Multilingual Informedia: 
Search and Summarization of Video Information on Demand from 
Combined English Language and Foreign Media Sources

SUBMITTED TO
Receiving Officer
e-mail address: “hpdstnrad@spawar.navy.mil”

Rich Laverty (619) 553-2918 laverty@nosc.mil
Frank Schindler (619) 553-2845 fschindl@nosc.mil
Bob Medearis (619) 553-6377 medearis@nosc.mil

SUBMITTED BY
Carnegie Mellon University
Pittsburgh, PA 15213

Principal Investigator
Howard Wactlar (412) 268-2571 wactlar@cmu.edu

Administrative Contact
Colleen Everett (412) 268-7674 everett@cs.cmu.edu

Contract/Financial Contact
Karen M. Faber (412) 268-5838 faber+@andrew.cmu.edu

Distribution authorized to DOD components only.
Do not distribute to DTIC or other data depositories.
Section 1

1.1 Report Title

Multilingual Informedia: Search and Summarization of Video Information on Demand from Combined English Language and Foreign Media Sources

1.2 Notices

Carnegie Mellon University will retain all rights to all intellectual property, hardware and software developed under this proposed contract, but will license them to the United States government at no charge for its own use.

1.3 Abstract

The Multilingual Informedia project develops automated systems and tools to enable multilingual and multimedia information capture, search, retrieval, summarization and reuse. Data from multiple information streams (newswires, TV and radio broadcasts, archival collections) in different languages is segmented, indexed and categorized for retrieval, summarization and on-demand integrated presentation.

1.4 Table of Contents; List of Tables, Figures and Illustrations.

Section 2
2.1 Summary ................................................................. 3
2.2 Introduction ........................................................... 3
2.3 Methods, Assumptions and Procedures ....................... 4
2.4 Results and Conclusions ........................................... 8
2.5 Recommendations .................................................. 8

Section 3
3.1 Bibliography ............................................................ 9
Exhibit A: Research Papers .......................................... 11

1.5 Preface and Acknowledgements.

We acknowledge the contribution of our colleagues at the Interactive Systems Labs at the University of Karlsruhe, Germany, and the Interactive Systems Lab at Carnegie Mellon.

1.5 List of Symbols, Abbreviations, and Acronyms

VTLN - Vocal tract length normalization
MLLR - Maximum likelihood linear regression
LM - Language model
WER - Word error rate
OOV - Out of Vocabulary
Section 2

2.1 Summary

The purpose of the Multilingual Informedia project was to develop automated systems and tools enabling multilingual and multimedia information capture, search, retrieval, summarization and reuse. The system, built on the underlying Informedia Digital Video Library system concepts, technology and infrastructure, is designed to access textual, audio (radio) and video (TV) information, to index, categorize, retrieve, summarize and analyze it, in one or multiple languages. We focused primarily on the Serbo-Croatian language to demonstrate viability and practicality of proposed concepts.

We implemented and demonstrated a prototype system that was a multilingual browser of text, video and radio material that accepts English queries and returns the most relevant Serbo-Croatian, and English language reports or segments in their original language, in full or summary form. For example, this enables the analyst to compare divergent American and foreign reporting of the same event or topic. The semantic-expansion translation that we use reconstructs all consistent meanings of words and phrases in the English query, resulting in an expanded target language query without loss of recall, but at some cost in precision.

2.2 Introduction

From its inception in 1995, the Informedia project’s goal has been to allow search and retrieval in the video medium, similar to what is available today for text only. To enable this access to video content, speech recognition is used to provide a text transcript for the audio track; and image processing determines scene boundaries, recognizes faces and allows for image similarity comparison. Everything is indexed into a searchable digital video library, where users can ask queries and receive relevant news stories as results.

The Multilingual Informedia Project pursues a seamless extension of the Informedia approach to search and discovery across video documents in multiple languages. Previously, we successfully demonstrated that current speech recognizers allow accurate information retrieval for automatically processed English news TV broadcasts. The multilingual system should perform speech recognition on foreign language news broadcasts, segment it into stories and index the foreign data together with existing English news data. This first multilingual prototype should easily be extensible to other languages.

There are three components to the Multilingual Informedia system that differ significantly from the original Informedia system:

- The speech recognizer recognizes a foreign language, specifically Serbo-Croatian.
- A keyword-based translation module transforms English queries into Serbo-Croatian, allowing a search for equivalent words in a joint corpus of English and Serbo-Croatian news broadcasts.
- English topic labels for the foreign language news stories allow a user to identify a relevant story in the target language.

