 details the application of QA technology to retrieve precise answers in video database. Section 5 presents the experimental results. Sections 6 & 7 respectively outline related work and conclude the paper.

2. SYSTEM ARCHITECTURE OF VideoQA

VideoQA aims to provide precise video answers to simple factoid questions posed over the news video collection. It is naturally used in a personalized video setting in which a user may request for details of different aspects of news and their summaries. It will be an essential component of future information systems.

During the preprocessing stage, VideoQA performs video story segmentation and classification, as well as video transcript generation and correction. During question answering, VideoQA employs modules for: question processing, query reinforcement, transcript retrieval, answer extraction and video summarization. Figure 1 gives the system architecture of VideoQA.

Given the news video collection, the pre-processing stage "prepares" the video for later answer retrieval. We analyze the raw video using a two-level story segmentation scheme as proposed in [4]. The basic unit of analysis is the shots, and we employ multi-modal analysis involving visual, audio and textual features. Briefly, we model each shot using high-level object-based features (face, video text, and shot type), temporal features (background scene change, speaker change, motion, audio type, and shot duration), and low-level visual feature (color histogram). At the shot level, we employ the Decision Tree to classify the shots into one of 13 genre types of: *Intro/ Highlight, Anchor-person, 2-anchor-person, Meeting/ Gathering, Speech/Interview, Live-reporting, Still-image, Sports*, *Text-scene, Special, Finance, Weather, and Commercials*. We then perform HMM analysis to detect story boundaries using the shot genre information, as well as time-dependent features based on speaker change, scene change and key phrases. The resulting video story may contain shots of different genre types. For example, a general news story typically contains shots of type *Anchor-person, Live-reporting* and *Speech/Interview*; while a sports story includes shots of type *Sports* and *Text-scene*. The overall story segmentation scheme is shown in Figure 2. Preliminary results based on small set of test video [4] showed that we could achieve an F_1 measure of about 89% in story segmentation, and that the use of only text-based segmentation technique [12] is not effective for this task. We are currently testing our approach on the full 120-hour of news video from TREC [24].

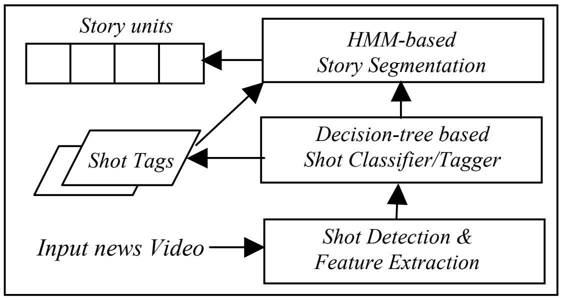

Figure 2: Overview of video story segmentation system

The pre-processing stage also generates the text transcripts of video by performing the speech recognition on the audio track. However, the transcripts contain numerous speech recognition errors, which cause many words, especially *names,* to be wrongly recognized as one or more similar sounding words. To correct such errors, we retrieve the news articles from the news web sites, and extract a list of possible names and their correlations. We then employ a phonetic-based matching technique to match the likely names in the name list to those terms in the transcript to correct the speech recognition errors.

In the question answering stage, we first perform question analysis to extract the key terms in the question; the type of questions and its likely answer targets; the type of video genre, and the implicit duration constraint. We then go to the news web sites to retrieve the latest related news and derive the words related to the queries. These terms, especially the named entities, provide the context to the queries. We expand the original query using these extracted terms to form a new query, ie. $q^{(0)} \rightarrow q^{(1)}$. Next, we match the new query $q^{(1)}$ against the (corrected) transcripts at the story level, and perform QA analysis to retrieve the relevant sentences. We ensure that the retrieved sentence(s) cover video shots of the expected genre type. For example, for speech query, we expect the video to be of type *Speech* with a detected face. The analysis helps to remove those retrieved passages associated with video genre of the unlikely types. For example, "*Saddam Hussein*" cannot be associated with sports video. Finally, we generate a single or multi-sentence video summary. The following sections described the details of our approaches

3. PROCESSING OF NEWS TRANSCRIPT

In this work, we use the Sphinx-III speech recognition engine [21] to transcribe the synchronized news transcript segmented at sentence level (separated by acoustic "paragraph" break through silence). Sphinx III is a large-vocabulary, speaker-independent, continuous speech recognizer developed at CMU. Because of the complexity of human vocal track, and the differences across different speakers, dialects, transmission distortions, and speaking environments [16], many errors incurred during speech recognition. Several studies [11, 14, 25] have shown that typical spoken document retrieval (SDR) systems could tolerate up to 30% speech recognition errors with no significant reduction in the accuracy of retrieval. These results, however, are not applicable to QA, which concerns with the retrieval of sentences rather than documents. At document level in SDR, many important terms are often repeated many times and are thus more likely to be recognized correctly at least once. Thus the overall retrieval at the document level tends to be good. For QA, relevant sentences will not be retrieved if such errors are not corrected at the sentence level.

One main type of speech recognizer error is substitution, where one or more wrong similar sounding words are "substituted" in place of the correct word. This type of error occurs most frequently on named entities (NEs) such as the names of persons, organizations, locations and events etc., as most NEs tend to be out-of-vocabulary words. Examples of such errors, as listed in Figure 3, include: *pneumonia* → *new area*; and *Jose Maria Aznar* → *Jose Mari ask not*. For ease in discussion, we use the term *Answer Target (ATs*) to collectively denote the name entities, which include also dates and numbers etc.

