Roadmaps and Resource Maps: Coordinating Language Resources for Cooperative Technology Development and Evaluation

Christopher Cieri and Kazuaki Maeda

Linguistic Data Consortium, University of Pennsylvania
{ccieri|maeda}@ldc.upenn.edu

Abstract

Advances in speech technologies increase demand for linguistic data in more languages with more sophisticated annotation. In speech recognition research, corpora of dozens of broadcast hours, hundreds of conversations and tens of millions of words of text are replaced by thousands of broadcast hours, tens of thousands of conversations and billions of words of text. The scope, scale and schedule of language resource development continues to evolve and the availability of high quality language remains a central issue for speech researchers. Although advancements in desktop computing allow researchers to create corpora in their offices, most cannot maintain the scale of corpus development their research requires. Instead they seek to find, re-annotate, reuse and re-distribute existing corpora. International data centers support grass-roots corpus creation by providing infrastructure, tools and best practices. However, such data centers must assume a more central role in creating and coordinating the creation of language recourse and must collaborate with their peer organizations to identify and address the needs of research communities and commercial developers. Initiative such as the ENABLER (European National Activities for Basic Language Resources) Project recognize and have created technology roadmaps to begin to address this need. The present paper will describe language resource activities in the United States focusing on those that coordinate effort across multiple areas of inquiry.

1. Introduction

The Linguistic Data Consortium was founded in 1992 within the context of the common task research management model employed by the U.S. Advanced Research Projects Agency (ARPA, DARPA). Under this model, multiple research organizations cooperate and compete to solve groups of related problems, as defined collaboratively with the sponsor, using shared data and evaluation methods. In 1992, the principle problem LDC was tasked to address was the distribution of existing linguistic databases and the coordination of new data creation efforts. LDC was also required to become self-supporting within five years.

Long before that initial phase ended, LDC had indeed become self-supporting and had identified a new challenge. Successes under the common task methodology coupled with promising results from statistical machine learning created demand for language which exceeded international capacity. What was needed was an organization devoted to large scale data collection. In 1995, LDC added data collection to its mission and began working on the CALLHOME project. CALLHOME created resource kits to support technology development for large vocabulary conversational telephone speech recognition. The corpora, in Japanese, Mandarin, Egyptian Colloquial Arabic, Spanish, German and English contain 200 transcribed, thirty-minute telephone conversations with lexicons that provide pronunciation, frequency information and, where necessary, morphological analysis and romanization of all surface forms in the transcripts. In 1998, responding to new demand from its members, LDC again expanded its mission to include annotation of linguistic data and development of tools and best-practices.

LDC’s most recent challenge, for highly coordinated linguistic resources, springs from two DARPA programs, TIDES and EARS. TIDES (Translingual Information Detection, Extraction and Summarization) began in 2000 to create the component technologies of a multilingual, multimodal news understanding system. In 2002, EARS (Effective, Affordable, Reusable, Speech-to-text) began to create high accuracy, multi-channel speech to text systems whose output will be both readable and richly annotated to support further processing. Although both TIDES and EARS employ common task methodology, their tasks are broadly defined and very ambitious requiring intensive, multilingual resource creation with careful coordination across research areas – in short, resource mapping. In the paragraphs that follow, we will suggest that, and describe efforts embracing the idea that, international data centers must adopt a broad mission which encompasses tools and best practices as well as data and must commit themselves to meaningful international coordination of resource creation efforts.

2. Talkbank

Talkbank is a National Science Foundation funded (BCS-998009, KDI, SBE), interdisciplinary project seeking to use new computational technologies to foster fundamental research into human and animal communication. LDC's role in Talkbank has been to
address the needs of multiple communities for linguistic data, standards and tools for creating that data and infrastructure for finding and sharing it. In 2003, LDC published corpora in four Talkbank areas: sociolinguistics, sign and gesture, morphology and conversation analysis with a corpus of animal communication to follow in 2004. Talkbank subsidizes the cost of distributing the first 50 or 100 copies of its corpora.

2.1. Annotation Graph Toolkit
To address the needs for standards and tools, LDC has created the Annotation Graph Toolkit (AGTK). The data model used in this toolkit is based on the Annotation Graph (AG) framework proposed by Bird and Liberman (2001). An annotation graph is a directed acyclic graph where nodes are optionally anchored with offsets and arcs can be labeled with multi-field records. Bird and Liberman demonstrated that a great variety of existing linguistic annotations can be represented with the AG model. AGTK provides an implementation of the AG model (AGLIB) as well as various software components for creating linguistic annotation tools. The AG API allows users to create, modify and manipulate AG objects. AGLIB provides two file formats, one XML-based format and one tabular format, for storing AG objects. LDC has also created plug-in file I/O modules for a variety of existing file formats. AGTK is open-source and available at http://agtk.sf.net.

