CLVQ: Cross-Language Video Question/Answering system

Yu-Chyeh Wu
National Central University,
Jhongli City, Taoyuan,
Taiwan
bcbb@db.csie.ncu.edu.tw

Chia-Hui Chang
National Central University,
Jhongli City, Taoyuan,
Taiwan
chia@csie.ncu.edu.tw

Yue-Shi Lee
Ming Chuan University,
Gwei Shan District, Taoyuan,
Taiwan
leeys@mcu.edu.tw

Abstract

Multi-Language information retrieval promotes users to browse documents in the form of their mother language, and more and more peoples interested in retrieves short answers rather than a full document. In this paper, we present a cross-language video QA system i.e. CLVQ, which could process the English questions, and find answers in Chinese videos. The main contribution of this research are: (1) the application of QA technology into different media; and (2) adopt a new answer finding approach without human-made rules; (3) the combination of several techniques of passage retrieval algorithms. The experimental result shows 56% of answer finding. The testing collection was consists of six Discovery movies, and questions are from the School of Discovery web site.

1. Introduction

Question/Answering on multi-media is a new research issue in recent years. Traditional Q/A systems aim to find exact and short answers in the form of phrases or sentences. Text REtrieval Conference (TREC) have launched QA track to evaluate performances among participants. Finding out exact answers instead of document lists became the main goal of the QA task. However, most of current QA systems return their answers just in text-type. But more and more users attempt to explore multimedia data as they retrieve documents as usual.

As we know, video is the most typical prototype of multimedia, since it contains not only texts, but also voice, and image parts. However, there are few researches [8] [16] focused on this issue. (Oh and Bandi, 2002) defined three kinds of video data:

(1) Produced type: like news, movies.
(2) Raw type: like the surveillance, and traffic video.
(3) Medical type: like the Echocardiogram.

The produced type of data often contains rich and diverse information. For example, the Discovery movie, Napoleon, described several important events that occurred in Napoleon’s life. They included the experience of his growing, his wife Josephine, and some important fights when he became the emperor of France.

The second type is to monitor the abnormal situations among all frame sequence. Most of image processing researchers addressed large work on motion detection and pattern recognition. Their target goal is to detect the exceptional behaviors in a given frame sequence, which was called key frame detection. For example, [10] developed system, which could online cluster similar frames and find the key frames. The medical type of video data is usually used to identify organs or objects in animals.

A Discovery movie contains rich information and knowledge since it explores a specific topic and discovers the hidden science for its main topic. It is useful for general populations to watch the educational videos. However, to construct a QA system based on the Discovery knowledge base aims to return a list of shots and short answers instead of pure text media. When querying a question, for example: “What is the diameter of the Comet?” users often want to inquire not only the text-type answers, but also would like to watch the objects in further. Our main idea is to provide such a QA system that could couple with traditional nature-and-cross language questions and displays answers in shots of video.

Generally speaking, there are four important components in QA systems:

(1) Question analysis
(2) Document retrieval
(3) Candidate extraction (Passage retrieval)
(4) Answer extraction

The question analysis component determines the
question type and finds out keywords. In other words, given a question, it will be categorized into question category, and lists the relevant keywords. Secondly, the document retrieval component retrieves a list of relevant documents, which contain at least one keyword. Then, each document is decomposed into passages, and candidate extraction component tries to remove irrelevant passages, and retains/ranks the answer candidates, which include the answers. Finally, the answer extraction component searches the possible answers among these passages.

The cross-language QA system requires three fundamental problems. First is the video processing, i.e. video Optical Character Recognition (OCR); several related researches [8] [13] [17] had been done on extracting texts in videos. The second problem is video segmentation. Traditional QA system aims to return a list of short answers, but the short answers in video type maybe too short to watch. Thus we adopt a simple segmentation approach to split videos into several units. When watching the answers, CLVQ will display relevant answers in short passages.

In addition, we need to solve the translation and text problems between English-Chinese. There is a main problem of Chinese word segmentation, differ from English, there are no bounds between terms. In order to make our cross-language QA system more precisely, we extracted keywords from answers and use a term translation between English and Chinese. Thus, when answers are extracted by the QA system, we need to apply to Chinese video data; further discussions will be shown in section 4.

The remainder of this paper is organized as follows: Section 2 describes the overview of our system which contains three parts: OCR module, Translation module, and QA module. Section 3 introduces OCR module for identifying words from images. Section 4 describes the translation strategy in this paper, and section 5 explains to the QA module based on the results of OCR and machine translation. Section 6 discusses experimental results. Finally concluding remarks and future work are given in section 7.