We built upon an existing technological base at Carnegie Mellon that includes:

- **Speech recognition** (JANUS and SPHINX). Both are state-of-the-art large vocabulary continuous speech recognition systems and are both available as a client server architecture.
Multilingual Informedia - Final Technical Report

- **Speech translation.** The JANUS system, now handles two channel spontaneous human-human dialogs without push-to-talk buttons.
- **Information retrieval.** CMU developed the Pursuit engine used in LYCOS, one of the most universally adopted and widely used web search engines.
- **Statistical language modeling**, for speech recognition and translation, proven methods in Pangloss, Sphinx, etc.
- **Informedia digital video library**, built at CMU including automated indexing of text, audio and video, full-retrieval, initial summarization (gisting) of video segments
- **Linguistic resources**, dictionaries, phrase books in German, Spanish, French, Japanese; Serbo-Croatian accessible, including access to Serbo-Croatian, German, and other broadcasts and resources.
- **Machine translation,** KANT (large-scale accurate interlingual translation within domain), Pangloss and EBMT, example-based machine translation (general-purpose translation assistants).
- **Summarization:** Initial metrics and methods for automated on-demand summary generation for text and gisting for video.

2.3 Methods, Assumptions, and Procedures

**Serbo-Croatian Speech Recognition**

The Serbo-Croatian broadcast news recognizer was based on a Serbo-Croatian dictation system built using the JanusRTk speech recognizer developed at Carnegie Mellon. We started with a context dependent system that was trained on the broadcast news data alone, later including standardized dictation data. This dictation system yielded a 28.2% word error rate on the dictation test set. The initial broadcast news test yielded a surprising drop in recognition performance to 73.6% WER due primarily to poor and noisy recordings. We improved our models through a series of iterative studies and experiments with segmentation, adaptation, vocal tract length normalization, and language model interpolation, reaching much improved WER of 36%, while simultaneously decreasing our out of vocabulary rate.

For training, we obtained audio data from speech recording and transcription, and text acquisition from internet news texts. The audio data for the dictation system was collected in Croatia and Bosnia-Hercegovina, and read into the system by native speakers. The read utterances were checked against the original text to eliminate major errors and mark spontaneous effects. We retrieved text data from 20 different internet television, radio, newspaper and news agency sources for use in language modeling.

Since Serbo-Croatian is a highly diacritical language, this caused significant problems with both audio and text data collection and language modeling. For audio data, we used a reversible mapping into ASCII text. For example, Ć and č become C1 and c1 in ASCII text. For text data, many sites simply map diacritics onto their corresponding non-diacritical letter; for example diacritic ć and Ć becomes c, d becomes d, š is replaced with s, and ž with z, which was not a problem for native speakers. However, for language modeling we had to insert the diacritics at the right position. We solved this problem in two ways.

1. We utilized a digital Serbo-Croatian→English dictionary provided by the Diplomat Project at CMU’s Language Technologies Institute. This helped in converting texts without diacritics and for (cross language) information retrieval.
2. We also developed an algorithm based on word similarity and letter trigram models which inserts the diacritics at the right position with a fairly high accuracy level of 95%.

The overall result of being able to utilize text with no diacritics doubled the amount of text data available for language modeling.

The other data collection difficulty we encountered was large data transcription times. Transcription time varied between 13 to 18 hours per news broadcast (approx. 40 minutes) because there are no closed caption or teletext available, speakers speak very fast, many noisy segments take longer to transcribe, and acoustic labeling of the segments consumes a lot of time.

The combination of acoustic material from both the dictation and the broadcast news system resulted in the ability to train a fairly reliable automatic transcription system which generates output for the multilingual Informedia database.

In contradiction to English broadcast news systems, we did not benefit from clustering of segments, but interpolated language models on a word basis again helped boost recognition performance, while the interpolation with class based (inflection oriented) models did not lead to an improvement. Morpheme based approaches also did not decrease the word error rate.

The improvement gained from multiple language models remained at a level of about 1.5% over a single language model. Comparing the results for the dictation and the broadcast news language models, one can observe that there is little difference between them (in terms of recognition improvement), although the latter model is based on much more text data. These results were promising early on because they were already very close (in terms of performance) to English broadcast news systems, which are trained on much more acoustic data.

