# Robust Knowledge Discovery from Parallel Speech and Text Sources

F. Jelinek, W. Byrne, S. Khudanpur, B. Hladká. CLSP, Johns Hopkins University, Baltimore, MD.
H. Ney, F. J. Och. RWTH Aachen University, Aachen, Germany
J. Cuřín. Charles University, Prague, Czech Rep.
J. Psutka. University of West Bohemia, Pilsen, Czech Rep.

## 1. INTRODUCTION

As a by-product of the recent information explosion, the same basic facts are often available from multiple sources such as the Internet, television, radio and newspapers. We present here a project currently in its early stages that aims to take advantage of the redundancies in parallel sources to achieve robustness in automatic knowledge extraction.

Consider, for instance, the following sampling of actual news from various sources on a particular day:

**CNN:** James McDougal, President Bill Clinton's former business partner in Arkansas and a cooperating witness in the Whitewater investigation, died Sunday while serving a federal prison term. He was 57.

**MSNBC:** Fort Worth, Texas, March 8. Whitewater figure James McDougal died of an apparent heart attack in a private community hospital in Fort Worth, Texas, Sunday. He was 57.

**ABC News:** Washington, March 8. James McDougal, a key figure in Independent Counsel Kenneth Starr's Whitewater investigation, is dead.

**The Detroit News:** Fort Worth. James McDougal, a key witness in Kenneth Starr's Whitewater investigation of President Clinton and First Lady Hillary Rodham Clinton, died of a heart attack in a prison hospital Sunday. He was 57.

**San Jose Mercury News:** James McDougal, the wily Arkansas banking rogue who drew Bill Clinton and Hillary Rodham Clinton into real estate deals that have come to haunt them, died Sunday of cardiac arrest just months before he hoped to be released from prison. He was 57.

**The Miami Herald:** Washington. James McDougal, the wily Arkansas financier and land speculator at the center of the original Whitewater probe against President Clinton, died Sunday.

.

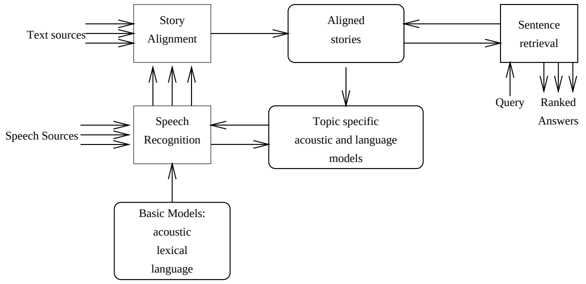

**Figure 1: Information Flow in Alignment and Extraction**

We propose to align collections of stories, much like the example above, from multiple text and speech sources and then develop methods that exploit the resulting parallelism both as a tool to improve recognition accuracy and to enable the development of systems that can reliably extract information from parallel sources.

Our goal is to develop systems that align text sources and recognize parallel speech streams simultaneously in several languages by making use of all related text and speech. The initial systems we intend to develop will process each language independently. However, our ultimate and most ambitious objective is to align text sources and recognize speech using a single, integrated multilingual ASR system. Of course, if sufficiently accurate automatic machine translation (MT) techniques ([1]) were available, we could address multilingual processing and single language systems in the same way. However MT techniques are not yet reliable enough that we expect all words and phrases recognized *within* languages to contribute to recognition *across* languages. We intend to develop methods that identify the particular words and phrases that both can be translated reliably and also used to improve story recognition.

As MT technology improves it can be incorporated more extensively within the processing paradigm we propose. We consider this proposal a framework within which successful MT techniques can eventually be used for multilingual acoustic processing.

## 2. PROJECT OBJECTIVES

The first objective is to enhance multi-lingual information systems by exploiting the processing capabilities for resource-rich languages to enhance the capabilities for resource-impoverished language. The second objective is to advance information retrieval and knowledge information systems by providing them with considerably improved multi-lingual speech recognition capabilities. Our research plan proceeds in several steps to (i) collect and (ii) align multi-lingual parallel speech and text sources, (iii) exploit parallelism for improving ASR within a language, and to (iv) exploit parallelism for improving ASR across languages. The main information flows involved in aligning and exploiting parallel sources are illustrated in Figure 1. We will initially focus on German, English and Czech language sources. This section summarizes the major components of our project.

## 2.1 Parallel Speech and Text Sources

The monolingual speech and text collections that we will use to develop techniques to exploit parallelism for improving ASR within a language are readily available. For instance, the North American News Text corpus of parallel news streams from 16 US newspapers and newswire is available from LDC. A 3-year period yields over 350 million words of multi-source news text.

In addition to data developed within the TIDES and other HLT programs, we are in the process of identifying and creating our own multilingual parallel speech and text sources.

*FBIS TIDES Multilingual Newstext Collection*
For the purposes of developing multilingual alignment techniques, we intend to use the 240 day, contemporaneous, multilingual news text collection made available for use to TIDES projects by FBIS. This corpus contains news in our initial target languages of English, German, and Czech. The collections are highly parallel, in that much of the stories are direct translations.

*Radio Prague Multilingual Speech and Text Corpus*
Speech and news text from Radio Prague was collected under the direction of J. Psutka with the consent of Radio Prague. The collection contains speech and text in 5 languages: Czech, English, German, French, and Spanish. The collection began June 1, 2000 and continued for approximately 3 months. The text collection contains the news scripts used for the broadcast; the broadcasts more or less follow the scripts. The speech is about 3 minutes per day in each language, which should yield a total of about 5 hours of speech per language.