ATs are essential to performing QA at sentence level as most queries on news tend to involve news worthy subjects. The errors in ATs must be corrected to improve QA accuracy. One obvious approach to correct the substitution errors is to convert the words to phonetic sounds and match the phonetic sequence of words at syllable level. A similar approach has been used to expand [5, 22, 23] or correct [28] spoken language queries for effective information retrieval. As there may be a large number of similar sounding words, a straight-forward application of phonetic matching may result in low accuracy [17]. Thus we need to constrain the list of terms to be matched in order to ensure accuracy. Fortunately, in news domain, we are able to extract a list of possible ATs from recent news articles available on the news web sites. This is similar to the approach taken in [10, 28]. In addition, we use OCR output of video text in video news stories to help correct the speech recognition errors. This section discusses our approach to correct most speech recognition errors in ATs to support QA.

3.1 Associating Answer Targets to Video Transcripts

The basic premise in transcript error correction is that recent news articles share many common terms with the video transcripts, especially for ATs. This is reasonable as news worthy ATs of up to a week away tend to re-appear in current news. Thus the first task is to retrieve a list of recent news articles from the news web sites, $\underline{D}_{all}$. The news web sites we used include Alta Vista and CNN news sites.

The problem of error correction in news transcript can be expressed as:

$$p(s_l \mid a_{jk})\ \ p(\underline{A}_j \mid D_j)\ \ p(\underline{D}_i \mid T_i) \qquad (1)$$

where T_i is the news transcript i with $T_i = (t_{i1}, t_{i2}, .., t_{ip})$ of p terms; $\underline{D}_i \in \underline{D}_{all}$ is the list of news articles relevant to T_i; $D_j \in \underline{D}_i$, and $\underline{A}_j$ is the list of ATs contained in D_j; $a_{jk} \in \underline{A}_i$ is an AT; and s_l is a sequence of one or more consecutive terms in T_i. Equation (1) expresses the process of correcting errors in T_i through the use of $\underline{A}_j$ derived from the relevant news articles $\underline{D}_i$.

First, $p(\underline{D}_i|T_i)$ is estimated by computing the similarity $Sim(T_i,D_j)$ between the news article D_j and T_i using the cosine similarity measure [19]. We select the top m articles with $Sim(T_i,D_j)>\sigma_1$.

Second, $p(\underline{A}_j|D_j)$ is obtained by simply selecting the list of ATs that appear in D_j, which in turn are related to T_i. The list of ATs, which we denote as $a_{jk}\in\underline{A}_j$, are obtained using the shallow parsing tools from UIUC [18] to extract the noun phrases and the tool developed by [7] to extract the name entities.

Third, we use the list of ATs in $\underline{A}_j$ as the basis for phonetic sound matching of terms in transcript T_i at the syllable level. The basic problem of estimating $p(s_l \mid a_{jk})$ is then to select an $a_{jk}\in A_i$ to replace a sequence of terms $s_l\in T_i$ that maximizes the probability:

$$\arg\max_{s_l\in T_i \wedge a_{jk}\in A_i} \; p(s_l | a_{jk}) \qquad (2)$$

where s_l contains one or more consecutive terms in T_i.

3.2 Correcting Transcript Errors

We now discuss how we estimate measure (2) using phonetic matching technique. As substitution error in speech recognition is typically a homonym problem involves the "substitution" of an AT by one or more simpler similar sounding words in T_i, our system concentrates on matching the set of ATs in $\underline{A}_j$ to single or multiple terms in transcript T_i.

One example of substitution error is the AT "*pneumonia*" that was wrongly recognized as two simpler words or a phrase, "*new area*". Their respective phonetic strings as defined in the phonetic dictionary of Sphinx system are: *<N AH M OW N Y AH>* and *<N Y UW>*, *<EH R IY AH>*. It is clear that at the phonetic level, both set of strings are highly similar. By observations and through experimentations, we found that the similarity between the phonetic representations of two strings can be established on the basis of: (a) the similarities of their first and last syllables; and (b) the number of correct syllable matches in the occurrence sequence. In addition, the two phonetic strings should have approximately the same length.

Hence, given two phonetic strings x and y of approximately the same length, we derive two measures to compute their similarity as follows (see also Figure 3):

a) String Boundary Similarity, which measures the similarity in the starting (start())and ending (end()) phonetic sounds.

$$S_b(x, y) = \begin{cases} 1, & \text{if start(x)=start(y) and end(x)=end(y);} \\ 0.5, & \text{if start(x)=start(y) xor end(x)=end(y);} \\ 0, & \text{otherwise.} \end{cases} \qquad (3)$$

b) Longest Common Sub-sequence (LCS) Similarity. This computes the number of phonetic matches between two strings in their occurrence order using the LCS algorithm [9], and normalizes the measure by the phonetic length of string x.

$$S_l(x,y) = LCS(x,y)/|x| \qquad (4)$$

The overall similarity between $a_{jk}\in\underline{A}_j$ and the phrase $s_l\in T_i$ is:

$$S_p(a_{jk}, s_l) = \alpha_b\, S_b(a_{jk}, s_l) + \alpha_l\, S_l(a_{jk}, s_l) \qquad (5)$$

where $\alpha_b+\alpha_l=1$ are weights in which we set $\alpha_b=\alpha_l=0.5$.

In many cases, the AT may contain multiple words, such as the AT "*tony blair*". In this case, we compute the $S_p^r(x_k^r,y_k^r)$ for each word *r* in AT with a term or phrase of equivalent length in the transcript, and compute the overall similarity as:

$$p(s_l | a_{jk}) = S_p(a_{jk}, s_l) = \frac{(\gamma)^{N_a-1}}{N_a}\sum_{r}^{N_a} S_p^r(a_{jk}^r, s_l^r) \qquad (6)$$

where a^r_{jk} and s^r_l are sub-elements of a_{jk} and s_l respectively, and N_a is the number of words in the AT to be matched. $\gamma>1$ is a constant to give higher weight to multiple word matches.

