Linguistic Resources for Meeting Speech Recognition

Meghan Lammie Glenn and Stephanie Strassel

Linguistic Data Consortium, University of Pennsylvania,
3600 Market Street, Suite 800, Philadelphia, PA 19104 USA
{mlglenn, strassel}@ldc.upenn.edu
http://www.ldc.upenn.edu

Abstract. This paper describes efforts by the University of Pennsylvania's Linguistic Data Consortium to create and distribute shared linguistic resources – including data, annotations, tools and infrastructure – to support the Rich Transcription 2005 Spring Meeting Recognition Evaluation. In addition to distributing large volumes of training data, LDC produced reference transcripts for the RT-05S conference room evaluation corpus, which represents a variety of subjects, scenarios and recording conditions. Careful verbatim reference transcripts including rich markup were created for all two hours of data. One hour was also selected for a contrastive study using a quick transcription methodology. We review the two methodologies and discuss qualitative differences in the resulting transcripts. Finally, we describe infrastructure development including transcription tools to support our efforts.

1 Introduction

Linguistic Data Consortium was established in 1992 at the University of Pennsylvania to support language-related education, research and technology development by creating and sharing linguistic resources, including data, tools and standards. Human language technology development in particular requires large volumes of annotated data for building language models, training systems and evaluating system performance against a human-generated gold standard. LDC has directly supported NIST's Rich Transcription evaluation series by providing both training and evaluation data and related infrastructure. For the Rich Transcription 2005 Spring Meeting Recognition Evaluation, LDC provided large quantities of training data from a variety of domains to program participants. Additionally, LDC produced both quick and careful reference transcripts of evaluation data to support automatic speech-to-text transcription, diarization, and speaker segmentation and localization in the meeting domain. Finally, in the context of this program LDC has undertaken creation of specialized annotation software that supports rapid, high-quality creation of rich transcripts, both in the meeting domain and in a wide variety of other genres.

2 Data

2.1 Training Data

To enhance availability of high-quality training data for RT-05S, LDC distributed twelve corpora that are part of the LDC catalog for use as training data by evaluation

S. Renals and S. Bengio (Eds.): MLMI 2005, LNCS 3869, pp. 390 – 401, 2006.

participants. The data included not only three corpora in the meeting domain, but also two large corpora of transcribed conversational telephone speech (CTS) as well as one corpus of transcribed broadcast news (BN). All data was shipped directly to registered evaluation participants upon request, after sites had signed a user agreement specifying research use of the data. The distributed training data is summarized in the table below.

RT-05S Training Data Distributed by LDC

Title	**Speech**	**Transcripts**	**Volume**	**Domain**
Fisher English Training Part 1	LDC2004S13	LDC2004T19	750+ hours	CTS
Fisher English Training Part 2	LDC2005S13	LDC2005T19	750+ hours	CTS
ICSI Meeting Corpus	LDC2004S02	LDC2004T04	72 hours	Meeting
ISL Meeting Corpus	LDC2004S05	LDC2004T10	10 hours	Meeting
NIST Meeting Pilot Corpus	LDC2004S09	LDC2004T13	13 hours	Meeting
TDT4 Multilingual Corpus	LDC2005S11	LDC2005T16	300+ hours	BN

2.2 Evaluation Data

In addition to training data, LDC developed a portion of the benchmark test data for this year's evaluation. The RT-05S conference room evaluation corpus includes ten meeting sessions contributed by five organizations or consortia: AMI (Augmented Multi-Party Interaction Project), CMU (Carnegie Mellon Institute), ICSI (International Computer Science Institute), NIST (National Institute of Standards and Technology), and VT (Virginia Tech). The sessions contain an average of six participants. In all but one case, head-mounted microphone recordings were available; the one exception is a speaker participating in the recording session by teleconference. The meetings represent a variety of subjects, scenarios and recording conditions. The RT-04 meeting evaluation corpus, also transcribed by LDC, covered a broader set of meeting activities, including simulated meetings and game playing (for instance a game of Monopoly or role playing games). The RT-05S conference room corpus on the other hand contains more typical business meeting content [1]. As a result, LDC transcribers found the RT-05S corpus easier to transcribe.

