Linguistic Resources for Effective, Affordable, Reusable Speech-to-Text

Stephanie Strassel

Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
strassel@ldc.upenn.edu

Abstract

This paper describes ongoing efforts at Linguistic Data Consortium to create shared evaluation resources for improved speech-to-text technology. The DARPA EARS Program (Effective, Affordable, Reusable Speech-to-Text) is focused on enabling core STT technology to produce rich, highly accurate output in a range of languages and speaking styles. The aggressive EARS program goals motivate new approaches to corpus creation and distribution. EARS research sites require multilingual broadcast news and telephone speech, transcripts and annotations at a much higher volume than for any previous technology program. In response to these demands, LDC has developed new corpora for training and evaluating speech-to-text systems in English, Arabic and Chinese and to support systems that distinguish speakers, identify and repair disfluencies and punctuate a text to improve readability.

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 to current technology. Initially, the focus languages are English, Chinese and Arabic, with expansions possible in future years. Within EARS, researchers require not tens but hundreds and thousands of hours of speech data plus corresponding manual 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 human language technology 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, LDC has created and published more than 288 linguistic databases, and has accumulated considerable experience 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.

Data requirements and collections

The EARS program supports several common task evaluations. Administered by the National Institute of Standards and Technology (NIST) under the Rich Transcription Evaluation heading (NIST 2004), the specific research tasks are broadly categorized as supporting either Speech-to-Text (STT) or Metadata Extraction (MDE). While STT emphasizes getting the words right, MDE is concerned with structuring STT output to be maximally readable for humans and downstream automatic processes. In 2003 and 2004, STT tasks cover broadcast news and telephone speech in English, Mandarin and Arabic. Metadata evaluations are currently limited to English, though pilot explorations have begun in Chinese and Arabic.

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 allow for 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.

Additionally, a new telephone speech collection platform, named the Fisher protocol (Cieri et.al. 2003), 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. Fisher-style platforms are currently operating in English and Arabic, with Chinese collection slated to begin in early 2004.

These broadcast and telephone collections are transcribed and annotated to support the full range of EARS research tasks. In 2003 and through the first quarter of 2004, LDC produced thousands of hours of training data, plus tens of hours of development and evaluation data for EARS for each language and evaluation area. In addition to the evaluation 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 changes from year to year, 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. Researchers also require development data to evaluate system performance in advance of the official common task evaluations. Typically the previous year's evaluation set serves as DevTest data, but in some cases this is supplemented with additional data created explicitly for system development. Finally, a small amount of

benchmark data is dually transcribed and annotated to establish inter-annotator agreement rates.

Speech-To-Text

Speech-to-Text 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.

Careful transcription of benchmark data

For purposes of evaluating STT technology, system output must be compared with high-quality 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 human effort to carefully transcribe one hour of speech. The careful transcription effort involves multiple passes over the data (LDC 2004b). 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. Senior transcribers then conduct quality checks on the data to ensure the completeness and accuracy of the transcripts and consistent application of the markup. Further automatic and manual scans over the data identify common errors, conduct spelling and syntax checks, standardize the spelling of personal, organization and other names across the transcripts, and validate the data format.

Quick transcription of training data

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 (QTR) 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.

The approach taken during QTR is to limit the amount of time annotators may spend with a given speech file (LDC 2003). 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. Post-processing handles spell checking, syntax checking and automatic 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, five to 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 QTR process. QTR is the default transcription process for English telephone speech training data, and is currently being explored for both Arabic and Chinese data.

Metadata

The goal of the EARS 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 can 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 defined a MDE annotation task to create both training and test data (LDC 2004c). Working with a careful, verbatim transcript (e.g., reference data created for STT), 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, where a speaker corrects or alters his original utterance, or abandons it entirely and starts over. Both fillers and edit disfluencies can be 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, 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:

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	*Speaker A*: I had this job that I just lost. It wasn't the best job I've ever had, but it paid the bills.

Figure 1: Standard STT vs. Rendered Text Output

LDC also provides forced alignment output along with the MDE annotation, relying on a locally-developed FA system that creates word-based alignment for each word within the transcript.

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. In 2003 alone, the annotation task definition underwent over 15 major revisions to accommodate the MDE program's evolving research goals. In early 2004 another set of major changes was adopted to incorporate lessons learned during the first year of MDE, and to address concerns about inter-annotator consistency.

Because the first year of MDE research saw an ever-evolving task definition and a compressed timeline for data production, having task-specific, highly customized and easily modifiable annotation tools was essential. LDC developed a MDE annotation toolkit using the Annotation Graphs model (Maeda and Strassel 2004). The basic annotation tool is displayed below:

