Online Speech and Language Resources

LREC Tutorial

Research and development in speech and language technology are facing a fundamental paradigm shift. More and more speech and language resources, tools and even entire workflows are becoming accessible online in the context of the emerging research infrastructures.

In this tutorial, leading experts in the field of speech and language resources will discuss both opportunities and challenges of online resources. They will present state of the art technology for metadata, web services, persistent identifiers, and human and machine readable online repositories. Showcases and real world applications will illustrate how these technologies can be put into use.

The target audience of the tutorial are professionals from speech and language technology development, research and higher education.

1. Outline of the presentations

- Resource and Tool Creation
  - Metadata: specification, creation and use of metadata
  - Webservices: architecture and guidelines for webservices, case study of an existing webservice
- Online access to Resources
  - Persistent Identifiers: motivation, concepts and implementation
  - Repositories: architecture, harvesting, browsing, case study of an existing repository

Resource and Tool Creation

Metadata describes speech and language corpora, tools and other linguistic resources so that they can be indexed. A metadata schema must fulfil contradicting requirements: it must be sufficiently precise to adequately describe a resource, but it must be general enough to cover the wide variety of existing resource types; it must be stable for long-term access, but it must be easily adaptable to new types of resources or technical developments; it must be fine-grained to capture the relevant details of a resource, but at the same time efficient and easy to use by humans and automated processes. Finally, a metadata schema should be as theory-neutral as possible with respect to the primary resources to allow for broad application across disciplines.

Webservices are increasingly becoming popular in speech and language processing. Many tools have been freely available for years, but the efforts and technical expertise needed to install or run them locally prevented their wide adoption. Webservices offer an elegant solution: a tool runs on a server, and remote clients, e.g. web browsers, standalone annotation tools or application programs, access and exploit these services via the net.

In the tutorial, we present a component based and self-describing metadata schema built on the foundation of agreed standards and terminology in the field and show how tools can be used to generate metadata descriptions with minimum effort. Furthermore, the design and implementation of webservices, and their description with metadata, is presented in some detail; as a case study, the automatic speech segmentation system of the BAS will be used.

Persistent Identifiers and Repositories

Speech and language resources, as well as tools and processing workflows, evolve over time. Persistent identifiers provide a way of assigning a unique and immutable identifier to a specific version of a resource, and they may be used to refer to this resource independently of its physical storage location or means of access.

Repositories provide controlled access to language and speech resources and services both to humans, e.g. via a browser, as well as to automated processes, e.g. search engines or harvesters. Repositories require a minimum set of metadata, a flexible and powerful storage management, and access authorization, amongst other features. Although there exist software packages with repository functionality, they require considerable technical expertise to maintain.

In the tutorial, we present the alternative schemes for obtaining and maintaining persistent identifiers. Furthermore, we discuss the far-reaching consequences - including the benefits - of providing persistent identifiers for one's own resources, with a special focus on versioning and long-term storage. With respect to repositories we give an overview of existing software solutions including the integration of repositories and content management systems, and discuss in some detail the technical aspects of querying and harvesting language and speech repositories. As a case study, we will present the repository and services of the CLARIN-D centre Leipzig.

2. Presenters

Christoph Draxler, Bavarian Archive for Speech Signals, LMU Munich, head of the corpus and tools group. He has developed a number of speech tools, e.g. SpeechRecorder, WebTranscribe, and percy, and he was responsible for the collection of several large-scale speech databases, e.g. SpeechDat II and SpeechDat-Car (German), Ph@ttSessionz, VOYS

Thomas Eckart, Natural Language Processing Group, University of Leipzig, research associate. He has worked in several projects on the usage of large written language resources in the eHumanities and in infrastructure projects. He is co-developer of the Virtual Language Observatory (VLO).

Daniel Jettka, Hamburger Zentrum für Sprachkorpora, Hamburg, research associate. He has implemented the HZSK Repository for Spoken Language Corpora, created webservices for the conversion and visualization of transcription formats, and was part of several projects dealing with the creation and curation of spoken language corpora.

Alexander Geyken, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, head of DWDS and DTA project groups. He has coordinated the compilation of two large reference corpora for historical and contemporary written German (DWDS-Kernkorpus of the 20th century and the DTA core corpus (c17-c19)).

Dieter van Uytvanck, Max-Planck-Institute of Psycholinguistics, Nijmegen, is a research infrastructure specialist at The Language Archive. He has been involved in technical infrastructure building for LRT purposes since 2008.