3 Transcription

3.1 Careful Transcription (CTR)

For purposes of evaluating transcription technology, system output must be compared with high-quality manually-created verbatim transcripts. LDC has already defined a careful transcription (CTR) methodology to ensure a consistent approach to the creation of benchmark data. The goal of CTR is to create a reference transcript that is as good as a human can make it, capturing even subtle details of the audio signal and providing close time-alignment with the corresponding transcript. CTR involves multiple passes over the data and rigorous quality control. Some version of LDC's current CTR specification has been used to produce test data for several speech technology evaluations in the broadcast news and conversational telephone speech domains in English, Mandarin, Modern Standard and Levantine Arabic as well as other languages over the past decade. The CTR methodology was extended to the meeting domain in 2004 to support the RT-04 meeting speech evaluation, and was used in producing this year's conference room evaluation corpus [2].

Working with a single speaker channel at a time (using head-mounted microphone recordings where available), annotators first divide the audio signal into virtual segments containing speaker utterances and noise. At minimum, the audio is divided into individual speaker turns, but long speaker turns are segmented into smaller units. Speaker turns can be difficult to define in general and are particularly challenging in the meeting domain due to the frequency of overlapping speech and the prevalence of side conversations that occur simultaneously with the main thread of speech. Further, speakers may utter comments under the breath that are difficult to distinguish from non-speech sounds, even when listening to a head-mounted microphone signal. Transcribers are therefore generally instructed to place segment boundaries at natural breakpoints like breath groups and pauses, typically resulting in segments of three to eight seconds in duration. In placing segment boundaries, transcribers listen to the entire audio file in addition to visually inspecting the waveform display, capturing any region of speech (no matter how minimal) as well as isolating certain speaker noises including coughs, sneezes, and laughter. Breaths are not specifically captured unless they occur around a speaker utterance. Transcribers are instructed to leave several milliseconds of silence padding around each segment boundary, and to be cautious about clipping off the onset of voiceless consonants or the ends of fricatives. Meeting segmentation practices do not differ substantially from those for other domains, but additional care is taken to create segment boundaries that respect the natural flow of the conversation, particularly with respect to the speaker turn issues mentioned above.

After accurate segment boundaries are in place, annotators create a verbatim transcript by listening to each segment in turn. Because segments are typically around five seconds, it is usually possible to create a verbatim transcript in one listen; but difficult regions that contain speaker disfluencies or other phenomena may warrant several reviews. No time limit is imposed, but annotators are instructed to utilize the "uncertain transcription" convention if they need to review a segment three or more times. A second pass checks the accuracy of the segment boundaries and transcript itself, revisits sections marked as uncertain, and adds information like speaker identity, background noise conditions, plus special markup for mispronounced words, proper names, acronyms, partial words, disfluencies and the like. A final pass over the transcript is conducted by the team leader to ensure accuracy and completeness. The individual speaker channels that have been transcribed separately are then merged together. Senior annotators listen to the merged files and use the context of the full meeting to verify specific vocabulary, acronyms and proper nouns as required. Further automatic and manual scans over the data identify regions of missed speech, correct common errors, and conduct spelling and syntax checks, which identify badly formatted regions of each file.

3.1.1 Quality Control

The meeting domain presents a number of unique challenges to the production of highly accurate verbatim transcripts, which motivates the application of quality control procedures as a part of the multi-pass strategy described above. One such challenge is the prevalence of overlapping speech. In meetings, overlap is extremely frequent, accounting for well over half the speech on average. Even when transcribing from the individual speaker recordings, capturing overlapping speech is difficult. Other speakers are typically audible on close-talking microphone channels, and transcribers must focus their attention on a single speaker's voice while simultaneously considering the context of the larger conversation to understand what is being said. During all stages of transcription, transcribers and team leaders devote extra attention to overlapping speech regions.

