ELSEVIER

Speech Communication 33 (2001) 1–4

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Editorial

Speech annotation and corpus tools

1. Introduction

In the last 20 years, there has been a pressing need to develop speech and language corpora as training and testing material for a wide range of speech technology applications. This has been coupled with a growing interest in the speech community to develop models of spoken language that are based on corpora that are increasingly representative of natural, spontaneous speech.

The growth in the use of speech corpora has benefited in the last 10 years from the establishment of data centres, such as the Linguistic Data Consortium (LDC), the European Language Resources Association (ELRA), the Japanese Language Resource Consortium (GSK: Gengo Shigen Kyouyuukikou), and multi-site annotation initiatives, such as the ToBI system for prosodic annotation and the DAMSL system of discourse annotation. Today hundreds of annotated speech corpora exist and are used worldwide, and the demand for richly annotated corpora is growing.

The growth in the use of corpora has, however, not been matched by the development of a standard set of tools for creating, editing, annotating and querying corpora: as a result, many laboratories have developed their own systems for corpus annotation and analysis, precisely because existing tools are ill equipped to cope with the increasing size and range of applications for which corpora were constructed. A wealth of formats and tools have sprung up around this enterprise, a diversity which at once facilitates and frustrates progress. The linguistic annotation page (www.ldc.upenn.edu/annotation) and a series of international workshops have drawn attention to the scale of ongoing activity, to the existence of diverse approaches to similar problems and of similar approaches to diverse problems. Despite the explicit formats and well-documented user interfaces, insights about the structure of the annotations themselves are often buried in coding manuals, internal data structures and file formats.

There are pressing needs to document data models and tool requirements, to identify notational and functional equivalences among different approaches, to report on new approaches to core representational problems, and to describe new domains and empirical problems which stretch our conceptions of the models. These needs are the focus of this special issue. The papers gathered here address a broad range of theoretical and practical issues concerning the representation of annotations, the structure of annotated speech corpora, and the design, analysis and implementation of tools for creating, browsing, searching, manipulating and transforming annotations and annotated speech corpora.

2. Themes of the papers

2.1. Scalability

To what extent are annotation tools and formats adapted to dealing with very large corpora? A number of papers touch on this issue. One approach is to use the Extended Markup Language (XML) as the data model (McKelvie et al.; Jacobson et al.), which allows data annotation to benefit directly from new developments in indexing, storage, query and transformation of large XML databases. Another approach is

PII: S0167-6393(00)00066-2

to provide a relational representation for annotation data (Bird and Liberman; Cassidy and Harrington), so that existing highly-optimized relational database technology can be applied to annotated corpora.

Apart from the annotation data itself, a separate scalability issue concerns the speech data itself. Barras et al. set out reasons why existing annotation tools were too cumbersome and inadequate to transcribe large quantities of broadcast news data. In response to this, they developed a system called 'Transcriber' which is optimized for the transcription of speech recordings lasting hours.

Another dimension to scalability concerns tool development. It is becoming impractical to develop new software for each new corpus project. McKelvie et al. employ XSLT to create task-specific user interfaces without the need for new programming. The projects described in three other papers (Barras et al.; Bird and Liberman; Cassidy and Harrington), are converging on a common API in the context of ATLAS, *Architecture and Tools for Linguistic Analysis Systems* (www.nist.gov/speech/atlas/).

2.2. Adaptability

Barras et al. have summarised succinctly some other design criteria that are more likely to ensure that a system is likely to be used for a wide range of corpus construction and annotation tasks. These include the possibility of porting the system to different platforms, developing procedures that will cut down on the annotation time (e.g., through an appropriate development of the graphical user interface – Barras et al.; McKelvie et al.; Jacobson et al. – or by automating components of the annotation – Syrdal et al.). Many authors also agree that speech annotation software is likely to reach a wider audience if it can be downloaded and used without being burdened by complicated (and costly) licensing arrangements.

2.3. Representing multilayered structures

Speech is inherently a multilayered activity in which parallel streams of overlapping symbolic information are transmitted from speaker to hearer. Accordingly, most of the papers grapple with the problem of how to provide a computational architecture for representing speech at multiple levels of granularity, with various layering methods for representing logically independent annotations as linked but autonomous.

XML provides mechanisms for representing structured information linked to signals. For text annotation the markup is typically embedded in the stream of characters. However, for speech annotation the signal data is usually left in its original format and simply *referenced* from the annotation using a sample number or time offset, a technique called remote or stand-off markup. This method permits annotations to reference other annotations, and not just signals, and provides a natural solution to the challenge of representing multiple intersecting hierarchies in XML. Consider a sequence of three words w_1, w_2, w_3 where the first two words form some constituent a and the last two words form some constituent b. The necessary marked-up representation would be something like the following: <a>$w_1$$w_2$$w_3$. This representation is not well-formed XML because the tags are not properly nested.

McKelvie et al. represent independent hierarchies and layers using separate files which contain their own XML structures. Independent layers are synchronized by hyperlinks to common lower-level layers. Jacobson et al. are able to represent their multilayer structures in a single XML hierarchy, and do not need to use the above method.

Taylor et al. group their basic units into layers using binary relations; two units related by the transitive closure of some relation r are said to be in layer r. Cassidy and Harrington provide a type hierarchy, associating each of their basic elements with one of the types, and restricting the associations between their layers to those licensed by the hierarchy. Bird and Liberman permit their annotation graphs to be disjoint, allowing different layers to be maintained independently. Each layer is a well-formed annotation on its own, and multiple layers can be combined into a single multilayered annotation using a union operation.