2. Overview of CLVQ

The cross-language video QA diagram of CLVQ is given in Figure 1. While videos (like movies or news) as raw data input to the CLVQ, the OCR module is trigger to recognize all of the caption words at first. Then, we use other translation resource to help to translate the Chinese text into English. Finally, an English-based QA module is built on the translated text. When input a nature language question, the CLVQ will return a list of answers, and the positions of them in videos.

Figure 2 is one of the images of the “Napoleon”. We treat all of the words which appear in the same frame as a sentence, for instance, all of the words (Chinese mean: “拿破崙波拿巴”; English means: “the Napoleon Bonaparte”) in Figure 2 would be viewed as a sentence. After words identification, the QA module will return a list of answers in the form of sentence like Figure 2.

For example, a question (from the School of Discovery1): “Analyze Napoleon's role in the French Revolution and his speedy rise to power. What were his talents?” inputs to CLVQ, then the QA module will start to find the answers in the video corpus. After QA processing, the answers will be listed in decreasing order i.e. the higher possibility the sentence, the higher rank it is. Figure 3 shows the interface of CLVQ. The left part of Figure 3 is the answer player, which can display the shot of the answer that user choose.

Figure2. A frame image of extracted from “Napoleon”

1 http://school.discovery.com/
3. OCR Module

OCR Module processes all of the input frame sequences and identifies all of the caption words. Figure 4 shows the key processing flow of this module.

Similar to other video processing researches [8] [13], the first step of OCR module is to decompose input video into list of frames, then the kernel module starts to with the following steps: filtering, representing, and character recognizing. These three steps will be described in the next section.

3.1. Filtering

Filtering is an important processing before identify characters. Briefly speaking, this stage could remove the noisy blocks. We employ some of the well-known techniques, like: large non-relevant areas removing, extraordinary line deletion and multi-frames cleaning, details can be seen in [8] [13].

For example, Figure 5 (a) shows the result of binarizing and removing large black areas from Figure 2. Figure 5 (b) is the result after extraordinary line deletion. Images which pass through filtering techniques can kithe the text area clearly.

3.2 Representing

It should be noted that the OCR task was built on single character recognition. In order to separate each single word, we made up some heuristics to split each character. The single word identification can be considered as the classification task; each character can be represented to a vector space model, then, this vector will be the used to find the category of this vector which was constructed as the training corpus.

In CLVQ, we use several kinds of features to
3.3. Character recognizing

Each single character can be viewed as a vector, and this step is used to find the most relevant characters in training set. As mentioned in 3.2, the OCR task is another kind of classification. To solve the classification problem, an algorithm is used to find the most relevant category which suits for a testing vector. There are many satisfied classification algorithms, like Support Vector Machines (SVM), Neural Network [12], k-Nearest Neighbors (kNN) [2], and Decision Trees. In this paper we used a kNN to categorize each character.

4. Machine Translation

The second problem of the cross-language video QA system is the strategy of machine translation. In this section, we will compare and discuss two common translation strategies.

4.1. Query Translation

The simplest translation strategy in a cross-language information system is to translate the query terms into another language in query time, and use a mono-lingual QA system to process the translated question. However, query translation usually suffers from the lack of context information of question terms, i.e. ambiguous senses among words. For example, Table 2 shows a translation result of a question from the School of Discovery.

<table>
<thead>
<tr>
<th>Video Title: “Peter the Great”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original English Question</strong></td>
</tr>
<tr>
<td><strong>Translate to Chinese</strong></td>
</tr>
<tr>
<td><strong>English means of the Chinese translation</strong></td>
</tr>
<tr>
<td><strong>Correct Chinese Sense</strong></td>
</tr>
</tbody>
</table>

The incorrect sense of the word “order” occurs in the question, and it has six meanings in Chinese. In this question, it should be assigned to the meanings of “punish”, but it was translated into the sense of “reserve”.

It is obviously that the more context information the ambiguous word, the higher quality the translation result. Comparing with the second translation strategy (explains in section 4.2), the context information is less than a full article. Thus, we select another approach to translate Chinese text.

4.2. Document Translation

In this approach, all text had been translated, thus the QA module just tries to find the answers within the translated text. (Wang and Orad, 2001) compared several strategies of machine translation in cross-language information retrieval (CLEF), the experimental results showed that the document translation is more precise than query translation. In this paper, we use the strategy of document translation, and translated all Chinese text into English. After translating, the QA system was built on the English text, users could ask their nature language question in English, CLVQ will list the possible answers both in English and Chinese. The framework of document translation was shown in Figure 1.