Transcription starts with the individual head-mounted microphone recordings, which facilitates the accuracy of basic transcription. Senior annotators listen to all untranscribed regions of individual files, identifying any areas of missed speech or chopped segments using a specialized interface. Some meetings contain highly specialized, technical terminology and names that may be difficult for transcribers to interpret. To resolve instances of uncertainty, final quality checks are conducted on a merged file, which conflates all individual speaker transcripts into a single session that is time-aligned with a mixed recording of all head-mounted channels, or a distant or table-top microphone channel. This merged view provides a comprehensive check over the consistency of terminology and names across the file, and is conducted by a senior annotator who has greater access to and knowledge of technical jargon. Senior annotators also check for common errors and standardize the spelling of proper nouns and representation of acronyms in the transcript. Transcription ends with multiple quality assurance scans, which include spell checking, syntax checking, which identifies portions of the transcript that are poorly formatted (for example, conflicting markup of linguistic features), and expanding contractions.

3.2 Quick Transcription

The careful transcription process described above was used to prepare benchmark data for purposes of system evaluation. In addition, LDC selected a one-hour subset of the evaluation data for transcription using Quick Transcription (QTR) methodology. The goal of the QTR task is simply to "get the words right" as quickly as possible; to that end, the QTR methodology automates some aspects of the transcription process and eliminates most feature markup, permitting transcribers to complete a verbatim transcript in a single pass over the data. The QTR approach was adopted on a limited scale for English conversational telephone speech data within the DARPA EARS program [3], with real-time transcription rates of seven to ten times real-time. Automatic post-processing includes spell checking, syntax checking and scans for common errors. 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 disfluent or overlapping regions of the transcript. 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.

Manual audio segmentation is an integral part of careful transcription, but is very costly, accounting for 1/4 or more of the time required to produce a highly-accurate verbatim transcript. To reduce costs in QTR, we developed AutoSegmenter, a process that pre-segments a speech file into reasonably accurate speaker segments by detecting pauses in the audio stream. AutoSegmenter achieves relatively high accuracy on clean audio signals containing one speaker, and typically produces good results on the head-mounted microphone channels. If the audio is degraded in any way, however, the quality of automatic segmentation falls dramatically, leading to large portions of missed speech, truncated utterances, and false alarm segments – segments that may have been triggered by noise, distortion, or other meeting participants. In the QTR method, segment boundaries produced by AutoSegmenter are taken as ground truth and are not altered or manually verified, since doing so would result in real-time rates far exceeding the target of five times real-time.

3.2.1 Quality Control

Quality assurance efforts are minimized for QTR, since the goal of this approach is to produce a transcript in as little time as possible. A quick quality assurance check was applied to the five transcripts were reviewed in a quick final pass, which involved a spell check, a syntax check and some basic formatting standardization including the removal of "empty" segments – that is, false alarm segments created by AutoSegmenter that contain no speech. (Typically these segments contain background noise or other speaker noise which under QTR is not transcribed.) Additionally, the contractions in each file were expanded. Transcripts were not reviewed for accuracy or completeness.

3.3 CTR vs. QTR: A Contrastive Study

With one hour of the conference room evaluation data transcribed using both the CTR and QTR methods, comparison of the resulting data is possible. Practical constraints of time and funding prevented us from providing a complete quantitative analysis of discrepancies during RT-05. While LDC's transcription toolkit does include processes to automatically compare and calculate agreement rates for multiple transcripts of the same source data, existing infrastructure assumes that segment boundaries are identical for transcripts being compared. The data created for RT-05 does not meet this requirement; CTR files contain manual segment boundaries while QTR files contain autosegments. However, a qualitative comparison is still possible.

In general terms, careful transcription offers maximum transcript accuracy, but it is time consuming and costly. Quick transcription by contrast is much more efficient, but does not maintain the same level of accuracy. Both methods may be called for to suit particular needs (for example, CTR for benchmark evaluation data; QTR for large-volume training data).

Comparison of the CTR- and QTR-produced transcripts of the five sessions reveals discrepancies in both segmentation practices and orthographic completeness. These categories are not orthogonal: many orthographic errors are caused by the automatic assignment of segment boundaries and the time constraints imposed in QTR. The following table shows the most common differences between QTR and CTR transcripts. Highlighting indicates higher accuracy or completeness.

