

Automated Question Answering From Videos: NLP vs. Pattern Matching

Jinwei Cao
University of Arizona
jcao@cmi.arizona.edu

José Antonio Robles-Flores
Arizona State University / ESAN
Jose.Robles@asu.edu

Dmitri Roussinov
Arizona State University
Dmitri.Roussinov@asu.edu

Jay F. Nunamaker, Jr.
University of Arizona
jnunamaker@cmi.arizona.edu

Abstract

We have explored the feasibility of automated question answering from video material used in conjunction with PowerPoint slides. We have tested separately and combined two recently becoming popular approaches to question answering: 1) the approach based on deep Natural Language parsing and 2) a self-learning probabilistic pattern matching approach. The advantages and shortcomings of each of the approach are discussed. We present the results of the comparison and our qualitative observations.

1. Introduction

Learning is an important way of transferring knowledge, and effective learning is highly desired in the current knowledge-based economy. In recent years, advances of information technology have affected learning dramatically. According to the IEEE Learning Technology Standards Committee (LTSC, <http://ltsc.ieee.org>), *e-learning*, referring to the use of computers and network technology to create, deliver, manage, and support learning at anytime, anywhere, has become widely adopted as a promising solution to lifelong learning and on-the-job workforce training. Thousands of online courses, including degree and certificate programs, are now offered by universities world-wide. For example, in 2001, MIT announced its commitment to make material from virtually all of its courses freely available on the Web for non-commercial use (<http://web.mit.edu/newsoffice/nr/2001/ocw.html>). It was predicted that U.S. online education student enrollment will top one million by 2005 [1]. At the

same time, companies are also increasingly relying on online corporate training. For example, Circuit City provides customized e-learning courses about product knowledge, sales skills, and customer service to over 40,000 associates and managers with the help from DigitalThink, an e-learning solution provider (<http://www.digitalthink.com>). According to IDC, e-learning vendors that sell into the enterprise space are expected to post more than \$23 billion in revenues by 2004. The fundamental value proposition of e-learning – access to quality education or training freed from the boundaries of time and location – is growing with the demand for higher education and professional training in the United States and worldwide.

In current online education or training programs, multimedia learning materials such as videotaped lectures and PowerPoint slides are more and more commonly provided as a way to help learners engage in the learning process. For example, in online courses provided in Stanford University (<http://scpd.stanford.edu/scpd/students/onlineClass.htm>), a video of an instructor is synchronized with his/her PowerPoint slides. Multimedia lectures are considered to be able to give students a perception of listening to a lecture in real-time, make them pay more attention to the learning task, and help them retain more information through vivid and rich presentations [2]. However, simply watching a mentor talking in a lecture video is still quite different from learning with a real mentor. An important factor of learning, “learning interactions”, is usually missing in multimedia online lectures and often results in higher dropout rates as compare to classroom learning [3]. For example, a student in classroom learning can ask questions and get them answered by the instructor. In a more ideal learning scenario that a student learns from a private mentor, not only can the student ask

questions at any time, but also the mentor will ask the student questions and assess the student's knowledge based on his or her answers and thus provide customized instructions, which may trigger more questions from the student.

Virtually any science fiction work depicting the future, from Spielberg to Asimov, includes scenes where people converse with a machine in natural language to get answers to their questions. This interaction has been a dream of artificial intelligence (AI) researchers as well as ordinary people since the invention of computers. Recent advances in Natural Language Processing (NLP) and AI in general have approached this dream world to the point where it mixes with reality. Several known futurists (people who make their living entirely by trying to predict "future") believe that computers will reach capabilities comparable to human reasoning and understanding of languages by 2020 [4].

The goal of automated Question Answering (QA) is to locate, extract, and represent a specific answer to a user question expressed in natural language. A QA system would take as input a question like "What is mad cow disease?" and it should get as output "Mad cow disease is a fatal disease of cattle that affects the central nervous system. It causes staggering and agitation." Modern Question Answering (QA) technologies rely upon many components, including document retrieval, semantic analysis, syntactic parsing and explanation generation. QA promises an important new way of information access for all, a natural step beyond the keyword query and document retrieval characteristic of today's information quests, such as those on the web.