We use a_{jk} to replace s_l in transcript T_i iff:

$$\arg\max_{k} \; p(s_l | a_{jk}) \;\wedge\; [p(s_l | a_{jk}) > \sigma_2] \;\wedge\; a_{jk}\in A_i \qquad (7)$$

Example 1:
Original NP: "pneumonia" <N AH M OW N Y AH>
Recognized string: "new area" <N Y UW> <EH R IY AH>
$S_b(x, y)=1$; $S_l(x, y)=2/7$;
Overall similarity is: *S_p(pneumonia)=0.5 * (1 + 2/7)*

Example 2:
Original NP: "tony blair" <T OW N IY> <B L EH R>
Recognized string: "teddy bear" <T EH D IY> <B EH R>
For tony: $S_b(x, y)=1$; $S_l(x, y)=2/4$;
For blair: $S_b(x, y)=1$; $S_l(x, y)=3/4$;
Hence: *S_p("tony blair") = [(S_p(tony) + S_p(blair)]* γ /2*

Example 3:
Original NP: "Jose Maria Aznar"
<HH OW Z EY> <M ER IY AH> <?>
Recognized string: "Jose Mari ask not"
<HH OW Z EY> <M AA R IY> <AE S K> <N AA T>
For jose: $S_b(x, y)=1$; $S_l(x, y)=1$;
For maria: $S_b(x, y)=1/2$; $S_l(x, y)=2/4$;
For aznar: $S_b(x, y)=0$; $S_l(x, y)=0$;
Hence: S_p("jose maria aznar") = [(S_p(jose) + S_p(maria)] $(\gamma)^2/3$*

Figure 3: Examples of NPs & the wrongly recognized strings

3.3 Correcting other Transcript Errors using Video Text

In the above examples, we look up Sphinx phonetic dictionary to obtain the phonetic representation of the terms in $\underline{A}_j$ and T_i. We, however, noticed that many non-Latin names are wrongly recognized. Because of the limitation of the version of Sphinx-III system that we are using, we cannot find the phonetic representations for many of these names, which tend to be out-of-vocabulary terms. Thus we cannot correct the errors in these names using the phonetic matching technique as described above. Fortunately, many such names also appear in video text during the news story. For example, the name "*Wen Jiabau*" is not in the phonetic dictionary of Sphinx-III, but it appears as video text: "*WEN JIABAU IS NOW CHINA TOP ECONOMIC OFFICIAL*" (see Figure 4a). Similarly, "*Blix*" is not in Sphinx-III dictionary but appears in the video text of the news story as "*BLIX, ELBARADEI INVITED BACK TO BAGHDAD*" (see Figure 4b).

(a) (b)

Figure 4: Video text appearing in video news stories

Given the video-text output (with about 25% character recognition errors), we adopt a greedy approach to approximately match the presence of ATs$\in \underline{A}_j$ in the video text, and in transcript T_i as follows:

a) We extract OCR output of video-text in video story i as $\underline{T}_{vtext}$. We look for possible ATs$\in \underline{A}_j$ in $\underline{T}_{vtext}$ by performing character (or letter) level matching using the LCS algorithm. This is to cater to possible OCR errors in video text recognition at the character level. We denote the ATs found as $\underline{A}_{vtext}$.

b) If $\underline{A}_{vtext}$ is found with sufficiently high confidence, we append it at the appropriate position in T_i. Otherwise, we append $\underline{A}_{vtext}$ at the end of transcript T_i.

The above procedure is "greedy" as it tries to append $\underline{A}_{vtext}$ into the appropriate position in T_i as much as possible. This is to maximize the recall in retrieving the transcripts and sentences, as we have other measures during QA analysis to remove wrong transcripts in order to improve precision.

4. VIDEO QUESTION ANSWERING

Given the segmented news video stories, each with the "corrected" news transcript, the next task is to perform QA during retrieval to select precise answers at the sentence level. We assume that a user wants to view a video summary in less than 30 seconds. Hence, by default, we return about 3 sentences for each question. For each question, we perform a series of analysis including question processing, question reinforcement, transcript retrieval and sentence extraction, and video summarization. Our approach is adopted from the one described in [29, 30]. The following sub-sections describe the details of each step.

4.1 Question Processing

Users may issue short questions like "*What is the score of the football match last night?*" or "*What are the symptoms of atypical pneumonia?*" For each question, we need to infer the intents of the users, both in terms of precise information needs, and the type of expected answer targets and video genre type. In general, the question $Q^{(0)}$ can be modeled as:

$$Q^{(0)} = Content + Constraint \quad (7)$$

$$Content := query\ words\ \underline{q}^{(0)};\ noun\ phrases\ \underline{n};\ named\ entities\ \underline{h}$$

$$Constraint := answer\text{-}target;\ video\text{-}genre\text{-}type;\ time\text{-}duration$$

The *Content* models the user's precise information needs, while *Constraint* specifies the expected answer types.

The question analysis aims to classify the question into one of 8 main question classes (or answer targets) and the expected video genre. The 8 question classes are shown in Table 1 as: *Human, Location, Organization, Time, Number, Object, Description and General*. The last answer target *General* is used to group questions that cannot be categorized into other classes. The possible video genre types are in one of the 14 types as explained in Section 2, with the addition of a *General News* video type. Both the answer target and video genre found are used to locate the precise sentences in the news transcript. Our system employs a rule-based question classifier to determine the answer target and video genre type and could achieve an accuracy of over 95% (see [29]). The examples are given in Table 1.

