Shared Resources for Robust Speech-to-Text Technology

Stephanie Strassel, David Miller, Kevin Walker, Christopher Cieri

Linguistic Data Consortium
University of Pennsylvania
strassel|damiller|walkerk|ccieri@ldc.upenn.edu

Abstract

This paper describes ongoing efforts at Linguistic Data Consortium to create shared resources for improved speech-to-text technology. Under the DARPA EARS program, technology providers are charged with creating STT systems whose outputs are substantially richer and much more accurate than is currently possible. These aggressive program goals motivate new approaches to corpus creation and distribution. EARS participants require multilingual broadcast and telephone speech data, transcripts and annotations at a much higher volume than for any previous program. While standard approaches to resource collection and creation are prohibitively expensive for this volume of material, within EARS new methods have been established to allow for the development of vast quantities of audio, transcripts and annotations. New distribution methods also provide for efficient deployment of needed resources to participating research sites as well as enabling eventual publication to a wider community of language researchers.

1. Introduction

The DARPA EARS Program (Effective, Affordable, Reusable Speech-to-Text) is focused on enabling core speech-to-text technology to produce rich, highly accurate output in range of languages and speaking styles. Aggressive program goals target substantial improvements on current technology. Initially, the focus languages are English, Chinese and Arabic, with expansions possible in future years. All human language technology demands large quantities of data for system training and development, plus stable benchmark data to measure ongoing progress. This presents a substantial challenge to the HLT community, because human annotation and corpus creation is costly. Within EARS, researchers require not tens but hundreds and thousands of hours of speech data plus corresponding transcripts and other types of annotation. The availability of high quality language resources is a critical issue for not only the EARS program but for HLT research in general.

The Linguistic Data Consortium (LDC) was founded in 1992 at the University of Pennsylvania, with seed money from DARPA, specifically to address the need for shared language resources. Since then, the LDC has created and published more than 225 linguistic databases, and has accumulated considerable experience and skill in managing large-scale, multilingual data collection and annotation projects. Responding to the need for more data in a wider variety of languages with more sophisticated annotation, LDC has established itself as a center for research into standards and best practices in linguistic resource development, while participating actively in ongoing HLT research. Within the context of EARS, LDC provides conversational and broadcast audio and transcripts, lexicons and texts for language modeling, and other types of complex annotation in all of the target languages.

2. Data Requirements

The EARS program will support several common task evaluations. Administered by the National Institute of Standards and Technology (NIST) under the Rich Transcription Evaluation heading, the specific research tasks are broadly categorized as supporting either Speech-to-Text (STT) or Metadata Extraction (MDE). Within 2003, STT tasks cover English, Mandarin and Arabic for both broadcast news and telephone speech data. Metadata evaluations are limited to English for 2003.

Data collection is a serious concern for EARS. The program goals mean that research sites require an order of magnitude more data than in the past. LDC has responded to this challenge with targeted broadcast news and telephone speech collections in all three EARS languages. A customized, locally developed broadcast news collection platform has expanded LDC's ability to capture broadcast data from a wide range of sources in a multitude of languages. System capacity allows for collection via an array of satellite dishes, cable television, web audio and shortwave and broadband antennae, all controlled through LDC's in-house system. Automatic processes control digitization of audio, removal of video signal where appropriate, closed caption download and creation of automatic speech recognition output in English, Chinese and Arabic. A 10 terabyte "Wall of Disk" provides for ongoing storage. Under this configuration, LDC currently collects over 25 hours of audio per day in English, plus tens of hours in Chinese and Arabic from at least 14 sources. System capacity allows for substantially higher collection targets.

Additionally, a new telephone speech collection platform, named the Fisher protocol, has been designed and implemented to support the goals of EARS. Within Fisher, the collection platform initiates calls to participants, pairing them with other subjects who have indicated their willingness to participate at the designated time. The platform can record multiple simultaneous conversations without operator intervention, and a single project database tracks participant information and call activity. Both the telephone and broadcast news collection platforms rely on off-the-shelf hardware to provide robust but portable solutions.

The data collection goals for EARS are ambitious. Within 2003 alone, LDC is targeting 10,000 hours of English, 3000 hours of Arabic and 1000 hours of Chinese broadcast audio. The telephone collection goals are 2000 hours of English

audio, plus smaller collections in Arabic and Chinese. This broadcast and telephone data is transcribed and annotated in a number of ways to provide training, development and evaluation data to support the full range of EARS research tasks.

3. Speech-to-Text

The Speech-to-Text task is the core EARS research task. The fundamental program goal is a substantial improvement in STT system performance, measured in terms of overall word error rate. In addition to requiring thousands of hours of audio data in support of this goal, sites also need corresponding transcripts in order to develop language models and provide for system training. Benchmark data is also needed to allow sites and program sponsors to measure performance on a stable test set. LDC is providing these annotated corpora in a number of ways.