Table 1. Common discrepancies between Quick and Careful transcripts

	QTR	**CTR**
transcription	word substitutions (e.g., **and** instead of %**um**)	careful word transcription
	no indication of speaker restarts, disfluencies	indication of speaker restarts, disfluencies
	lacking some punctuation, capitalization	standard punctuation, capitalization
	lacks special markup (for filled pauses, acronyms, mispronounced words, etc.)	contains special markup (for filled pauses, acronyms, mispronounced words, etc.)
	misinterpreted acronyms	acronyms verified
	misinterpreted, inconsistent transcription of technical jargon	technical jargon verified
segmentation	isolated breaths segmented (captured by AutoSegmenter)	no isolated breaths captured (in accordance with task specification)
	words dropped out of segment	careful word segmentation – no missed words
	split words (1- **it** 2-**'s**)	no split words (it's)

3.3.1 Orthographic Discrepancies

The quality and completeness of orthography and transcription content is necessarily lower with QTR, given the abbreviated real-time rate goals of this method. According to the task definition, QTR is an effort to "get the words right." A quick transcript will contain limited or no special markup, inconsistent capitalization and fewer punctuation marks. Meeting sessions contain specialized and sometimes highly technical content. During the CTR quality control process, senior annotators investigate the meeting context and relevant jargon to resolve any cases of uncertainty on a transcriber's part. However, during the quick transcription process, which targets a transcription rate of five times real-time, no time is allocated to researching specialized vocabulary. As shown in the example below, this can result in mis-

Table 2. Transcription discrepancies in ICSI_20010531-1030

QTR	CTR
689.110 692.550 me013: ((**roar**)) digits and and stuff like that. the me- the meeting meeting 692.790 694.240 me013: is uh later today.	689.075 691.400 me013: @**AURORA** digits, and and stuff like that. 691.400 694.250 me013: The mee- the meeting meeting is %uh later today.

transcribed segments, or the use of the "uncertain transcript" flag, denoted by double parentheses.

It is always possible for two transcribers to interpret non-technical speech differently. In CTR, these errors are typically eliminated through repeated quality assurance passes over the data, which specifically target accuracy and consistency across all speakers in a given session and resolution of cases of transcriber uncertainty. Consider the following example:

Table 3. Common transcription discrepancies in CMU_20050301-1415

QTR	CTR
157.130 161.370 fLDKKLH: I ((**officially**)) I don't know. I thought it was more around forty seconds though for that	156.750 161.325 fLDKKLH: **I've usually** -- I don't know, I thought it was more around forty seconds though for that.

This and the previous example show how the faster real-time rate in QTR affects transcription quality.

3.3.2 Segmentation Discrepancies

The limitations of automatic segmentation in the meeting domain become abundantly clear when comparing segments in QTR and CTR transcripts. Meeting data introduces its own set of hurdles, such as ambient noise and multiple simultaneous speakers. Adjusting the AutoSegmenter threshold to capture all speech and noise from the targeted speaker, while excluding noises and isolated breaths or other non-transcribed material, is extremely difficult. In light of such challenges, the automatically generated segment boundaries may chop off words, parts of sentences, or eliminate entire utterances. Inaccurate segmentation of the speech signal can change the meaning of

Table 4. Segmentation discrepancies in CMU_20050301-1415

QTR	CTR
224.810 226.180 fZMW: But uh yeah, I agree.	224.575 226.325 fZMW: That's what I (()) -- yeah. I agree.

the utterance itself, as in this example, where the QTR segment starts approximately .25 seconds later than the CTR segment.

The impact of low amplitude on segmentation and transcription in general can be significant. In the example below, even careful manual segmentation and transcription was made difficult by a weak audio signal. The QTR rendering of the excerpt below is extremely impoverished, lacking approximately 50% of the words captured in the CTR version.

Table 5. Segmentation/transcription discrepancies in VT_20050318-1430

QTR	CTR
800.470 800.980 rehg-g: Wright State.	800.450 804.150 rehg-g: ^Wright State student, he wants to continue at ^Wright State, so that's 804.400 805.800 rehg-g: that's his preference school.
801.650 803.310 rehg-g: he wants to continue at Wright State.	
804.580 805.360 rehg-g: That's his preference.	
[missed]	806.650 809.400 rehg-g: %um In biomedical engineering and %uh
[missed]	810.950 813.900 rehg-g: kind of interesting in his write up because he said he wanted to %uh
[missed]	814.800 816.725 rehg-g: design ^Luke ^Skywalker's hand.
[missed]	817.025 817.975 stephen-e: {laugh}
[missed]	817.600 818.525 rehg-g: It's like wow. {laugh}
820.490 821.250 rehg-g: said when whenever	819.325 825.250 rehg-g: Because you know every -- he said when whenever whenever someone asks him what he wants to do, what he's doing that's the easiest way to describe it.
821.510 825.080 rehg-g: Whenever someone asks him what he wants to do or what he's doing that's the easiest way to describe it.	