5. QA Module

In this paper, we adopt three important steps in our QA module, Question Analysis, Passage Retrieval, and Answer Selection. Figure 6 is the processing flow of the QA module. The functionality of each component in Figure 6 had explains in section 1. There is an optional component “Document Retrieval”, which is unnecessary in our experiments but useful in open-domain QA system. This is because the testing question from the School of Discovery is designed for a specific video, and answers can be looked up in this movie.

![Figure 6. Framework of QA Module](image)

Therefore, we just focused on finding out answers from a single video rather than multiple videos. We defined two displaying types of answers in this paper:

1. Single answer list from single document
2. Multiple answer lists from multiple documents

Single answer list and single document required a
given video to be extracted in local. The second displaying type of answer will extract all of the videos in database, and answers may be comes from different video in global. As mentioned above, the School of Discovery just provided questions for a certain video, thus the output of each answers will be formed in the first displaying type.

5.1. Question Analysis

The first step of QA processing is to identify the question type of a given question. The result of question analysis is used to find the possible answer type. The question type can reduce the answer candidates, for example, given the following question:

Question: “Who is Napoleon’s wife?”

The question is asked a “person”, therefore, the answer selection component will remove those answer candidates which does not contains any “person” information.

In addition, question analysis can be viewed as a classification problem [1] [7] [18]. (Li and Roth, 2002) defined 2-level hierarchical question types, there are six coarse and 48 fine classes, and about 5500 questions for training, 500 questions (from TREC-10) for testing. We had developed an instance-based question classification approach, which could achieve 82.7 % F1-measure on question classification. In this paper, we employ this component to classify input questions.

5.2. Passage Retrieval

Traditional document retrieval system, such like: google, yahoo…etc, aims to find the most relevant documents that respected to user’s queries. But the length of each document is unstable, from several tens of words to hundreds of thousand terms. It is unapplicable to extract a short answer from those long documents. Therefore, a passage retrieval component can analyze the relevant document which is retrieved from document retrieval and evaluate the relation strength between question and passages. Then answer extractor searches the possibly answers among these passages.

Currently, there are some researches had been addressed on the performance of passage retrieval algorithms [4] [6]. (Tellex et al., 2003) compared eight and one (combined approach) passage retrieval algorithms in their study; they reported three top-performance algorithms: IBM [4], SiteQ [6], and ISI [3]. Each of them had three common features:

1. (1) Density-based ranking
   2. (2) Keyword weighting
   3. (3) Query expansion

A density-based ranking method is different from traditional vector space model-based similarity scoring function. Traditional similarity scoring function, like Eclead, and Hamming distance, estimates the similarity between given two vectors. However a density-based ranking method, calculates the distance between matched keywords for each passage, when keywords appear frequently in a passage, it may contain the answer possibly. Secondly, the keyword weighting give different levels of matched keywords in different degree of weight. For example, the proper noun of the question in section 5.1 is “Napoleon”, if the retrieved passages which contain the keyword would get $W_1$ score; similarly, the word “wife” is another content words in this question, if a term in the passage which matched with “wife” will gets $W_2$ score. Third, the query expansion is useful for some definition questions, for example:

Question: “What is an atom?”

Because the lack of information of the keyword “atom”, the passage retrievers will back off a document retriever. In order to earn more information from this dried-up keyword, an external knowledge base should be used to extend extra knowledge of this word. In this case, we use WordNet, a remarkable thesaurus to expand all of the keywords, for the above example: the result of atom in WordNet is shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3. Translation result of question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query: atom</td>
</tr>
<tr>
<td>2 senses of atom</td>
</tr>
<tr>
<td>Sense 1</td>
</tr>
<tr>
<td>atom --&gt; ((physics and chemistry) the smallest component of an element having the chemical properties of the element)</td>
</tr>
<tr>
<td>substance, matter --&gt; (that which has mass and occupies space; &quot;an atom is the smallest indivisible unit of matter&quot;)</td>
</tr>
<tr>
<td>Sense 2</td>
</tr>
<tr>
<td>atom, molecule, particle, corpuscle, mote, speck --&gt; ((nontechnical usage a tiny piece of anything)</td>
</tr>
<tr>
<td>material, stuff --&gt; (the tangible substance that goes into the makeup of a physical object; &quot;coal is a hard black material&quot;; &quot;wheat is the stuff they use to make bread&quot;)</td>
</tr>
</tbody>
</table>

There are two senses in WordNet, all of the possible synonyms of query term “atom” is listed in Table 3.
The query expansion extracts all of the hypernyms, synonyms, and hyponyms from WordNet, (in this case, “substance”, “matter”, “material”, and “stuff”) would be added into the keyword list of this question. Then, the retrieval algorithm will use the extra information to improve the performance of passage retrieval.