3.1. Careful transcription

For purposes of evaluating STT technology, system output must be compared with high-quality manually-created verbatim transcripts. The cost of creating such careful transcripts is quite high. Transcription rates approach forty to fifty times real time, so that it requires forty or more hours of annotator effort to carefully transcribe one hour of speech. Additional time goes into project management as well developing data formats and customized annotation tools to facilitate the transcription process. The careful transcription effort involves multiple passes over the data. Annotators first manually segment speaker turns and (for broadcast data) story boundaries, as well as indicating smaller breakpoints within the audio stream that correspond to breath or pause groups. After accurate segment boundaries are in place, annotators create a verbatim transcript by listening to each segment in turn. A second pass checks the accuracy of the segment boundaries and transcript itself, revisits difficult sections, and adds information like speaker identity, background noise conditions, plus special markup for mispronounced words, proper names, acronyms, partial words and the like. Further scans over the data identify common errors, conduct spelling and syntax checks, and standardize the spelling of personal, organization and other names across the transcripts. LDC also provides forced alignment output with the final transcripts, relying on a locally-developed FA system that creates word-based alignment for each word within the transcript.

In support of the 2003 Speech-to-Text evaluation, LDC supplied nine hours of English data (three hours broadcast audio and transcripts, three hours Fisher-style telephone data, and three hours Switchboard-style data). For Chinese, data consisted of one hour of broadcast news and one hour of Mandarin CallFriend; Arabic data was one hour of broadcast news and one hour of Egyptian Arabic CallHome data.

In addition to data created for a specific evaluation (called the Current Data Set) the EARS program also incorporates a Progress Data Set. While the content of the Current Data Set will change from year to year within the program, the Progress Data Set will remain stable for the duration of the EARS program, providing a yardstick for measuring improvement in system performance over time and allowing new EARS participants to quickly compare their performance against existing technologies using stable benchmark data. The English-only Progress Data Set developed by LDC consists of three hours of Fisher speech and transcripts plus three hours of broadcast news data.

3.2. Quick Transcription

The cost of producing careful transcripts of the type described above for the large quantities of training material required by the EARS program is prohibitively expensive. In order to achieve the aggressive program goals, in particular the significant reduction of word error rate, technology developers require thousands of hours of transcribed training data. Realizing that community needs would far outstrip available resources within the existing framework, LDC and other members of the EARS community planned a Quick Transcription experiment whose purpose was to pare down transcription rates while retaining the level of quality required for system training and statistical modeling. A pilot Quick Transcription experiment in late 2002 produced transcripts for 185 Switchboard calls; feedback from the EARS research community indicates that the quality of the resulting transcripts is sufficiently high to allow for system training and development.

The approach taken during Quick Transcription is to limit the amount of time annotators may spend with a given speech file. Transcription rates are targeted at five times real time. Many of the extra features of careful transcription are removed so that annotators can focus on creating verbatim transcripts within the time constraints. Rather than manually segmenting speaker turns, an automatic process developed at LDC pre-segments a telephone call into high-accuracy turn boundaries. Annotators do not use punctuation, capitalization or most of the special markup adopted for the careful transcription task. Rather than executing three to four separate passes over the data, annotators complete the (close-to) verbatim transcript within one transcription pass. Automatic post-processing targets spell checking, syntax checking and scans for common errors. Specialized transcription tools allow the annotator to quickly move from turn to turn within the transcript; new tools are being developed that will automate certain procedures, removing the need for repetitive keystrokes and allowing the annotator to speed up audio playback. Team leaders monitor annotator progress and speed to ensure that transcripts are produced within the targeted timeframe.

The resulting quick transcription quality is naturally lower than that produced by the careful transcription methodology. Speeding up the process inevitably results in missed or mis-transcribed speech; this is particularly true for difficult sections of the transcript, including disfluent or overlapping speech sections. However, the advantage of this approach is undeniable. Annotators work, on average, ten times faster using this approach than they are able to work within the careful transcription methodology. LDC project managers continue to work with other members of the EARS community to develop new quality assurance measures, and to research how LDC annotators might utilize the best existing STT technology to improve both efficiency and quality in the Quick Transcription process.

To support ongoing STT research, LDC annotators will transcribe at least 300 hours of Fisher telephone speech using the Quick Transcription method in calendar year 2003; transcription of another 1700 hours will be managed by BBN. Smaller efforts are targeted to create Quick Transcripts for English broadcast data as well as Chinese and Arabic telephone speech.

4. Metadata

The goal of the metadata extraction evaluation is to enable technology that can take the raw STT output and refine it into forms that are of more use to humans and to downstream automatic processes. In simple terms, this means the creation of automatic transcripts that are maximally readable. This readability might be achieved in a number of ways: removing non-content words like filled pauses and discourse markers from the text; removing sections of disfluent speech; and creating boundaries between natural breakpoints in the flow of speech so that each sentence or other meaningful unit of speech might be presented on a separate line within the resulting transcript. Natural capitalization, punctuation and standardized spelling, plus sensible conventions for representing speaker turns and identity are further elements in the readable transcript.