Table 1: Question classification and possible video genres

* Show only the likely video genres for the specific question examples

Answer Target	Likely Video Genre	Example
Human	Anchor, meeting, speech, General-news	Who is the Secretary of State of the United States?
Location	Live report, Anchor, General-news	Where is Saddam Hussein hiding?
Organization	Live report, anchor	Report on the hospital that is at the center for SARS treatment in Singapore?
Time	Anchor, General-news	When did the Iraq war start?
Number	Finance	The expected GDP of Singapore this year?
	Sports, Text-scene	The number of points scored by Yao Ming?
	Weather, Text-scene	What is the highest temperature tomorrow?
Object	Anchor, Still-image, Text-scene	The kinds of bombs used in the current Iraq war?
Description	Anchor, Text-scene, General-news	What does SARS stand for?

The question analysis also extracts important content information that is crucial for later processing. Detailed analysis is performed here in order to get as much useful information as possible. The three kinds of word groups that we extract from the original question are:

a. ***Original Query Words:*** These include nouns, adjectives, numbers, and some non-trivial verbs that appear in the question string. For example: "*Which is the first company to find SARS patient in Singapore?*", the content word vector will be $\underline{\mathbf{q}}^{(0)}$**: (company, first, SARS, patient, Singapore).**

b. ***Base Noun Phrases:*** we use noun phrase recognizer to identity all base noun phrases appearing in the question. For the above example, the base noun phrase vector $\underline{\mathbf{n}}$**: ("SARS patient")**

c. ***Named Entities:*** They refer to noun phrases that represent Person, Organization, Location, Time, Number, and Object etc. For the above example, the named entity vector $\underline{\mathbf{h}}$**: ("SARS", "Singapore")**

Table 2 shows the analysis of the above two query examples.

Table 2: Question analysis

Question	*What is the score of the football match last night?*	*What are the symptoms of atypical pneumonia?*
$\underline{q}^{(0)}$	score, football, match, last, night	symptoms atypical pneumonia
$\underline{n}$	*football match, last night*	symptom, *atypical pneumonia*
$\underline{h}$	football	*atypical pneumonia*
Answer Target	Number	Description
Video Genre	Sports, Text-scene	General News, Anchor, Text-scene

4.2 Question Reinforcement

One of the major problems in processing short questions in QA is that the questions are imprecise and lack important context. It is necessary to utilize additional knowledge to extract the context of the questions. Our aim at this stage is to explore an approach to extract context that is applicable to general users. Since we are dealing with news, we again make use of recent news articles obtained from the news web sites to extract the context.

Given the short factoid question $\underline{q}^{(0)}$, we go to the news web sites to retrieve the top N_w latest news articles related to the question. For each term $q_i^{(0)} \in \underline{q}^{(0)}$, we extract the list of non-trivial words, $\underline{w}_i$, that are within the local context of n_s-sentence window of $q_i^{(0)}$. We compute the weights for all terms $w_{ik} \in \underline{w}_i$ based on the probabilistic support of their occurrences with $q_i^{(0)}$ as:

$$\frac{d_s(w_{ik} \wedge q_i^{(0)})}{d_s(w_{ik} \vee q_i^{(0)})} \qquad (8)$$

where $d_s(w_{ik} \wedge q_i^{(0)})$ gives the number of n_s-sentence windows that contain both w_{ik} and $q_i^{(0)}$; and $d_s(w_{ik} \vee q_i^{(0)})$ gives the number that contains either w_{ik} or $q_i^{(0)}$. We merge all $\underline{w}_i$ to form the context word list $\underline{C}_q$ for $\underline{q}^{(0)}$.

Various studies [2, 8] have shown that Web is useful at finding world knowledge by providing words that occur frequently with the original query terms in the local context. However, it lacks information on lexical relationships among these terms, such as synonyms. To derive the lexical knowledge, we use WordNet [15], which is an electronic dictionary, to get the gloss (definition of terms) words $\underline{G}_q$ and synset (synonym sets) words $\underline{S}_q$ for $\underline{q}^{(0)}$. In order to ensure that we do not assign words in $\underline{G}_q$ and $\underline{S}_q$ out of context, we restrict only to those terms in WordNet that appear also in $\underline{C}_q$ and $\underline{q}^{(0)}$. In other words, we use $\underline{G}_q$ and $\underline{S}_q$ to increase the weights of the terms that appear in both the Web and WordNet. The final weight of each term is normalized and the top N_c terms above the cut-off threshold σ_c are selected to derive the context word vector $\underline{q}^{(1)}$:

$$\underline{q}^{(1)} = \underline{q}^{(0)} + \{\text{top } N_c \text{ terms} \in \underline{C}_q \text{ with weights} >= \sigma_c\} \qquad (9)$$

where N_c is initially set to 20 in our experiments.

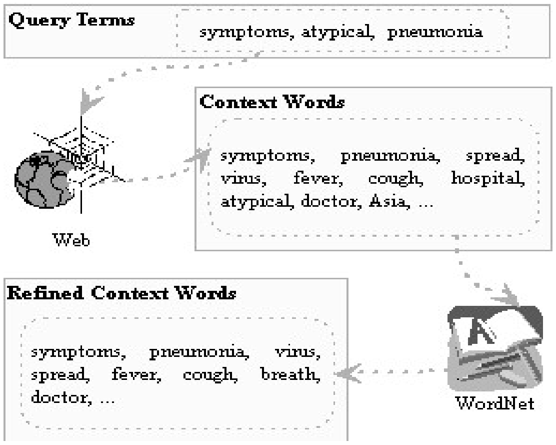
Query Terms
symptoms, atypical, pneumonia
Context Words
symptoms, pneumonia, spread, virus, fever, cough, hospital, atypical, doctor, Asia, ...
Web
Refined Context Words
symptoms, pneumonia, virus, spread, fever, cough, breath, doctor, ...
WordNet

Figure 5: Example of query reinforcement

$\underline{q}^{(1)}$ should contain more context words than $\underline{q}^{(0)}$ and hence is better able to identify the correct sentence(s) that answer the question. The overall query expansion process for the question *"what are the symptoms of atypical pneumonia"* is shown in Figure 5. The resulting $\underline{q}^{(1)}$ after the question expansion is: *"symptoms, pneumonia, virus, spread, fever, cough, breath, doctor"*.