The TREC question answering evaluation [5] is the motivating force behind a recent surge in question answering research. Systems participating in TREC have to identify exact answers to typically "factual questions" (who, when, where, what, etc.). Most of TREC QA systems are designed based on techniques from natural language processing, information retrieval and computational linguistics. For example, Falcon [6], one of the most successful systems, is based on a pre-built hierarchy of dozens of semantic types of expected answers, complete syntactic parsing of all potential answer sources, and automated theorem proving to identify the answers.

In addition to "deep" linguistic approaches, QA researchers explored more "shallow" approaches ground on pattern matching successfully used earlier for information extraction [7]. Pattern matching systems performed well in recent TREC QA competitions: the system from InsightSoft [7] won 1st

place in 2002 and 2nd place in 2001. Roussinov & Robles [8] studied a pattern based system that extends the redundancy based approach tried in [9] with the completely automatically learned patterns. The system achieved comparable performance with one of [9]. Its strong advantage over other techniques that it is completely trainable and does not require any manually developed rules or substantial linguistic resources.

However, most of current research studies focus on QA from a collection of text documents [10]. For the scenario of students asking questions to an instructor captured in video, we need QA from a collection of video files. Nevertheless, research on video-based QA is just initiated, with a focus on news video collection [10]. The algorithms used in these QA systems, such as video genre analysis, cannot be directly applied to the lecture video collection because the latter is quite different from news video. For example, lecture videos usually have very few scene changes (e.g. only a "talking instructor" on the screen); speeches in lecture videos are typically unscripted and spontaneous; and more importantly, speeches in lecture videos normally have many domain-specific vocabularies (e.g. UTP as Untwisted Pairs in the domain of data communication) which cannot be found from a general knowledge source such as WordNet [11].

In our study, we have explored the applicability of two text-based automated approaches, namely natural language based approach and pattern based approach, for the purpose of automated question answering from lecture videos. Our motivation was to first see if providing such a feature within a proof of concept prototype is feasible. Second, we wanted to compare two QA techniques which are inherently different using a set of test questions and correct answers to them. Our prototype system converts the sound track into transcripts by using speech recognition software and applies several enhancement methods including phonetic-based transcript error correction and external domain knowledge. Next section reviews the explored QA approaches, followed by the section describing our transcribing process, then by the section on our simulated comparison tests.

2. Technology involved

2.1. Natural language approach

As we mentioned earlier, most of modern Question Answering systems are designed based on techniques from natural language processing and information retrieval. In a typical QA system, a user submits a natural language question, e.g. "What is mad

cow decease?” Documents or passages are retrieved from the target collection by their relevance to the query using standard information retrieval techniques (e.g. those including words “decease” and “cow”). Most QA systems follow the following three steps [12]:

1. *Question Understanding* includes recognizing the type of a question, e.g. “Who is the CEO of IBM” expects the answer to be a person.

2. *Document Filtering* uses keywords in the question as query and some query expansion mechanisms (e.g. morphological or lexical, using WordNet) to perform retrieval. Only the retrieved documents are processed to extract candidate answers.

3. *Answer Extraction* performs a shallow parse of the returned documents to detect entities of the same type as the answer. Those entities are later treated as candidate answers and assigned scores based on the relevance between the processed query and sentences in the documents extracted in step 2. (E.g. “Samuel Palmisano”)

Zhang and Nunamaker [13] applied a natural language based QA approach to a collection of transcribed videos. In that approach, the transcripts of the manually segmented video clips are treated as text documents, and a similar three-step approach is used to identify the answers. However, unlike some TREC QA systems such as Falcon [6], it uses a template-based approach for question understanding and answer extraction. Such a template-based approach does not rely on complicated deep semantic analysis such as automated theorem proving in Falcon, and is believed to be more suitable for possibly spontaneous speech text. The approach uses a parser called Conexor iSkim [14]. Major verbs, nouns, noun phrases, named entities in the question are extracted from the iSkim output. Their synonyms are found from the WordNet dictionary. A query is formed using the original words, their synonyms and the named entities. The answer type of the question is also derived and the question is filled into a question template with nine slots, including answer type, question focus, person, organization, governor, objects, number, time, and location (Figure 1) [13].

```
< QUESTION TEMPLATE >:=
Answer Type (type of information
a question is looking for)
Question Focus (the core noun)
Person (named person)
Organization (named organizations)
Governor (key verbs)
Objects (other noun or noun phrases)
Number (numbers)
Time (year, date, etc.)
Location (country, region, city, etc.)
```

Figure 1. Question Template. From [13].