AGTK also provides tools using AGLIB for creating and annotating a variety of linguistic corpora. TableTrans is a spreadsheet style annotation tool for coding speech and animal communication recordings. MultiTrans supports the transcription of multi-speaker and/or multi-channel recordings, such as the audio from meetings collected with multiple microphones. Using AGLIB components including plug-in I/O functions and shared display widgets, LDC programmers have been able to more rapidly develop fully-featured annotation tools that are highly customized to task and thus permit more efficient annotation.

2.2. OLAC
The Open Language Archives Community (OLAC) is an international partnership of 25 archives that have created a virtual library of language resources by “(1) developing consensus on best current practice for the digital archiving of language resources, and (2) developing a network of interoperating repositories and services for housing and accessing such resources.” LDC has supported OLAC by contributing to the definition of its standards and the implementation of the OLAC infrastructure. Standards include the OLAC Metadata schema for describing linguistic resources. The infrastructure includes software for data aggregating, repository editing, archive registration and service registration. LDC also created and hosted the initial OLAC website.

3. TIDES
The DARPA TIDES program seeks to create a news understanding system capable of responding to queries in the operating language by searching multilingual, multimodal news archives, identifying and summarizing relevant documents translating them into the operating language while maintaining links to the original source material. The TIDES operating language is English though TIDES technologies process Chinese and Arabic as well. Selecting three target languages allowed research sites to make rapid progress and compare results while encouraging language independent solutions. Experiments in Cebuano and Hindi (Cieri, Maxwell and Strassel forthcoming) have demonstrated the language independence of TIDES technologies.

TIDES divides its main goal into four research areas: finding relevant news stories (Detection), identifying important entities, relations and events in those stories (Extraction), synthesizing information both within and across stories (Summarization) and translating stories or summaries into the operating language (Translation). Each area contributes to the overall TIDES goal, yet each has its own resource needs and historical practice. LDC’s challenge was to create and distribute enormous volumes of high quality linguistic data to serve the needs of both the individual research areas and the program as a whole.

3.1. TIDES Resources
All TIDES language resources are based upon a core set of raw data news text, broadcast news audio and transcripts and parallel news text harvested from the Internet or created under contract. The Gigaword News Text corpora (English: 1-58563-260-0, Chinese: 1-58563-261-9, Arabic: 1-58563-271-6) provide orders of magnitude more data than previously available in order to support robust language modeling of even uncommon phenomena. Parallel text, harvested from the Internet and sentence aligned, support translation and translingual information retrieval. The Chinese-English Translation Lexicon (1-58563-238-4) Buckwalter Arabic Morphological Analyzer (1-58563-257-0) provide coverage of surface forms appearing in important subsets of TIDES text. The English (1-58563-163-9), Chinese (1-58563-204-X) and Arabic (1-58563-261-9) Treebanks, created under the direction of Mitch Marcus, Martha Palmer and Mohamed Maamouri, respectively.

The number following any mention of a published linguistic database is its ISBN.
respectively, provide lexical, morphological and syntactic tagging of subsets of TIDES texts to support the development of automatic taggers and parsers.

In addition to these core resources, LDC has created corpora specifically to train and evaluate systems in the multiple TIDES research areas. The TDT corpora, built for the Detection community, contain transcribed, story-segmented broadcast news audio and newswire annotated to indicate relevance of all stories to sets of event-based topics selected at random from the corpus. LDC has published the corpora produced during three of the five TDT phases (TDT-2 Multilingual Text: 1-58563-183-3, English Audio: 1-58563-160-4, Chinese Audio: 1-58563-184-1, TDT-3 Multilingual Text: 1-58563-193-0, English Audio: 1-58563-185-x, Chinese Audio: 1-58563-186-8). Extraction research is based upon the ACE Corpora (e.g. 1-58563-270-8) which contains news text and transcribed audio with annotations of entities and the relations among them. To support machine translation evaluation, LDC has created a new genre, the Multiple Translation Corpus. These corpora contain news stories segmented into sentences and then translated by multiple commercial agencies. The best of the resulting human translations serve as reference material for evaluating MT system output from both TIDES and commercial off the shelf systems.

All TIDES raw data and all annotations have been licensed for distribution both within and beyond TIDES. The time epochs of all data types overlap with the news text providing a ten year superset. All TIDES participants have equal access to the data via LDC regardless of whether they are Consortium members. The latest version of every TIDES test set is reserved for internal use until it is superseded. Because all data is licensed for use across all TIDES areas, TIDES researchers often benefit from data produced by other areas. The TDT corpora were developed to support Detection research but have been reannotated for use by the Extraction, Summarization and Translation communities within TIDES as well as by the EARS community among others.