To support these goals, LDC has defined a MDE annotation task to create both training and test data. Working with a careful, verbatim transcript (e.g., reference data created for the STT evaluation), annotators identify a range of metadata phenomena that affect the representation of the rendered transcript. Metadata phenomena include four types of fillers: filled pauses like "uh" and "um", discourse markers like "you know", asides and parentheticals, and editing terms like "sorry" and "I mean". The second metadata feature is edit disfluencies, which are portions of speech in which a speaker's utterance is not complete and fluent; instead the speaker corrects or alters the utterance, or abandons it entirely and starts over. Both fillers and edit disfluencies are removed from the rendered transcript; their removal does not affect the content or flow of the discourse.

Annotators further identify SUs (alternately semantic units, sense units, syntactic units, slash units or sentence units); that is, units within the discourse that function to express a complete thought or idea on the part of a speaker. As with disfluency annotation, the goal of SU labeling is to improve transcript readability, here by creating a transcript in which information is presented in small, structured, coherent chunks rather than long turns or stories. There are four types of sentence-level SUs: statements, questions, backchannels and incomplete SUs. To enhance inter-annotator consistency, the annotation task also identifies a number of sub-sentence SU boundaries (coordination and clausal SUs).

An example of a readable transcript created by such annotation follows:

Table 1: Standard STT vs. Rendered Text Output

Original STT Output	Rendered Text Output
um but the job that i ju- i had this job that i just lost you know it wasn't like it wasn't the best job i've ever had but it sti- it like it paid the bills	I had this job that I just lost. It wasn't the best job I've ever had, but it paid the bills.

Data annotated for fillers, disfluencies and SUs under the EARS Program includes hundreds of hours of English conversational telephone speech, a smaller amount of English broadcast news, plus tens of hours of pilot data in Mandarin Chinese and Egyptian Arabic. A major challenge of the metadata task has been creating annotation guidelines that allow for a team of non-expert annotators to achieve high levels of inter-annotator consistency while maintaining maximal efficiency. Customized annotation tools and rigorous quality assurance measures including ongoing annotator training are core features of the task.

5. Data distribution and publication

Much of the material described above is based upon large volumes of text and speech best collected from commercial providers. Commercial sources may require the negotiation of agreements that permit the distribution of data to researchers while constraining the use of the material to linguistic education, research, and technology development. LDC coordinates all necessary intellectual property arrangements for multiple research programs, including EARS, to make resources gathered in this way available to the broader research communities.

In order to allow for expedited delivery of data to a limited number of research sites participating in the EARS common task evaluations, LDC has developed a new data distribution method known as ECorpora. ECorpora target expedited delivery of training and development data in support of formal technology evaluations. The training material described above has been or will be distributed to EARS program participants using this ECorpora method. Upon the conclusion of the formal task evaluation, pending negotiations with research sponsors and program coordinators, LDC publishes data more broadly to permit access to these valuable resources to all communities working in linguistic education, research, and technology development.

6. Conclusions

Shared resources are a critical component of human language technology development. New research programs like EARS require updated approaches to data collection, annotation and distribution to support ambitious goals. LDC is engaged in ongoing efforts to provide crucial resources for improved speech-to-text technology to program participants as well as a larger community of language researchers, educators and technology developers.

7. References

[1] Cieri, Christopher, David Miller and Kevin Walker, "From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields", 2003.

[2] EARS, DARPA Program in Effective, Affordable, Reusable Speech-to-Text, *http://www.darpa.mil/iao/EARS.htm*, 2002.

[3] Linguistic Data Consortium, EARS Project Page, *http://www.ldc.upenn.edu/Projects/EARS*, 2003.

[4] Linguistic Data Consortium, ECorpora Policy, *http://www.ldc.upenn.edu/Membership/ECorpora*, 2003.

[5] National Institute for Standards and Technology, NIST RT-03 Spoken Language Technology Evaluation, *http://www.nist.gov/speech/tests/rt/rt2003*, 2003.

[6] Miller, David, Linguistic Data Consortium Quick Transcription Specification, 2003.

[7] Strassel, Stephanie, Linguistic Data Consortium Simple Metadata Annotation Specification, *http://www.ldc.upenn.edu/Projects/MDE/SimpleMDE/Guidelines/SimpleMDE_V4.0.pdf*, 2003.

[8] Strassel, Stephanie, Linguistic Data Consortium RT-03 Transcription Specification, *http://www.ldc.upenn.edu/Projects/Transcription/rt-03/Guidelines/RT-03V2.3.pdf*, 2003.