The basic query is then sent to a Boolean information retrieval engine. The search results are processed to extract answers. Sentences in the retrieved documents are then parsed using iSkim and transformed into sentence templates (*ST*) in a similar form as the question template (*QT*). Similarity between the *QT* and the *ST* is calculated based on the combination of the following three factors [13].

Matched_Slots_Score (MSS) compares the slot values of QT with STs.

Same_WordSequence_Score (SWS) computes the number of keywords in the question that appear in the same sequence in the current sentence.

AnswerType_Found_Score (AFS) checks if either Answer Type or Question Focus of the question is found in the current sentence.

Finally, a sliding-window method is used to calculate the total similarity between the question and each five-sentence window in the document. The highest score is taken as the relevancy score of the document and the top relevant documents are returned.

This natural language based approach works well in a limited domain such as lecture videos. In this scenario, the interpretations of questions are clear and the deep parsing and understanding of sentences is feasible in real time since the set of retrieved documents is small.

2.2. Pattern based approach

We adapted the pattern based QA approach described in Roussinov & Robles [8] to our system. While searching for an answer to a question (e.g. “Who is the CEO of IBM?”) their approach looks for matches to certain patterns. For example “The CEO of IBM is Samuel Palmisano.” matches the pattern “\Q is \A .” where \Q is a question part (“The CEO of IBM”) and \A = “Samuel Palmisano“ is the text that forms a candidate answer. The approach automatically creates and trains up to 200 patterns for each type of a question (examples of types of questions are what is, what was, where is, etc.) based on a training set of given question-answer pairs. Through training, each pattern is assigned the probability that the matching text contains the correct answer. This probability is used in the ranking the candidate answers. \A, \Q, \p (punctuation mark) and * (a wildcard that matches any words) are the only special symbols used in their pattern language. Figure 2 summarizes their approach.

