MULTI-LINGUAL INFORMEDIA: A DEMONSTRATION OF SPEECH RECOGNITION AND INFORMATION RETRIEVAL ACROSS MULTIPLE LANGUAGES

A.G. Hauptmann, P. Scheytt, H.D. Wactlar, and P. E. Kennedy

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

ABSTRACT

The Multilingual Informedia Project demonstrates 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 new system performs speech recognition on foreign language news broadcasts, segments it into stories and indexes the foreign data together with existing English news data. This first multi-lingual prototype could easily be extended to other languages.

1. INTRODUCTION

The Informedia [5] project's goal is to allow search and retrieval in the video medium, similar to what is available today for text only. To enable this access to video, speech recognition is used to provide a text transcript for the audio track, image processing determines scene boundaries, recognizes faces and allows for image similarity comparisons. Everything is indexed into a searchable digital video library [4,6], where users can ask queries and receive relevant news stories as results.

2. THE NEW COMPONENTS OF MULTI-LINGUAL INFORMEDIA

There are four components in the Multi-lingual 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.

- Dynamic user annotations permit runtime addition of searchable user comments.

3. SERBO-CROATIAN SPEECH RECOGNITION: DICTATION AND BROADCAST NEWS

The Serbo-Croatian speech recognizer is trained on a weighted combination of 15 hours broadcast news data and 12 hours of read newspaper articles (audio parameters: 16 kHz/16 bit/mono) [1][3]. The quinphone-based context-dependent system consists of 4000 codebooks, which are modeled by left-to-right HMMs with 16 diagonal Gaussians. During preprocessing Mel-frequency cepstral coefficients are extracted every 10ms. The final feature vector is computed by a truncated LDA transformation of a concatenation of MFCCs, the short-term energy, their first and second order derivatives, and the zero crossing value.

Vocal tract length normalization and cepstral mean subtraction are used to extenuate speaker and channel differences. The pronunciation dictionary was created automatically by a simple grapheme-to-phoneme tool. The phone set consists of 30 phones, 4 noise and a silence model. The final system was developed on the basis of a recognizer for read speech.

We use segment-based MLLR adaptation and multi-corpora language model interpolation for testing. We retrieved 12 M words of Serbo-Croatian news texts from the Internet, filtered them, performed a diacritic conversion algorithm and normalized Serbian and Croatian dialectic variants to obtain a unique representation for every word. We further use the Hypothesis Driven Lexicon Adaptation (HDLA) to reduce the OOV rate due to rapid vocabulary growth [1][2]. The Serbo-Croatian speech recognizer finally yields a performance of 26.0% WER on broadcast news and 20.9% WER on dictation data. The next section will describe the translation component of the system.

4. THE INFORMEDIA QUERY-TRANSLATION FACILITY

The Informedia query-translation facility comes into play when the user wishes to search for a topic in a collection of broadcasts in a language he or she does not speak, and hence requires a capability for translating the query. The current version attempts to translate large chunks of phrases it finds in the query.

An earlier version looked up each word of the source-language query in a machine-readable bilingual dictionary and concatenated the results into an output string. The dictionary frequently gave multiple alternatives for the target-language equivalent of the source-language word. In such cases all the alternatives were included in the output string. In other cases a source-language word may not appear in the dictionary. Such words are simply omitted from the translated query.

The latest version takes advantage of multi-word phrase entries in the machine-readable dictionary. It parses the source-language query for phrases using a simple recursive algorithm. It first scans the dictionary for a translation(s) of the entire query as one phrase; if that fails it searches for phrasal translations of substrings one word smaller, then one more word smaller than that, and so on. The first phrasal translation thus obtained (or set thereof if there are multiple alternative translations) is kept as part of the output string, and the process is recursively invoked on the pieces of the query preceding and following the substring just translated. The recursion continues until a set of chunks and individual words is produced covering the query string, for which translations have been found for all the chunks and may or may not have been found for the individual words. The concatenated results become the output string.

The query-translation facility will work with any language pair so long as a bilingual machine-readable dictionary is available in the format the program understands. Our work so far has focused on English and Serbo-Croatian.

Future plans include stopword lists in English and Serbo-Croatian, stemming logic for Serbo-Croatian, compilation of test corpora in English and Serbo-Croatian, and a feature to route queries to one of the publicly available translation engines on the World-Wide Web. The latter feature will gain cross-lingual search capabilities over several languages at one stroke. The DIPLOMAT machine translation system developed here at Carnegie Mellon University will also get put to use for query translation among the language pairs it supports.

5. FOREIGN LANGUAGE TOPIC DETECTION

After initial experiments with the Serbo-Croatian news processed by the multi-lingual 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 Serbo-Croatian news story with an English language topic identification.

The topic identification was done using the query translation facility to translate the whole story into English words. This translation became the *topic quer*y. Separately, we had indexed over 50000 English language news stories, which had manually assigned topics assigned to them. Using the SMART information retrieval system, we now used the translated *topic query* to retrieve the most relevant 10 labeled English stories. Each of the topics for the retrieved labeled stories 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 Serbo-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.

6. DYNAMIC USER ANNOTATIONS

A final feature in the multi-lingual Informedia system is the ability for a user to add an annotation to a particular news story. This capability is intended to permit a user to write a note, comment, or perhaps even a full translation of the news story as an annotation. The annotation is immediately added to the searchable text index. No re-indexing of the full video library is necessary, all annotations are added incrementally. The annotation is then part of the searchable index for the complete digital video library. The purpose of this capability is to allow others then to retrieve the news story based their own or others comments, and to share the results of labor intensive work such as high quality manual translation.

7. SUMMARY

While the multi-lingual Informedia system is at this point only a concept demonstration, it illustrates the components that are necessary for a full cross-language broadcast news video library system. The system demonstrates the seamless extension of the Informedia approach to search and discovery across video documents in multiple languages. Through the use of a foreign language speech recognizer, a phrase-translation capability, an English topic label for foreign language news stories and incrementally added and searchable user annotations, the system points the way towards some of the capabilities one could expect in a fully-functional multi-lingual Informedia video library.

REFERENCES

1. Geutner, P., Finke, M., Scheytt, P., Waibel, A., and Wactlar, H., “Transcribing Multilingual Broadcast News Using Hypothesis Driven Lexicon Adaptation.” To appear in *BNTUW-98 Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne VA, February 1998.

2. Geutner, P., Finke, M., Scheytt, P., Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News, Technical Report, Carnegie Mellon University, Pittsburgh, PA, CMU-LTI-97-155, December 1997.

3. Scheytt, P., Finke, M., Geutner, P., Speech Recognition on Serbo-Croatian Dictation and Broadcast News Data, Technical Report, Carnegie Mellon University, Pittsburgh, PA, CMU-LTI-97-154, December 1997.

4. Witbrock, M.J., and Hauptmann, A.G., “Speech Recognition and Information Retrieval”, Proceedings of the 1997 DARPA Speech Recognition Workshop, Chantilly, VA, February 2-5, 1997.

5. Wactlar, H.D., Kanade, T., Smith, M.A. and Stevens, S.M. “Intelligent Access to Digital Video: Informedia Project”. *IEEE Computer*, **29**(5) May 1996, p.p. 46-52.

6. Hauptmann, A.G. and Witbrock, M.J., *Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval*, In Maybury, M. (ed.), "Intelligent Multimedia Information Retrieval", AAAI Press, 1997.