4.3 Transcript Retrieval &Answer Extraction

Given the expanded question $\underline{q}^{(1)}$, we perform similarity retrieval to obtain a ranked list of transcripts. Starting from the top ranking transcript T_i, our aim is to identify sentences in T_i that best answer the question. If such sentences cannot be found in T_i, we move down to next transcript in the ranked list. The best sentence that answers the question should contain: (a) important term groupings from the question such as $\underline{q}^{(0)}$, $\underline{n}$, $\underline{h}$ *and* $\underline{q}^{(1)}$ (see Table 2); (b) phrase of type answer target; and, (c) cover video shot of the expected genre type. Thus for each sentence $Sent_{ij}$ in transcript T_i, we match $Sent_{ij}$ with the following entities derived from the question analysis:

- noun phrases $\underline{n}$: w_{in} = percentage of overlap between $\underline{n}$ and $Sent_{ij}$
- named entities $\underline{h}$: w_{ih} = 1 if there is a match between $\underline{h}$ and $Sent_{ij}$; and 0 otherwise.
- original query words $\underline{q}^{(0)}$: w_{io} = percentage of term overlap between $\underline{q}^{(0)}$ and $Sent_{ij}$.
- expanded query words: w_{ie} = percentage of term overlap between $\underline{q}^{(1-0)}$ and $Sent_{ij}$, where $q^{(1-0)} = \underline{q}^{(1)} - \underline{q}^{(0)}$.
- Answer target: w_{ia} = 1 if $Sent_{ij}$ contains a phrase of type answer target; and 0 otherwise.
- Video genre: w_{iv} = 1 if $Sent_{ij}$ overlaps with shots of the right genre type; and 0 otherwise.

The final score for the sentence is:

$$S_{ij} = \sum_k \alpha_k W_{ik} \qquad (10)$$

where $\sum \alpha_k = 1$ and $w_{ik} \in \{w_{in}, w_{ih}, w_{io}, w_{ie}, w_{ia}, w_{iv}\}$. If the score s_{ij} of the top ranking sentence $Sent_{ij}$ in transcript T_i is above the threshold σ_s, then transcript T_i is deemed to contain the answer. We select the top *K* sentences in T_i as the answer to the query.

In our experiment, we select the top 3 sentences, which correspond to a video summary of less than ½ minute. Users may request for longer summary by requesting for more sentences. For the query "*What are the symptoms of atypical pneumonia?",* the top 3-sentence window selected by the QA engine is:

S_1: *"He and his two companions are now in isolation and the one hundred and fifty five passengers on the flight were briefly quarantined."*

S_2: *"Symptoms include high fever, coughing, shortness of breath and difficulty breathing."*

S_3: *"But health officials say there's no reason to panic."*

S_2 is the top ranking sentence selected.

4.4 Video Summarization

Given the set of news transcript sentences extracted, we perform dynamic news video summarization to generate the video answer. The list of sentences corresponds to appropriate audio fragments in the video story. This gives the constraint on the duration of the video to be shown. The task here is to extract appropriate visual segments, to be shown along with the (audio) sentences, that are both informative and interesting to the users. These two criteria suggest that even though we should show visual segments correspond to the selected transcript sentences; we should also pack as much variety of video genre types within the news story as possible. Also, we should avoid showing too much *Anchor-person* shots [6], even though many selected news transcript sentences are likely to come from such shots.

The algorithm to generate the visual summary, subject to the duration constraint imposed by the selected (audio) transcript sentences, is as follows:

a) We remove those shots that are shorter than 3 seconds in duration.

b) We pick those that overlap in duration with the selected transcript sentences and place them in the *candidate* list.

c) We group video shots of the same genre type together, and perform clustering of shots within each genre type by using the 64-bin color histogram of the key frame for each shot. For each cluster, we select ***two*** representative shots for inclusion in the *candidate* list. We select one representative shot near the centroid, and the other near the boundary. This is to eliminate duplicate shots, and ensure variety in the selection of candidate shots in each cluster.

d) We compute the weight of each shot in the *candidate* list based on two criteria. First, whether it overlaps with the selected transcript sentences (w_t=1) or not (w_t=0). Second, whether it is of more interesting genre type (w_g). We set the "*interesting index*" of each genre type, w_g, to a value ranging from 1 (high interest) to a smaller value. For example, for general news, the shot genres with high "*interesting index*" are: *Live-reporting, Meeting/Gathering*; while for sports, the high "*interesting index*" genres are: *Sports, Live-reporting, Text-scene*. The *Anchor-person, 2-anchor-person* shots have low "*interesting index*". In the current system, we pre-defined the "*interesting index*" of each genre. Further studies in this area need to be carried out.

e) We assign the weight to each shot in the *candidate* list using the following formula:

$$Wt(shot_i) = \beta_t w_{ti} + \beta_g w_{gi} \tag{11}$$

We set β_t=β_g=0.5 in this test in order to strike a balance between choosing visual segments that fall within the transcript duration and those that are "interesting" to the users.

f) Finally, we select the top N_s candidate shots, where N_s is set to the largest integer less than the duration divided by t_s. This is a heuristic to ensure that we can pack a sufficient number of interesting shots, each with a duration of about t_s (set to 6) seconds. We trim the shots to fit into the duration of the audio constraint.