Reducing Word Error Rates

When transcribing broadcast news data in other languages than the most common and well studied ones, problems of broadcast quality recognition are complicated by 1) the lack of available acoustic and language data (since closed captions are typically not available), and 2) the excessive vocabulary growth in heavily inflected languages that lead to unacceptable OOV (out of vocabulary) rates. While full-form word entries lead to excessively large vocabularies (and OOV-rates), the use of morpheme-based dictionaries also offers little relief. The combination of arbitrary morphemic affixes by way of a morphemic language model leads to an over-generation of illegally inflected word hypotheses and thus increases error rates. This is especially true for languages like Serbo-Croatian and German. As Serbo-Croatian is characterized by rapid vocabulary growth due to a large number of possible word inflections, we have to deal with out-of-vocabulary rates up to 13%. This makes OOV-words a major source of recognition errors in multilingual broadcast news.

Hypothesis Driven Lexical Adaptation (HDLA) has already been shown to decrease high OOV-rates significantly by using morphology-based linguistic knowledge. We introduced another approach to dynamically adapt a recognition lexicon to the utterance to be recognized. Instead of morphological knowledge about word stems and inflection endings, we used distance measures based on Levenstein distance. Compared to the use of morphological knowledge, our distance-based approach offers the distinct advantage that no expert knowledge about a specific language is required, no definition of complex grammar rules is necessary. Instead, grapheme sequences or the phoneme representation of words are sufficient to apply our HDLA algorithm easily to any new language. With our proposed
technique we were able to decrease OOV-rates by more than half from 8.5% to 4%, thereby also improving recognition performance by an absolute 4.1% from 29.5% to 25.4% word error rate.

Foreign Language Topic Detection

After initial experiments with the Croatian news processed by the Multilingual Informedia system, it became clear that returning a foreign language result to the user was not sufficient. The users were unable to tell if a particular news clip was actually relevant to their query, or if it was returned due to poor query translation or inadequate information retrieval techniques. To allow the user at least some judgment about the returned stories, we attempted to label each Croatian news story with an English-language topic.

The topic identification was done using the Informedia translation facility to translate the whole story into English words. This translation became the topic query. Separately, we had a collection of about 35,000 English language news stories from Prime Source Media, which had topics manually assigned to them. The algorithm for the topic-labeling module is based on a k-nearest neighbor (KNN) strategy. Using the SMART information retrieval system, we now used the translated topic query to retrieve the ten most relevant labeled English stories. Each of the topics for the labeled stories that were retrieved was weighted by its relevance to the topic query and the weights for each topic were summed. The most favored topics, above a threshold, were then used to provide a topic label for the Croatian news story. This topic label allows the user to identify the general topic area of an otherwise incomprehensible foreign language text and determine if it is relevant at least in the topic area.

To test the accuracy of the topic assignment, 24 news stories broadcast in 1996 by Croatian Television in Zagreb were manually transcribed. A typical story, which was manually translated into English, is given in the following paragraph:

“Palestinian leader Yassir Arafat gave an announcement in Cairo, after his talk with the Egyptian President Mubarak and the main secret agent Magid of the Arab League, that in spite of the Israelis protests he intends to proclaim independent Palestinian state in the area of the West Coast and Gaze. Palestinian State is not Israel's problem, but the Arabian and international, says Arafat who believes all contracts up to date are only temporary, reports Reuter. The proclamation for the state should be the crowning of the peace talks. It is not necessary to point to the Israelis opposition regarding this matter”

For the above story, human transcribers marked the following topic labels as relevant:

“Administration; Arafat, Yasir; Israel; Jewish-Arab relations; Middle East; Middle East peace negotiations; Occupied territories; Palestine Liberation Organization; Palestinian Arabs; Palestinian self-rule areas.”

The same story translated using DIPLOMAT and after stopword removal and stemming was:

“palestinian leader yasser arafat declar talk league proclaim independ countri territori zapadne palestinian stat question arabic report theodor agreements probat countri not peac negotiations opposit point view expire”

The phrase-based translation system produced the following translation (again after stopword removal and stemming):
Recall and relevance of the automatically generated topics were measured for this test set of 24 news stories. Each story had an average of 9.6 manually assigned relevant topics. We compared those topic labels with the topics generated by the KNN method. In other words, of the topics that the KNN method generated, how many were the same as the ones assigned by a human (precision) and how many of the human assigned topics did we correctly assigned using the KNN method (recall). To measure our topic classification effectiveness, we use a widely accepted metric called the F-measure.