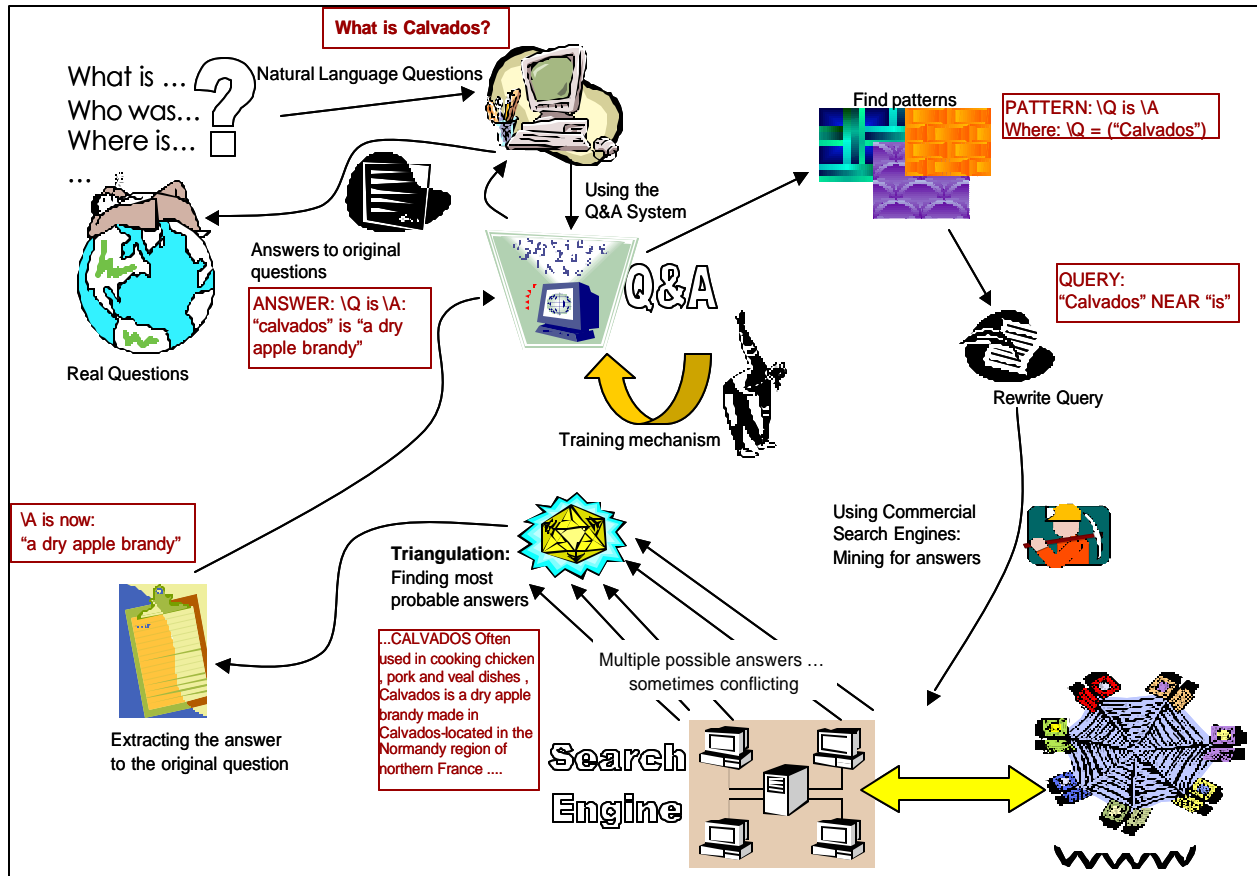


Figure 2. The general Web QA approach. From [8].

Answering the question “Who is the CEO of IBM?” demonstrates the steps of our algorithm:

Type Identification. The question itself matches the pattern who is $\backslash Q$?, where $\backslash Q$ = “the CEO of IBM” is the question part and “who is” is the type identifier.

Query modulation (although present in the original approach) is not necessary in our case since our system just scans each sentence in the transcripts.

Answer Matching. The sentence “Samuel Palmisano recently became the CEO of IBM.” would result in a match and produce a candidate answer “Samuel Palmisano recently”.

Answer Detailing produces more candidate answers by forming sub-phrases from the initial candidate answers. Our sub phrases do not exceed 3 words (not counting “stop words” such as a, the, in, on) and do not cross punctuation marks. In our example, the detailed candidate answers would be Samuel, Palmisano, recently, Samuel Palmisano, Palmisano recently.

When there are fewer than 5 matches are found, the system resorts to the “fall-back” approach by simply retrieving the segments according to their relevance to the question. For this, the questions and text segments are represented as normalized to unit length vectors in the classical vector space model [15] and the dot product is used as the relevance score.

2.3. Lecture Video Specific Issues

Figure 3 presents the overall framework of our video QA system. We have performed several implementation steps necessary specially for video-based QA. Specifically, we have added phonetic-based transcript error correction and enhanced video transcripts with the text from the power point slides that accompanied them, thus treating the slides as additional domain knowledge. Because lecture videos usually have very few scene changes the most information is in the video sound track. Because teachers typically use PowerPoint slides to outline key

concepts during their lectures and usually provide students the slides in electronic form, it is reasonable to use the text extracted from these slides as a resource of domain knowledge to improve the answer extraction performance.

3. Evaluation, Limitations and Future Research

We compare our two approaches using the archived videos of a professional training course