In this paper, we construct the passage retrieval algorithm based on the three features, Table 4 mentions to our passage retrieval algorithm of CLVQ.

Table 4. Passage Retrieval Algorithm of CLVQ

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identifies Name Entities and Question Focus term in given question.</td>
</tr>
<tr>
<td>2</td>
<td>Extracts all of the content words of synonyms, hypernyms, and hyponyms in WordNet.</td>
</tr>
<tr>
<td>3</td>
<td>Weights each term of each passage according to the weighting score definition.</td>
</tr>
<tr>
<td>4</td>
<td>Calculates the density score of each passage.</td>
</tr>
<tr>
<td>5</td>
<td>Select top N passages as the answer candidates.</td>
</tr>
</tbody>
</table>

Weighing Score Definition:

- \( W_1 \): Name Entity match
- \( W_2 \): Question Focus term match
- \( W_3 \): Question term exact match
- \( W_4 \): Stem match
- \( W_5 \): Synonym match
- \( W_6 \): Hyponym match
- \( W_7 \): Hypernym match

The density score function of \( S_i \) (according to (Lee et al., 2001)) is shown as follows:

\[
Score(S_i) = \sum_{j=1}^{k} \frac{W(t_j) + W(t_{j+1})}{\text{dist}(j, j+1)^{1/2}} \times \frac{k}{k-1}
\]

\( W(t_j) \) is the weighting score of term \( j \); the \( k \) is the number of matched terms between question terms and passage terms. The component “\( \text{dist}(j, j+1) \)” means the number of words between matched term \( j \), and matched term \( j+1 \).

In this paper, the passage retrieval algorithm returns top 30 passages as the answer candidates, and answer candidates searches the final answers from these 30 passages.

5.3. Answer Selection

Passage retrieval provides a list of answer candidates, and the main purpose of answer selection is to search the exact answer among these candidates. There are various kinds of answer selection types:

- (1) keyword-based
- (2) Pattern-based
- (3) Rule-based
- (4) Learning-based

The simplest way of answer selection is to find the pre-defined keywords in answers. Second, the pattern-based approach is similar to the first type, but it contains more context information near by the keyword. Third, rule-based answer selection makes use of a syntactic parser, and analyze each sentence with its syntactic relations. Fourth, the learning-based method, tries to build up a knowledge model based the matched relations between question terms and passages, and learns to find out answers from pre-constructed knowledge model.

The first three types required to made large number of patterns, rules, even keywords manually. This is a heavy work for constructing a human-made knowledge base. However, there are some difficulties in the learning-based answer selection strategy, like the knowledge representation, and the resource of training data. Furthermore, in this paper, we adopt an answer selection strategy, which combines some advantages of these approaches without human-made rules.

Type (1) (2) (3) extract answers in the form of pattern match; while type (4) used a learning approach to build up a knowledge model to find the answers. Our idea is to extract the expected answer types of each question class from training data. Expected answer types are often associated with specific name entity classes, thus we need to construct a many-to-many mappings between name entities and question types. Figure 7 illustrates some of the constructed mapping relations.

For example, the question “Who is Napoleon’s wife?” was assigned to the question class: “HUMAN: Individual”, and the expected answer type in the form of name entity is “PERSON”, then, we search all of the passages, which contain this entity class as the final answer.

Previous study [8] [11] [17] constructed the mapping relations manually. However, we use a simple statistic technique to build such a mapping relations instead of human construct. The main idea is to amount each of the entity frequency in question class; if the probability...
of a name entity class is higher than a threshold, then makes a connection between this question and name entity class. In this paper, we collect 492 answers from TREC-10 QA main task, and construct a mapping relation map by calculating the probability of each name entity for each question class.

6. Experimental Result

6.1. Dataset

Video films used for testing our approach are the descriptive video films from Discovery. All the features of our testing corpus were listed in Table 5. At the beginning, all films were converted into MPEG-1 format, with resolution 352x240 pixels per frame. Most of the qualities of the video frames are satisfied with filtering (as mentioned in 3.1).