4. EARS
The DARPA EARS (Effective, Affordable, Reusable Speech-to-Text) program is developing robust speech recognition technology for a range of languages and speaking styles whose outputs will be substantially richer and more accurate than currently possible. Like TIDES, EARS technologies are required to process English, Chinese and Arabic but should be portable to new languages. EARS speaking styles include broadcast news (BN), conversational telephone speech (CTS) and meetings. EARS research areas include Speech-to-Text (STT), Metadata Extraction (MDE). The goal of the STT group is to provide an order of magnitude reduction of word error rates particularly in CTS. The MDE community strives to process transcripts in order to identify speakers, to identify and thus punctuate meaningful units and to identify and thus remove disfluencies.

4.1. Fisher Collections of Conversational Telephone Speech
LDC provides number of different types of resources to support EARS research. Naturally EARS researchers use the Gigaword News Texts and previously published corpora of transcribed BN and CTS as well as pronunciation lexicons. However, in order to develop systems that are robust in the face of variation in speaker and channel, it was necessary to create new kinds of data for EARS. Existing telephone collections such as CALLHOME, designed to jump-start speech recognition in 6 languages, CALLFRIEND, designed to support language identification, and Switchboard, designed to support speaker identification, were each suboptimal for EARS purposes. Working with the EARS community, LDC created a new protocol, Fisher, to collect relatively small numbers of short calls from large numbers of participants who do not know each other and who discuss assigned topics in order to assure a diverse vocabulary. The Fisher English collection contains 16,446 ten-minute calls from 14,385 subjects yielding 2,741 hours of total audio with roughly equal number of males and females and balance across three age groups and four major U.S. dialect regions. A Fisher collection of Levantine Arabic is underway and will yield 1000 ten-minute conversations. A Fisher Chinese collection of 1200 conversations is in planning.

4.2. Quick Transcription of Conversational Telephone Speech
In order to achieve the kind of performance gains EARS requires, the program is providing a steady stream transcribed audio to researchers. The first year goal was to create transcripts of 2000 hours of English CTS. From the Fisher calls, LDC transcribed or contracted the transcription of eight minutes of each of 15,000 calls. Previous DARPA programs in CTS use transcription specifications that required up to 50 hours of effort to transcribe one hour of audio. Such rates are clearly impractical for the volume of audio EARS required. Working with EARS sites LDC created a new Quick Transcription (QTr) specification that requires between 5 and 7 hours of effort for each hour of audio. Under the QTr specification, an automated process segments a phone conversation into utterances of two to eight seconds duration for which a human provides a verbatim transcript in a single pass. The result is checked for spelling and normalized. Although automatic
segmentation sometimes misses utterances and although humans sometimes make transcriptions errors (especially in long or disfluent sections), EARS researchers are finding that the large increase in volume more than compensates for the small decrease in quality. The transcripts also offer, for the first time, the opportunity to create language models directly from an adequate volume of CTS. LDC is also engaged in a similar experiment to provide quick transcription of broadcast news.

4.3. Speech-to-Text

Although EARS sites will rely upon quick transcriptions for much of their training material, technology evaluation still requires very high quality transcripts. LDC produces one or two hours of carefully transcribed BN, CTS and meetings in English, Chinese and Arabic twice each year to support development-test and formal evaluations of EARS systems. The latest version of each yearly evaluation set is reserved for the internal use of the program until superceded. Since EARS has not had its second yearly evaluation, none of the evaluations sets are publicly available at this time. In addition one evaluation set will be held in reserve until the end of the program to provide a stable benchmark against which progress of EARS system may be gauged.

4.4. Metadata Extraction

Metadata Extraction (MDE) research, targets systems that refine the raw output of Speech-to-Text systems into forms that are more useful to humans and downstream processes.

In 2003, LDC defined an annotation task to support MDE research and created approximately 75 hours of English training, development and evaluation data. The MDE annotation task requires annotators to flag non-content words like filled pauses and discourse markers, identify and characterize sections of disfluent speech, and create boundaries between natural breakpoints in speech (SUs). LDC created an MDE annotation tool using AGTK which allowed programmers to effectively and quickly modify the tool and data format in response to evolving task definitions and demands for greater annotation speed and accuracy. Users highlight relevant spans of text, play the corresponding segments, and record annotation decisions with a few mouse clicks or keystrokes. Annotators verify their decisions by viewing the resulting “cleaned-up” transcript that removes fillers and disfluencies and displays each SU as a separate line of text.

Like TIDES, EARS research requires close coordination among the raw data sources, annotations created for training purposes and those created to evaluate STT and MDE technologies.

5. Conclusions

The projects described above each support the idea that international data centers must adopt a broad mission which includes creation and distribution of linguistic data, tools and best practices and the coordination of such efforts across projects in order to identify and address the needs of research communities and commercial developers.

6. References