Figure 6 illustrates the process of summarization. Suppose that there are 6 sentences (represented by S1 to S6) and 5 shots (AN is the *Anchor-person* shot, MT1, MT2, and MT3 are *Meeting/Gathering* shots, and SP is the *Speech/Interview* shot) in the story. Assume that our QA engine selects sentences S2 and S4, and the visual summarizer selects MT1 and MT3 from the cluster of 3 *Meeting* shots, and SP of *Speech* genre. Eventually, we trim visual segments from MT1, SP, and MT3 to provide the summary.

Figure 7 shows the video summary extracted for the query: "*What are the symptoms of atypical pneumonia?"*.

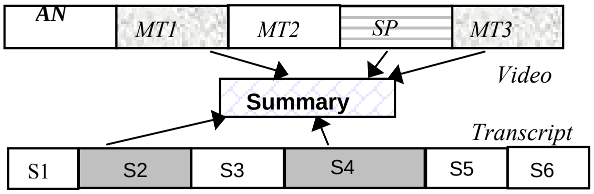

Figure 6: A scenario of general news video summarization

* Only keyframes of shots are shown here.

Figure 7: Video summary of the "*pneumonia*" example

5. EVALUATION

To test the effectiveness of the system, we selected 7 days of CNN news video from 13-19 March 2003 for our test. We used two half-hour news segments per day, giving rise to a total of 350 minutes of news video. After performing news video story segmentation, we extracted a total of 175 video stories. For testing purposes, we also retrieved about 600 news articles per day from the Alta Vista news web site [1] during those 7 days. The total number of articles used is about 4,000. They are used as external resource during QA.

We designed 40 TREC-style [26] questions related to the 7-day news. The questions belong to 8 different question classes at various difficulty levels. The list of questions is given in Figure 8. Among the 40 questions, 28 of them are general questions and are asked everyday during the test period. Most of these general questions will give different answers when posed on different days. The rest of 12 questions (those marked with an asterisk in Figure 8) are date-specific and are relevant only to the specific day. This gives rise to a total of 208 questions. Most questions have answers in the video corpus. Only 4 questions do not have an answer (NIL-answer questions).

5.1 Impact of Video Transcript Correction

We first test the ability of our technique to correct the text recognition errors using a combination of phonetic-based and video-text-based matching techniques based on the name-list extracted from on-line news articles. From the Sphinx-III output of news transcripts for the 7-day news, we manually counted the number of ATs (3,155) and those that are wrongly recognized (1,227). Out of those in error, 762 can be found in Sphinx-III phonetic dictionary. Hence we can attempt to correct the errors in these ATs using the phonetic-based matching technique. The rest of those in error that are not found in the Sphinx's phonetic dictionary (465) would have to be corrected using the OCR-based matching technique.

The results of error corrections are summarized in Table 3. Overall, we are able to correct 68% of errors, while introducing 513 new errors (or false positives). The results suggest that our transcripts error correction technique is effective. The ability to correct 68% of errors will greatly enhance the ability of subsequent QA process to retrieve precise answers.

Table 3: Statistics & performance of transcript correction steps

# of ATs found in 175 news trasncripts	3,155
# of ATs with errors in the news transcripts	1,227
Process 1:	
# of ATs with errors found in Phonetic Dictionary	762
# of AT errors corrected by Phonic Matching	529
Accuracy of Process 1	529/762
Process 2:	
# of ATs with errors not found in Phonetic Dictionary	465
# of additional AT errors corrected by OCR matching	308
Accuracy of Process 2	308/465
% of errors corrected	68.2%
# of false positive	513

1) Who is the British Prime Minister?
2) Who is elected to be China's President?
3) Who is the President of the United States?
4) What is the name of the former Premier of China?
5) What is the name of the new Premier of China?
6) Who will pay the heaviest tallies?
7) Who was arrested in Pakistan?
8) Which musician called off his US tour?
9) When will NASA resume shuttle flights?
10) When will Germany, France and Russia meet?
11) When is the funeral of DjinDjic?
12) Which are the three countries involved in the summit today?
13) Where was the summit held?
14) Which city is the capital of Central African Republic?
15) Which are the three major war opponent countries?
16) To whom US withdrew the aid offer?
17) Which country vowed to veto the resolution today?
18) Which country's compromise proposal was rejected by US?
19) Where is Kashmir Hotel?
20) Where did Iraq invite the chief weapons inspectors to?
21) Which city has the largest anti war demonstration?
22) Where did a AL QUEDA suspect arrested?
23) How many people attended the rally in San Francisco?
24) What is the cost of war?
25) How many people were killed in a Kashmir Hotel?
26) How many people participated in the rally in Madrid?
27) How many people were killed by the new pneumonia?
28) What are the symptoms of the atypical pneumonia?
29) What sanction did President Bush lift?
30) What was the name of the space shuttle broken apart in February?
31) Which rally shows the support for President Bush?
32) What is the official name for the mysterious pneumonia?
33) Which company tests their new passenger profiling system?
34) Name one Jewish holiday.
35) What is the British stance?
36) How did Serbs Prime Minister die?
37) How is the anti-war protest in Madrid?
38) How is tomorrow's weather?
39) What is the conflict between US and Turkey?
40) What does the WHO call the new pneumonia?

Figure 8: List of textual questions for news video

5.2 Question Answering Performance

Next, we evaluate the ability of the QA system in retrieving the correct sentence(s) containing the answer. As long as the correct answer is contained in one of the three returned sentences, we consider the answer to be correct. In order to assess the effects of transcript error corrections, we test the performance of our QA system on: (a) the raw transcripts without error correction; and (b) the transcript with error correction.

Tables 4-6 tabulate the results. The results in Table 6 indicate that without transcript error corrections, our QA system could achieve a QA accuracy of about 55%. However, with the corrected transcript, the QA accuracy improves drastically to about 73%. The results demonstrate that our transcript error correction is useful, and it helps in realizing a usable video QA system.