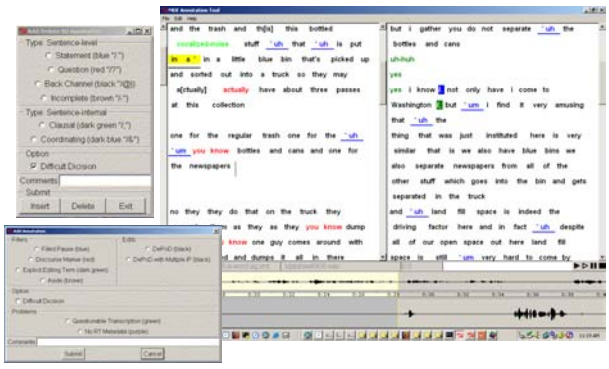

Figure 2: MDE Annotation Tool

The tool features separate modes for filler and edit disfluency tagging, SU tagging, and second passing as well as a rendered text output function. The MDE toolkit also includes an adjudication tool that allows the user to review the output of dually annotated data, compare results and adjudicate differences. The tool takes two annotation files for the same source data, highlights the differences and asks for the adjudicator's judgment on each discrepancy. Approximately 15% of all training data, as well as most of the development and evaluation data created to support MDE is dually annotated by an independent annotator, and files are compared to establish a baseline inter-annotator agreement rate. The files are then adjudicated to produce gold standard data for system evaluation and ongoing annotator training.

Data distribution and publication

Whereas normal LDC publications require several weeks or months to produce, data for common task evaluations must be distributed to researchers very soon after it has been collected or annotated, given the fixed evaluation timeline. In order to allow for expedited delivery of data to a limited number of research sites participating in formal evaluations, LDC has developed a new data distribution method known as ECorpora (where "E" stands for experimental). ECorpora do not involve the same level of rigorous data validation and documentation that a general release publication demands, but still require tracking of user agreements and data recipients as well as attention to intellectual property rights (IPR) issues. ECorpora utilize LDC's regular publications mechanisms thus allowing for the necessary tracking, but are produced using a "fast-track" approach that enables a turnaround time of several days rather than weeks or months. The ECorpus production process is illustrated in Figure 3 below.

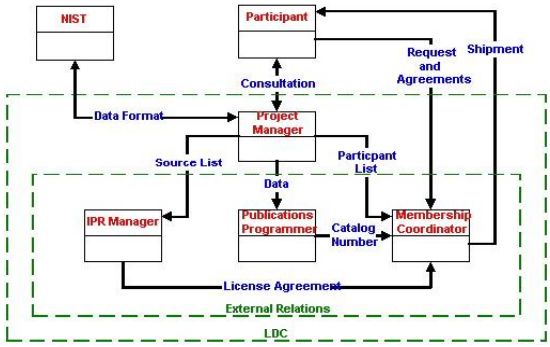

Figure 3: E-Corpus Production and Distribution Process

Linguistic data developed for EARS takes full advantage of the ECorpus distribution method. In 2003, 19 ECorpora were produced by LDC for EARS, while 22 are slated for release in the first half of 2004.

Much of the material developed for the EARS Program is based upon large volumes of text and speech collected from commercial data providers. Commercial sources typically 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.

The volume of broadcast news training data (over 10,000 hours in 2003 alone) makes licensing and re-distribution of this data prohibitively expensive. Instead, this data is made available to EARS program participants using a lending library approach, in which the data is shipped to sites on drive arrays

and returned to LDC at the conclusion of the evaluation. Wherever possible, however, data is distributed more broadly. Upon the conclusion of the formal task evaluation, pending negotiations with research sponsors and program coordinators, LDC publishes data as part of its regular catalog to provide access to these valuable resources to all communities working in linguistic education, research, and technology development.

Conclusions

The table below summarizes benchmark and training data distributed to EARS sites in 2003 or scheduled for production and distribution in 2004 (LDC 2004a).

Language	Test Set	Data Type	Data for 2003-2004 (hours)		
			Evaluation	DevTest	Training
English	Progress	telephone	3		
		broadcast	3		
English	STT	telephone	9	3	1920
		broadcast	6	6	7860
Chinese	STT	telephone	2	2	200
		broadcast	2	0.5	900
Arabic	STT	telephone	2	2	120
		broadcast	2	0	1600
English	MDE	telephone	4	4	86.5
		broadcast	4	4	40.25

Figure 4: Data Resources for EARS

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.

References

Cieri, Christopher, David Miller and Kevin Walker (2003). From Switchboard to Fisher: Telephone Collection Protocols, their Uses and Yields. In Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland, September 2003.

DARPA (2004). EARS Program (Effective, Affordable, Reusable Speech-to-Text) Website. [http://www.darpa.mil/ipto/Programs/ears/index.htm]

Linguistic Data Consortium (2003). Quick Transcription Guidelines. [http://www.ldc.upenn.edu/Projects/Transcription/quick-trans/index.html]

Linguistic Data Consortium (2004a). EARS Project Page. [http://www.ldc.upenn.edu/Projects/EARS]

Linguistic Data Consortium (2003b). RT-03 Careful Transcription Specification. [http://www.ldc.upenn.edu/Projects/Transcription/rt-03/RT_Transcription_V2.2.pdf]

Linguistic Data Consortium (2004c). Simple Metadata Annotation Specification V6.0. [http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.0.pdf]

Maeda, Kazuaki and Stephanie Strassel (2004). Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium. To appear in Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, May 2004.

National Institute for Standards and Technology (2004). NIST Rich Transcription Language Technology Evaluation. [http://nist.gov/speech/tests/rt/index.html]