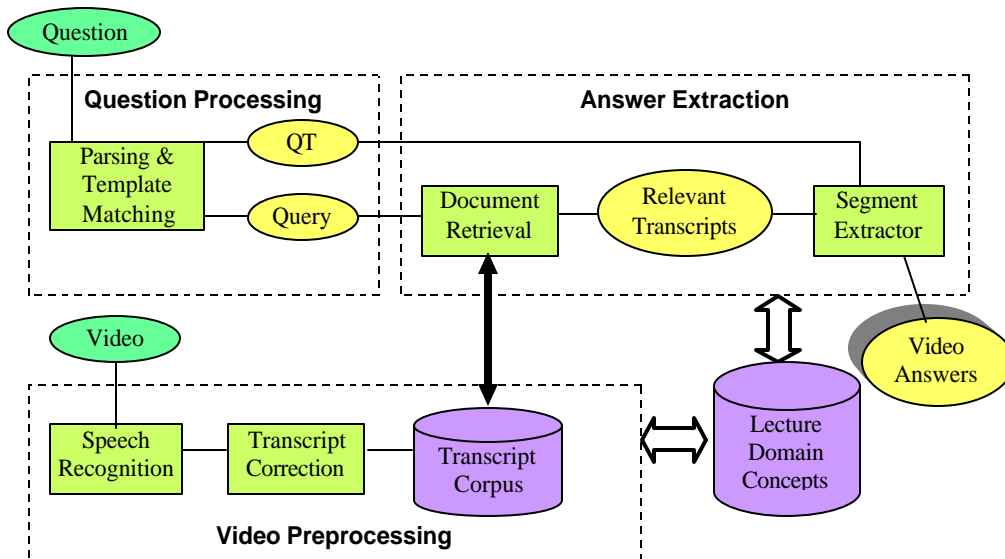


Figure 3. Overall Prototype Architecture.

The text transcripts of lecture videos are generated by a speaker-independent speech recognition tool: Virage VideoLogger® (<http://www.virage.com>). This tool also generates time stamps that synchronize the video stream with the transcribed text at word level. However, speech recognition errors could incur in the video transcripts, and these errors may greatly reduce the accuracy of retrieval. We solve this problem by doing transcript correction based on phonetic matching, a method described in [10]. A list of domain concept words is developed based on content in the domain knowledge base. Particularly, these domain concept words are extracted from the text of the PPT slides that are associated with the video. Words in transcripts are converted to phonetic sounds and are compared to the phonetic sequence of the words in this list. Similar sound words in the transcripts are changed to the one in the word list. Finally, the corrected video transcripts are indexed and stored in a database.

Once the text answers are identified, the links to the video segments that contain them are presented to the user in the rank order of decreasing relevance and can be played upon request.

(“Deception Detection”). There are two lectures (videos) in this collection, and they are pre-segmented into 100 short clips (segments). This is a relatively small collection, but human generated transcripts are available for comparison purpose. We are now in the process of developing a larger video lecture corpus for future evaluation.

To evaluate the two approaches, a PhD student who is an expert in the respective domain (“Deception Detection”) has created a set of 30 questions and manually identified the best answer segment for each of the questions. We assumed that there is only one correct answer for each question. Although various metrics have been explored, we used mean reciprocal rank of the answer (*MRR*) according to the following formula [12] to evaluate the two approaches.

$$MRR = \frac{1}{\# \text{ questions}} \sum_{i=1}^{\# \text{ questions}} \frac{1}{\text{answer}_i \text{ rank}}$$

Where answer_i rank is the rank of the first correct answer for the question i , and if the answer is found at multiple ranks, the best rank will be used. If no relevant answer is found in the top 5, the score for that particular question is zero. The highest *MRR* score is 1 and the lowest is 0. For example, *MRR* score of 0.5

can be roughly interpreted as “in average” the correct answer being the second answer found by the system.

MRR is closely related to the average precision measure used extensively in document retrieval. The drawback of this metric is that it is not very sensitive since it only considers the first correct answer, ignoring what follows.

For the natural language based approach, we have compared the results using three different sets of transcripts, and for each set, we compare the method of using PPT slides in answer extraction to the one without using PPT slides. Our results are listed in Table 1.

We conclude that the NL based approach works best for human generated transcripts plus the help from PPT slides. QA based on the transcripts directly generated by the speech recognition software has even higher MRR than QA based on human generated transcripts only. This is encouraging and shows that it is possible to get satisfying retrieval results without the time-consuming human correction by just using PPT slides. However, it is surprising that although the MRR for QA on transcripts with automatic correction is higher than QA on transcripts without error correction, using PPT slides does not increase the MRR as much as on the transcripts without error correction. One possible explanation is that we use the same set of PPT slides in transcript correction so there is not much extra knowledge in answer extraction to boost the performance. Also, the transcripts may be over-corrected and new errors may be introduced. More test questions and larger test collections may be needed in future to arrive at more reliable conclusions.