Table 5. Features in video corpus

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of video films</td>
<td>30</td>
</tr>
<tr>
<td>Total number of sentences</td>
<td>16126</td>
</tr>
<tr>
<td>Total number of characters</td>
<td>141441</td>
</tr>
<tr>
<td>Average number of sentences per video film</td>
<td>537.53</td>
</tr>
<tr>
<td>Average number of words per video film</td>
<td>4714.7</td>
</tr>
</tbody>
</table>

The short description of discovery movie could be viewed on the web site (http://www.discovery.com).

6.2. Questions

We collect 47 questions of six films from the School of Discovery web site. The films are “Ice Age”, “Peter the Great”, “Byzantine Empire”, “Napoleon”, “The Great Wall”, and “Bears”. The MRR (Mean Reciprocal Rank) scoring function is used to evaluate the performance of the QA result in TREC-QA task. MRR was measured over the top five response answers, and give higher scores for higher rank.

6.3. Results on OCR

Table 6 lists the OCR performance on inside and outside testing results, and each character had been identified by human recognition. The OCR testing data: “Napoleon” which describes important events of Napoleon and some great wars that lead by him. For outside test, the OCR module identifies each character with trained and non-feedback character database; while inside test makes use of all the character database contains original and feedback ones. Comparing with advanced researches [12], the OCR results in lower performance, because we did not employ a sophistical filtering and feature selection approaches. Due to the OCR error, the quality of machines translation, and QA performance were limited. In the future, we will combine more advanced OCR approaches, and combine the sound, and image features, to decrease the damages for the downstream components.

Table 6. OCR result on “Napoleon”

<table>
<thead>
<tr>
<th>Test name</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inside test</td>
<td>92.1%</td>
<td>7.6%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Outside test</td>
<td>81.5%</td>
<td>16.3%</td>
<td>2.0%</td>
</tr>
</tbody>
</table>

6.4. QA results

The experimental result is listed in Table 7, and the MRR value, for example: “The Great Wall” was 0.146 ((1/4+1/2+1/10+1/6+1/4+1/5)/10); while the accuracy is 0.6 (6/10).

Table 7. QA performance

<table>
<thead>
<tr>
<th>Video name</th>
<th>MRR</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ice Age</td>
<td>0.156</td>
<td>0.571</td>
</tr>
<tr>
<td>Peter the Great</td>
<td>0.309</td>
<td>0.63</td>
</tr>
<tr>
<td>Byzantine Empire</td>
<td>0.324</td>
<td>0.428</td>
</tr>
<tr>
<td>Napoleon</td>
<td>0.426</td>
<td>0.666</td>
</tr>
<tr>
<td>The Great Wall</td>
<td>0.146</td>
<td>0.6</td>
</tr>
<tr>
<td>Bears</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Because such a cross-language QA system integrated technologies from many research areas, however there are many errors occurs and affect the performance of CLVQ. These main errors are:

1. Character identification error during the OCR stage. In this paper, the accuracy of OCR is about 81.2%; in the future we will employ more sophistical technology to minimize these errors.
2. Translation error causes the ambiguous words in English documents. Word Sense Disambiguation (WSD) is the main problem in machine translation system. The more rich the context, the more accurate the WSD is. An example of WSD error can be seen in Table 2. We will improve WSD performance with external thesaurus, like: Chinese-English bilingual dictionaries.
3. The misclassification during the question analysis stage, and the mapping error in the answer selection stage. Due to the length of limitation, we did not report our approach of question analysis, which achieve 82.7% F1-measure on TREC-10 testing data, as opposed to 80.2% of (Zhang, and Lee, 2003), 82.0% of (Hacioglu, and Ward, 2002) and 84.2% of (Li, and Roth, 2002). The classification error of the answer mapping relation causes the answers cannot found. We also
try to combine more general but effective answer selection approaches to decrease the errors at QA stage.

7. Conclusion

Many users are interested in searching for information in different media, such like: text, image, sound, and video. However current video retrieval systems are designed to return a list of video documents rather than exact answers. Besides, there is not such a multimedia system can process queries in different language. In this paper, we adopt a cross-language video QA system: CLVQ, which aims to find the exact answers that corresponding to the users’ nature and cross language questions. We collect six discovery movies and testing with the public questions that come from the School of Discovery web sites. Average speaking, the accuracy of the returning top five answers could achieve about 56%, while 28.1% is the MRR value. In the future, we would improve some key-components, like: OCR, and machine translation, to make better qualities on QA.

References