2.4. Querying corpora

In both the MATE (McKelvie et al.) and the EMU (Cassidy and Harrington) systems, query languages have been developed for retrieving information from multilevelled speech corpora. Both systems allow complex combinations of sequential and hierarchical queries that can then be used to extract whatever signal file data is available in the corpus (e.g., for a tones-and-break-indices type of annotation: find all phrase-final syllables that are immediately preceded by bitonally accented syllables in an H–H% type of intermediate/intonational phrase). One of the ways in which the systems differ is that the retrieval of symbolic information and signal file data has to be accomplished as two separate instructions in EMU, but not in MATE: that is, a single query in MATE could be used to refer to fundamental frequency values for particular kinds of syllables, but in EMU, it would be necessary first to find the syllables and then to pass those syllables to separate commands for obtaining their F0 values. MATE also has a set of primitives for extracting information from temporally overlapping hierarchical structures, which are particularly relevant for annotating spontaneous speech corpora.

2.5. Discourse-prosody interface

The development of annotated corpora of spontaneous speech has enhanced various research initiatives at the interface between natural language and speech processing. One such area, which is essential for improving the naturalness of text-to-speech systems, is discourse structure and text analysis and its relationship to prosody. In Syrdal et al.'s paper, grammatical category, punctuation and sentence length are used to predict intonation phrase boundaries and some of these, as well as a prediction of 'given' and 'new' information are used to mark words as pitch-accented. Beyond contribution to various theoretical issues at the discourse-prosody interface, their system has the potential to save transcription time which can take anything up to 30 minutes in a manual ToBI transcription of a 10 second utterance. Their 'semi-automatic method', in which tones and break indices were automatically predicted from a discourse and grammatical analysis of the text (requiring the transcribers to adjust these subsequently if necessary), speeded up transcription time considerably compared with the 'scratch' method in which no ToBI labels were provided. They also showed that the transcribers' choice of labels was not biased when they were presented with a set of labels by the 'semi-automatic' method.

The extent to which transcription time is improved must also depend on the accuracy with which the system can predict ToBI labels from a discourse/text analysis: this is the main area of investigation of Stirling et al. Their analysis of the correspondences between the two coding schemes (the HCRC system and the 'Switchboard' version of DAMSL) is also relevant to the annotation of discourse per se. As far as the discourse-prosody mapping is concerned, they show quite a close correspondence between dialogue act boundaries and major prosodic boundaries (breaks of 3 or 4 corresponding to an intermediate or intonational phrase) in both coding schemes in an analysis of several spontaneous speech dialogues. They also show that pitch reset at prosodic boundaries was often associated with discourse boundary strength and initiating a new discourse event.

3. Future directions

In the light of the papers in this collection, and the state of the field more generally, we can see a number of key areas where work is underway and where we can expect to see intensive activity in the near future.

A number of powerful general-purpose frameworks have been developed, which often include explicit XML formats for data storage and interchange, and application programming interfaces (APIs). Analysis of the formats and APIs, as well as identification of the substantive differences and the needs which

motivated them, will contribute to a deeper understanding of the nature of speech annotation and, we believe, to a widespread convergence between the existing models of speech annotation. Special purpose and domain-specific tools will be more easily incorporated in the general purpose frameworks, and will benefit from the expanded infrastructure for reusing data and software. The greater opportunities for sharing will facilitate the creation of very richly annotated corpora which combine the expertise of scholars working remotely from each other and using different platforms.

Another key development will be in the database area. As annotated corpora become larger and more complex, researchers will move away from conventional storage in plain unindexed text files. Storage in a relational or semistructured database, along with appropriate indexing, query languages and web-based annotation servers, will become the norm, with flat files only used for interchange and archive purposes.

The development of annotated speech corpora will continue to have a widespread impact both on speech and language technology research, for which they provide primary training and testing material, and on basic research in phonetics and linguistics. Speech corpora provide an effective way of testing theories of speech and language without each laboratory having to devote extensive resources to creating and annotating individual task-specific databases. As corpora become more comprehensive including, e.g., a wider range of speaker types and styles, they are likely to be used increasingly in speech and language research. However, we evidently need to develop more streamlined and cost-effective methods for corpus creation, as well as better tools for creating, editing, annotating and querying corpora. Moreover, these methods and tools must allow theories of speech and language to be easily represented and deployed by users who do not have an extensive background in computational techniques. Corpora will also continue to become more widespread in basic research if tools are developed that can be easily shared between users and integrated with graphical and statistical packages for visualization and analysis. Annotated corpora continue to provide one of the most fundamental links between speech and language technology research on the one hand and basic research in phonetics and linguistics on the other. The tools and techniques that are discussed in this special issue make an important contribution to consolidating the links between these two areas.

Acknowledgements

The editors are grateful to over 50 people who undertook timely and insightful reviews of the submissions. The papers by Bird and Liberman, and Cassidy and Harrington, were reviewed independently by the *Speech Communication* editors and timed to appear with this issue.

Steven Bird
University of Pennsylvania, Linguistic Data Consortium
3615 Market St., Suite 200
Philadelphia, PA 19104-2608, USA
E-mail address: sb@ldc.upenn.edu

Jonathan Harrington
Macquarie Centre for Cognitive Science and
Speech and Language Research Centre
Macquarie University Sydney, Australia
E-mail address: jmh@shlrc.mq.edu.au