Our initial analysis of the Radio Prague corpus suggest that only approximately 5% of the stories coincide in topic, and that there is little, if any, direct translation of stories. We anticipate that this sparseness will make this corpus significantly hard to analyze than another, highly-parallel corpus. However, we expect this is the sort of difficulty that will likely be encountered in processing ‘real-world’ multilingual news sources.

## 2.2 Story-level Alignment

Once we have the multiple streams of information we must be able to align them according to story. A story is the description of one or more events that happened in a single day and that are reported in a single article by a daily news source the next day. We expect that we will use the same techniques used in the Topic Detection (TDT) field ([5]). Independently of the specific details of the alignment procedure, there is now substantial evidence that related stories from parallel streams can be identified using standard statistical Information Retrieval (IR) techniques.

**Sentence Alignment** As part of the infrastructure needed to incorporate cross-lingual information into language models, we are employing statistical MT systems to generate English/German and English/Czech alignments of sentences in the FBIS Newstext Collection. For the English/German sentence and single-word based alignments, we plan to use statistical models ([4]) [3] which generate both sentence and word alignments. For English/Czech sentence alignment, we will employ the statistical models trained as part of the Czech-English MT system developed during the 1999 Johns Hopkins Summer Workshop ([2]).

## 2.3 Multi-Source Automatic Speech Recognition

The scenario we propose is extraction of information from parallel text followed by repeated recognition of parallel broadcasts, resulting in a gradual lowering the WER. The first pass is performed in order to find the likely topics discussed in the story and to identify the topics relevant to the query. In this process, the acoustic model will be improved by deriving pronunciation specifications for out-of-vocabulary words and fixed phrases extracted from the parallel stories. The language model will be improved by extending the coverage of the underlying word and phrase vocabulary, and by specializing the model’s statistics to the narrow topic at hand. As long as a round of recognition yields new information, the corresponding improvement is incorporated into the recognizer modules and bootstrapping of the system continues.

*Story-specific Language Models from Parallel Speech and Text*
Our goal is to create language models combining specific but sparse statistics, derived from relevant parallel material, with reliable but unspecific statistics obtainable from large general corpora. We will create special *n-gram* language models from the available text, related or parallel to the spoken stories. We can then interpolate this special model with a larger pre-existing model, possibly derived from training text associated to the topic of the story. Our recent STIMULATE work demonstrated success in construction of topic-specific language models on the basis of hierarchically topic-organized corpora [8].

Unlike building models from parallel texts, the training of story specific language models from recognized speech is also affected by recognition errors in the data which will be used for language modeling. Confidence measures can be used to estimate the correctness of individual words or phrases on the recognizer output. Using this information, *n-gram* statistics can be extracted from the recognizer output by selecting those events which are likely to be correct and which can therefore be used to adjust the original language model without introducing new errors to the recognition system.

*Language Models with Cross-Lingual Lexical Triggers*
A trigger language model ([6], [7]) will be constructed for the target language from the text corpus, where the lexical triggers are not from the word-history in the target language, but from the aligned recognized stories in the source language. The trigger information becomes most important in those cases in which the baseline *n-gram* model in the target language does not supply sufficient information to predict a word. We expect that content words in the source language are good predictors for content words in the target language and that these words are difficult to predict using the target language alone, and the mutual information techniques used to identify trigger pairs will be useful here.

Once a spoken source-language story has been recognized, the words found here there will be used as triggers in the language model for the recognition of the target-language news broadcasts.

# 3. SUMMARY

Our goal is to align collections of stories from multiple text and speech sources in more than one language and then develop methods that exploit the resulting parallelism both as a tool to improve recognition accuracy and to enable the development of systems that can reliably extract information from parallel sources. Much like a teacher rephrases a concept in a variety of ways to help a class understand it, the multiple sources, we expect, will increase the potential of success in knowledge extraction. We envision techniques that will operate repeatedly on multilingual sources by incorporating newly discovered information in one language into the models used for all the other languages. Applications of these methods extend beyond news sources to other multiple-source domains such as office email and voice-mail, or classroom materials such as lectures, notes and texts.

## 4. REFERENCES

[1] P. F. Brown, S. A. DellaPietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical translation. *Computational Linguistics*, 19(2), 1993.

[2] K. K. et al. Statistical machine translation, WS'99 Final Report, Johns Hopkins University, 1999. `http://www.clsp.jhu.edu/ws99/projects/mt`.

[3] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL'00*, pages 440–447, 2000.

[4] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In *EMNLP/VLC'99*, pages 20–28, 1999.

[5] Proceedings of the Topic Detection and Tracking workshop. University of Maryland, College Park, MD, October 1997.

[6] C. Tillmann and H. Ney. Selection criteria for word trigger pairs in language modelling. In *ICGI'96*, pages 95–106, 1996.

[7] C. Tillmann and H. Ney. Statistical language modeling and word triggers. In *SPECOM'96*, pages 22–27, 1996.

[8] D. Yarowsky. Exploiting nonlocal and syntactic word relationships in language models for conversational speech recognition, a NSF STIMULATE Project IRI9618874, 1997. Johns Hopkins University.