Another common automatic segmentation error is a form of truncation that occurs when complete utterances are captured by AutoSegmenter but are chopped in half in the presence of short pauses, as in the following example:

Table 6. Segment truncation in CMU_20050301-1415

QTR	CTR
67.640 68.950 fZMW: if it's longer than twenty sec-	67.425 70.350 fZMW: If it's longer than twenty seconds then they have a problem.
67.980 68.930 fLDKKLH: Mhm.	67.875 68.600 fLDKKLH: Mhm.
69.060 70.410 fZMW: -conds then I have a problem.	

In this example, AutoSegmenter detected a 0.1 second pause in the middle of the word "seconds," resulting in a halved word. Where the audio signal is clean and the amplitude is high, AutoSegmentation provides good results:

Table 7. Similarities between CTR and QTR in CMU_20050301-1415

QTR	CTR
131.810 137.910 mVHQQMY: I- is that due to uh speech recognition in general or just with whatever particular system they were using?	131.725 137.750 mVHQQMY: I- is that due to %uh speech recognition in general or just with whatever particular system they were using?
136.450 140.980 fZMW: We never had this problem though I was kind of surprised that they went back to this old system.	136.275 141.225 fZMW: We never had this problem though I was kind of surprised that they went back to this old system.

The level of accuracy demonstrated in the previous example makes a case in favor adopting a QTR-style approach to at least parts of the transcription process.

3.4 Transcription Rates

A fundamental challenge in transcribing meeting data is simply the added volume resulting from not one or two but a half a dozen or more speakers. A typical thirty-minute telephone conversation will require twenty hours or more to transcribe carefully (30 minutes, two speakers, 20 times real-time per channel). A meeting of the same duration with six participants may require more than 60 hours to produce a transcript of the same quality. LDC careful transcription real-time rates for the RT05S two-hour dataset approached 65 times real-time, meaning that one hour of data required around 65 hours of labor (excluding additional QC provided by the team leader). Examined in light of the number of total channels, however, the real-time rate for careful transcription per channel is around 15 times real-time, comparable with rates for BN and slightly less than that for CTS. Methods like Quick Transcription can cut these times considerably, but the volume of effort required is still substantial. The real-time rate for quick transcription of a one-hour dataset is about 18 times real-time; the real-time rate per channel is around four times real-time.

4 Infrastructure

Specialized software and workflow management tools can greatly improve both efficiency and consistency of transcription, particularly in the meeting domain. The nature of meeting speech transcription requires frequent jumping back and forth from a single speaker to a multi-speaker view of the data, which presents a challenge not only for the transcribers, but for the transcription tools they use. Current transcription tools are not optimized for this approach (or in many cases do not permit it at all). Further, different languages and domains currently require different tools (leading to lack of comparability across results). For the most part existing transcription tools cannot incorporate output of automatic processes, and they lack correction and adjudication modes. Moreover, user interfaces are not optimized for the tasks described above, in particular QTR. To support the demand for rapid, efficient and consistent transcription, LDC has created a next-generation speech annotation toolkit, XTrans, to directly support a full range of speech annotation tasks including quick and careful transcription of meetings. XTrans utilizes the Annotation Graph Toolkit [4, 5] whose infrastructure of libraries, applications and GUI components enables rapid development of task-specific annotation tools. Among the existing features, XTrans

- Operates across languages
- Operates across platforms
- Supports transcription across domains
- Contains customized modules for Quick Transcription, Careful Transcription and Rich Transcription/Structural Metadata markup
- Includes specialized quality control features; for instance speakerID verification to find misapplied speaker labels and silence checking to identify speech within untranscribed regions.
- Contains an "adjudication mode", allowing users to compare, adjudicate and analyze discrepancies across multiple human or machine-generated transcripts

As an added feature of great benefit to meeting transcription, XTrans allows users to easily move back and forth between the multi- and single-speaker views, turning individual channels on and off as required to customize their interaction with the data.