The results in Table 4 reveal that our QA system performs reasonably well on general questions.

Table 4: Accuracy over 196 (28*7) general questions

Transcript	*Correct Answers*	*Accuracy*
without error correction	110	56.1%
with correction	143	73.0%

Table 5: Accuracy of over 12 date-specific questions

Transcript	*Correct Answers*	*Accuracy*
without error correction	6	50%
with error correction	10	93.3%

Table 6: Overall accuracy of 208 questions

Transcript	*Correct Answers*	*Accuracy*
without error correction	116	55.8%
with error correction	153	73.6%

5.3 Discussion of Results

In developing such a system that integrates technologies and research from many fields, many sources of errors may incur and could affect the quality of the results. The main remaining sources of errors include:

- Errors in segmenting video sequence into story units, and in identifying sentence boundaries using the statistics of speech pauses. The cumulative error is about 18% for the news video that we are testing. More works need to be done to improve the accuracy of story segmentation using multi-modality analysis.
- Most non-Latin names are wrongly recognized. For those names that do not have equivalent phonetic representation in the dictionary, and also do not appear or wrongly recognized in video text, there is no means of correcting such errors. In fact, more than 10% of uncorrected ATs fall under this category. To further improve the performance of error correction, we need better video OCR tool and a speech recognizer that can be tuned to better handle non-Latin names. We should also explore the association of multimedia features with concepts, and the functions of speech recognizer such as Sphnix III to return n-best probable words for a speech utterance.
- Certain wrongly recognized words are hard to correct. For example, *Baghdad* is wrongly recognized as *burger*, and *Chile* as *chalet*. Both of these substituted words are meaningful single words and are thus hard to identify. To correct such error, we need to incorporate statistical language model to determine the probability of occurrence of certain words or ATs in the language context and constructs.

- The QA quality is quite good. It, however, uses only multi-modal features in the form of video genre type to constrain the answers to sentences that cover video shots of the correct genre. Further improvement is likely to come from text, and appropriate mid-level features and their association to concepts.

6. RELATED WORK

The main emphasis of this work is on the generation and correction of errors in video transcripts, and in performing QA to extract precise answers. Our work is related to other research on speech and video retrieval. One of the most related works is the well-known *Informedia* project, which covers most aspects of feature extraction, segmentation, and retrieval of news video [6, 27]. Similar to our approach, they also utilized news transcripts and external news articles to help correct feature extraction errors. In particular, they used the name lists extracted from the news transcripts and external news articles to improve the accuracy in video OCR [27], and in associating names to faces [20]. Our work differs from this in that we use the name list in several aspects: to constraint the phonetic search and OCR matching in correcting the speech recognition errors in the transcripts; and to induce context in the questions.

More recent work in *Informedia* project [6] introduced the idea of video collages as an effective interface for browsing and interpreting video collections. They extracted video stories, key phrases of news transcripts comprising mainly of names of person, location and organization; and other structured information. The system supports queries by users to retrieve information through map, text and structured information like date range. Differing from this work that supports database style search for information, we aim to generate correct transcript at the sentence level with the aim to perform flexible question-answering.

Our work is similar to other research in spoken document retrieval (SDR), which concerns with the retrieval of audio documents based on their speech recognition outputs. Several studies [11, 14, 25] have shown that such systems could tolerate up to 30% of speech recognition errors with no significant reduction in accuracy of document retrieval. There are several reasons for this result [25]. First, many important terms are repeated many times in the spoken documents and are thus more likely to be correctly recognized. Second, if there are many words in the question, missing one or two of them may still permit the retrieval of the documents. Thus the overall retrieval at the document level will still be fairly good despite of high recognition error rate. For QA, however, the situation is quite different as important sentences will be lost if such errors are not corrected at the sentence level.

The use of phonetic matching approach to correct speech recognition error is employed in audio document retrieval using speech queries. As one of the main sources of errors in speech recognition come from substitution, confusion matrix has been used to record confused sound pairs in an attempt to eliminate this error. Confusion matrix has been employed effectively in spoken document retrieval [23] and to minimize speech recognition errors [22]. However, such a method will bring in many irrelevant terms, when they are used directly to correct speech recognition errors [17]. Instead, many SDR systems use confused sounding pairs to expand spoken language query, or statistics of repeated sounds to correct substitution errors. In our research, we focus on using the known name list, along with phonetic and OCR text matching techniques to correct substitution errors.

7. CONCLUSION

Many users are interested in searching for *information*, while the current video retrieval engines are designed to return only *video documents*. There are many simple factoid questions posed over news video collection where the users expect to acquire the video segments containing short precise answers. Here we raise the topic of video question answering to support personalized news video retrieval. Users interact with systems using short natural language text with implicit constraints on the contents, duration, and genre of expected videos. Our system VideoQA returns the relevant news video fragments as the answers, supplemented by text version of latest news, summarized to the duration constraint specified by the users.

The realization of VideoQA system requires the integration of a range of technologies including: video story segmentation, speech recognition and correction to generate news transcripts, question-answering analysis to generate precise multi-sentence output, and summarization of news stories. The main contributions of this work are: (a) the extension of question answering technology to support QA in news video; and (b) the use of multi-modality analysis and external knowledge to extract news story boundaries, correct speech recognition errors, and perform precise question answering. Our preliminary tests on the 7-day of CNN news demonstrate that the approach is feasible and effective.

This work is only the beginning, more research needs to be carried out as follows. First, the system is designed to handle factoid questions. We need to extend the capability of the system to handle other types of questions such as exploratory and opinion questions. Second, we need to further reduce the speech recognition errors using name lists and other multimedia cues. Third, we need to extract better mid-level features and to associate these features with concepts in audio transcripts. This will help in identifying interesting video segments as answers. Lastly, as the current query is text-based, we will explore speech queries and incorporate interactions in next stage of our research.