Table 1. Results of our NLP based approach

MRR		
Transcript	Transcript Only	Transcript + PPT Slides
without error correction	0.423	0.524
With error correction	0.476	0.517
human generated	0.511	0.561

Our pattern based QA approached produced *MRR of 0.622*, which seems slightly higher than of NLP. However, after analyzing our log files, we observed that only 6 questions were answered by pattern matching and the rest were by our “fall-back” (retrieval) approach. Thus, the size of the test

collection was too small to reliably investigate the performance of patterns and their importance in the overall process. Although we have not yet performed the detailed analysis at this stage yet, the increased performance of the “fall-back” IR approach over our NLP approach may be explained by the use of a different stemmer, different relevance metrics or may be just an artifact of this particular set of questions. Thus, more tests again seem to be needed in future.

Since both NLP and Pattern-matching, approaches present limitations; we plan to combine the approaches in future. NLP seems to work very well in the interpretation of questions that are related to the specific domain (a video digital library, a specific database, etc.). On the other hand, pattern-matching performs well only when there is a large set of documents with redundancy (such as very large digital libraries, large intranets or the entire WWW) but it has reduced advantages when the domain is smaller.

Since students do not always only need to recall what was on the lectures but sometimes need to go beyond the lectures, we believe that using open domain QA on the Web as part of the process can provide a more thorough solution.

The application described can also be extended for different business applications: as a business intelligence tool, as a knowledge management tool, etc.

References

- [1] Gallagher, S. Online Distance Education Market Update: A Nascent Market Begins to Mature, *Eduventures*.
- [2] Agius, H.W., and Angelides, M.C. (1999). Developing knowledge-based intelligent multimedia tutoring systems using semantic content-based modelling, *Artificial Intelligence Review*, 13, pp 55-83.
- [3] O'Connor, C., Sceiford, E., Wang, G., and Foucar-Szocki, D. (2003). Departure, Abandonment, and Dropout of E-learning: Dilemma and Solutions, *TechLearn 2003 Conference*.
- [4] Lempert, R. J., Popper, S. W., Bankes, S. C. (2003). Shaping the next one hundred years: new methods for quantitative, long-term policy analysis, RAND, Santa Monica, CA.
- [5] Voorhees, E. M., Tice, D. M. (2000) Building a question answering test collection, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece.

- [6] Harabagiu, S., et al. (2000) FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of the Text Retrieval Conference (TREC-9)*.
- [7] Soubotin, M., & Soubotin, S. (2002). Use of patterns for detection of likely answer strings: A systematic approach. In *the Proceeding of TREC 2002*.
- [8] Roussinov, D., and Robles, J. (2004b). Web Question Answering: Technology and Business Applications. In *the proceedings of 2004 American Conference on Information Systems*. August 6 – 8, New York, NY.
- [9] Dumais, S., Banko, M., Brill, E., Lin, J., and Ng, A. Web Question Answering: Is More Always Better? (2002) *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland.
- [10] Yang, H, Chaisorn, L., Zhao, Y., Neo, S., and Chua, T. (2003). VideoQA: Question Answering on News Video. In *Proceedings of the ACM conference on Multimedia (Multimedia'03)*, Berkeley, CA, November 2-8.
- [11] Miller, G. (1990). WordNet: An On-line Lexical Database. In *International Journal of Lexicography*, 3, 4.
- [12] Voorhees, E.M. (1999). The TREC-8 Question Answering Track Report. *Eighth Text REtrieval Conference (TREC-8)*, Gaithersburg, MD.
- [13] Zhang, D., and Nunamaker, J. (2004). A Natural Language Approach to Content-Based Video Indexing and Retrieval For Interactive E-Learning, *IEEE Transactions on Multimedia*, 6, 3.
- [14] Voutilainen, A. (2000). Helsinki taggers and parsers for English. In J. M. Kirk (Ed.) *Corpora Calore: Analysis and Techniques in Describing English*. Rodopi, Amsterdam & Atlanta.
- [15] Salton, G. and McGill, M.J. (1983). Introduction to Modern Information Retrieval. New York. McGraw-Hill.