Two Data Views in XTrans

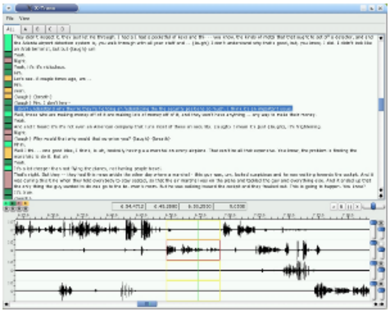

Fig. 1. Global speaker view in XTrans

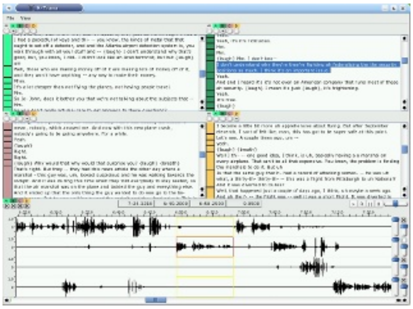

Fig. 2. Individual speaker view in XTrans

XTrans also automates many common annotation tasks, for instance removing the need for repetitive keystrokes and allowing the annotator to speed up audio playback. A timer function will also enforce transcriber efficiency by warning users (and reporting to managers) when transcription rates exceed the targeted real-time rate for a given task. As with LDC's current transcription tools, XTrans will be fully integrated into LDC's existing annotation workflow system, AWS. AWS controls work (project, file) assignment; manages directories and permissions; calls up the annotation software and assigned file(s) for the user; and tracks annotation efficiency and progress. AWS allows for double-blind assignment of files for dual annotation, and incorporates adjudication and consistency scoring into the regular annotation pipeline. Supervisors can query information about progress, efficiency and consistency by user, language, data set, task, and so on.

5 Future Plans and Conclusion

LDC's planned activities include additional transcription in the meeting domain as well as new data collection. Using existing facilities at LDC developed for other research programs, meeting collection is currently opportunistic, with regularly scheduled business meetings being recorded as time allows. Five hours of English meetings, three hours of meetings in Chinese and another two hours in Arabic have already been collected under this model. As new funding becomes available, we also plan to develop our collections infrastructure with additional head-mounted and lavaliere microphones, an improved microphone array, better video capability and customized software for more flexible remote recording control. While the current collection platform was designed with portability in mind, we hope to make it a fully portable system that can be easily transported to locations around campus to collect not only business meetings but also lectures, training sessions and other kinds of scenarios.

Future plans for XTrans include incorporation of video input to assist with tasks like speaker identification and speaker turn detection. We also plan to add a "correction mode" that will allow users to check manual transcripts or verify output of automatic processes including autosegmentation, forced alignment, SpeakerID and automatic speech recognition output. A beta version of XTrans is currently under testing, and the tool will be freely distributed from LDC beginning in late 2005 [6].

Shared resources are a critical component of human language technology development. LDC is actively engaged in ongoing efforts to provide crucial resources for improved speech technology to RT-05 program participants as well as to the larger community of language researchers, educators and technology developers. These resources are not limited to data, but also include annotations, specifications, tools and infrastructure.

References

1. Strassel, S., Glenn, M.: Shared Linguistic Resources for Human Language Technology in the Meeting Domain. Proceedings of the ICASSP 2004 Meeting Recognition Workshop. (2004) http://www.nist.gov/speech/test_beds/mr_proj/icassp_program.html

2. Linguistic Data Consortium: RT-04 Meeting Transcription Guidelines. (2004) http://www.ldc.upenn.edu/Projects/Transcription/NISTMeet/index.html
3. Strassel, S., Cieri, C., Walker, K., Miller, D.: Shared Resources for Robust Speech-to-Text Technology, Proceedings of Eurospeech (2003)
4. Bird, S., Liberman, M.: A formal framework for linguistic annotation. Speech Communication, (2001) 33:23–60.
5. Maeda, K., Strassel, S.: Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium. Proceedings of the 4th International Conference on Language Resources and Evaluation (2004).
6. http://www.ldc.upenn.edu/Projects/Transcription/Tools