8. REFERENCES

[1] Alta-Vista news web site (2003). http://news.altavista.com/

[2] Brill, E., Lin, J., Banko, M., Dumais, D. and Ng A. Data-intensive question answering. Proceedings of the 10th Text REtrieval Conference (TREC'2001), 2001, 393-400.

[3] Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. Proceedings of the 7th International WWW Conference. 1998.

[4] Chaisorn, L., Chua, T.S. and Lee, C.H. The segmentation of news video into story units. Proceeding of IEEE Int'l Conference on Multimedia and Expo-ICME 2002, Lausanne, Switzerland, Aug 2002.

[5] Chen, B., Wang, H.-M. and Lee, L.-S. Improved spoken document retrieval by exploring extra acoustic and linguistic

cues. Proceedings of the 7th European Conference on Speech Communication and Technology. 2001.

[6] Christel, M.G., Hauptmann, A.G., Wactlar, H.D. and Ng, T.D. Collages as dynamic summaries for news video. Proceedings of ACM Multimedia 2002, Juan-les-Pins, France, December 2002.

[7] Chua, T.S. and Liu, J.M. Learning pattern rules for Chinese named-entity extraction. AAAI'2002. Edmonton, Canada, Jul/Aug 2002. 411-418.

[8] Clarke, C., Cormack, G. and Lynam, T. Web reinforced question answering. Proceedings of the Tenth Text REtrieval Conference (TREC'2001), 673-680.

[9] Cormen, T.H., Leiserson, C.E. and Rivest, R.L. Introduction to algorithms. McGraw-Hill Book Company. 1990.

[10] Fujii, A., Itou, K. and Ishikawa, T. A method for open-vocabulary speech-driven text retrieval. Proceedings of EMNLP'02, 188-195. 2002.

[11] Hauptmann, A.G., Jones, R.E., Seymore, K., Siegler, M.A., Slattery, S.T. and Witbrock, M.J. Experiments in information retrieval from spoken documents. BNTUW-98 Proceedings of the DARPA Workshop of Broadcast News Understanding Systems. VA, Feb 1998.

[12] Hearst, M. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1), 33-64. March 1997.

[13] Hsu, W. H.-M. and Chang, S.-F. A statistical framework for fusing mid-level perceptual features in news video. ICME 2003, Baltimore, USA, July 6-9, 2003.

[14] Johnson, S.E., Jourlin, P., Moore, G.L., Sparck Jones, K. and Woodland, P.C. The Cambridge University spoken document retrieval system. Proceedings of ICASSP. 49-52. 1999.

[15] Leacock, C., Chodorow, M. and Miller, G. Using corpus statistics and WordNet for sense identification. Computational Linguistic, 24(1), 147-165. 1998.

[16] Lee, C.H. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication, 25, 29-47. 1998.

[17] Ng, K. Information fusion for spoken document retrieval. Proceedings of ICASSP'00, Istanbul, Turkey, Jun 2000.

[18] Roth, R. A SnoW-based shallow parser. A software down-loaded from: http://l2r.cs.uiuc.edu/~cogcomp/. 2003.

[19] Salton, G. and McGill, M.C. Introduction to information retrieval. McGraw Hill, 1983.

[20] Satoh, S., Nakamura, Y. and Kanade, T. Name-it: naming and detecting faces in news videos. IEEE Multimedia, Jan 1999, 22-35.

[21] Seymore, K., Chen, S., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishhankar, M., Rosenfeld, R., Siegler, M., Stern, R. and Thayer, E. The 1997 CMU Sphins-3 English broadcast news transcription system. Proceedings of the 1998 DARPA Speech recognition Workshop. 1998.

[22] Shen, L., Chai, H., Qin, Y. and Tang, D. Character error correction for Chinese speech recognition system. Proceedings of International Symposium on Chinese Spoken Language Processing Symposium. 136-138, 1998.

[23] Singhal, A. and Pereira, F. Document expansion for speech retrieval. Proceedings of the 22nd Annual International ACM SIGIR conference on Research and Development in Information Retrieval, 34-41. 1999.

[24] TRECVID. TREC video retrieval evaluation home page. http://www-nlpir.nist.gov/projects/trecvid/. 2003.

[25] Van Thong, J.-M., Moreno, P.J., Logan, B. Fidler, B., Maffey, K. and Moores, M. Speechbot: an experimental speech-based search engine for multimedia content on the web. IEEE Trans on Multimedia, 4(1), 88-96. 2002.

[26] Voorhees, E.M. Overview of the TREC 2002 question answering track. In notebook of the Eleventh Text REtrieval Conference (TREC'2002), 115-123. 2002.

[27] Wactlar, H.D., Hauptman, A.G., Christel, M.G., Houghton, R.A. and Olligschlaeger, A.M. Complementary video and audio analysis for broadcast news archives. Communications of the ACM. February 2000, 43(2), 42-47.

[28] Wang, G. Chua, T.S., Wang, Y.C. Extracting key semantic terms form Chinese speech queries for Web searches. 41st Annual Meeting of the Association of Computational Linguistics (ACL'03), Sapporo, Japan, July 2003, 248-255.

[29] Yang, H. and Chua, T.S. The integration of lexical knowledge and external resources for question answering. The Eleventh Text Retrieval Conference, TREC 2002. Gaithersburg, Nov 2003. 486-491.

[30] Yang, H., Chua, T.S, Wang, S. and Koh, C.K. Structured use of external knowledge for event-based open-domain question-answering". 26th Int'l ACM SIGIR Conference' 03. Jul/Aug 2003. Canada.