For this test set from Croatian television news, an evaluation which combined equal precision and recall yielded an F(1) measure value of 0.46 on manually translated Croatian television news stories. The machine translation experiment showed only a slight (8%) decrease in topic classification accuracy from 0.46 to 0.43, using our best available machine translation technology. Using phrase-based translation, the decrease was 31% to an F-measure value of 0.32. In all cases even the manual translation, the system performed worse than the English topic assignment on English language news stories reported in comparable experiments, which had an F-measure value at equal precision and recall of over 0.48.

In summary, we found the approach to be promising and these initial results to be encouraging. In particular, a high-quality machine translation provided topic assignment results comparable to perfect translations. The quick-and-dirty phrase translation system, however, showed noticeable degradation in the topic assignment.

**Relevant Informedia System Improvements**

**Map Interface**

Through automatic processing, descriptors are derived for the video to improve library access. A new extension to the video processing is the extraction of geographic references from these descriptors. The operational library interface shows the geographic entities addressed in a given story, highlighting the regions discussed at any point in the video through a map display synchronized with the video playback. The map can also be used as a query mechanism, allowing users to search the terabyte library for stories taking place in a selected area of interest.

Of course the real value to the geocoding and map interface is in displaying location information for video segments for which the producer has not previously added a map within the video data. Every news story does not have an embedded map that becomes part of the broadcast, but through our geocoding, maps can be automatically produced to reflect the areas discussed within each story. Another benefit is that the user can interact with the interface map using the toolbar icons to get additional detail, whereas no such interaction is possible with an image of a map encoded as part of the video stream.

The maps accompanying videos are not static displays. They are animated in synchronization with video playback. As places are discussed, they are highlighted on the map. For countries and administrative areas such as states or provinces, the areas contained within their respective polygon boundaries are highlighted by changing the color with which they are shaded. Areas covered at some time during a video segment are colored yellow; when the video frames during which an area is discussed are played, that area is then colored orange. For cities and other places, the marker color is
changed from blue to red and the label is shown. A glance at the map can then show the areas of current focus.

**Multimodal Search and Relevance Feedback**

We investigated multimodal queries, specifically, searching with text while also searching indicated geographic areas. A weighting scheme has been developed to combine results produced from these different modalities, with direct manipulation interfaces provided so that the user can adjust the weights to match specific needs. In addition, features have been extracted from the result set for use in relevance feedback, whereby the user informs the system which results are relevant to his needs. The features common to those results marked as relevant are used to reorder the result set accordingly. Experiments will be conducted in the next phase of Informedia research to measure the effectiveness of multimodal map and text queries and relevance feedback strategies, as compared to video information retrieval interfaces with text-only query, and text query with relevance feedback.

### 2.4 Results and Conclusions.

We built a dictation system which served as a baseline and prototype for the training of the actual broadcast news recognizer. We described the effects of different normalization and adaptation strategies, examined the effects of multi-corpora language model interpolation, and presented techniques to attack language specific problems such as rapid vocabulary growth and high OOV rate due to a large number of inflected word forms. The two distinct development lines for dictation and broadcast news were finally merged into one system for both tasks. Despite the fact that our system was based on a fairly low amount of acoustic training data, we yielded results that were comparable to the 1996 Hub-4 evaluation for English broadcast news: 29.5% WER for broadcast news, and 20.9% WER for dictation data.

As in all research, many challenges still remain. While this work was interesting and showed great promise, the lack of better machine translation became very apparent. At the very least, one would want to extract named entities from the foreign language data (People, Places, Organizations) and translate those into English. We should also attempt to generate English language summaries and titles, in addition to topic labels. The big remaining problem is the fact that retrieved results cannot be 'verified' unless the user speaks the target language.

### 2.5 Recommendations

There are some issues in the Multilingual Informedia client that would merit further work. One of these issues is to allow the combination of target language transcript and target language OCR, together with English language user annotations, titles and topics. Other issues include: How should a search proceed over video library data that has mixed language information? How should the architecture be modified to handle this? What about searches over larger collections of both English and foreign language video material? How should the search results from each language be combined?

In addition, non-language information (such as visual features, overlaid text faces, color, shapes, etc.) has added importance, as it is not subject to a language-specific interpretation. We need to further advance and better integrate visual feature extraction and image-based queries.
Section 3

3.1 Bibliography


Wactlar, H. New Directions in Video Information Extraction and Summarization, 10th DELOS Workshop, Santorini, Greece, June 24-25, 1999.


EXHIBIT A

Research Papers, attached.