Development of Data Analysis Talents

D-DATa

高度データ関連人材育成プログラム

WASEDA University

データ収集・活用技法集中セミナー

音声言語データベース構築編

シラバス・開催案内

2019年4月24日

D-DATaプログラム 運営事務局

データ収集・活用技法集中セミナーの目的

- 大量のデジタルデータを大量に収集できるようになる一方で、人間の生の行動データは、収集や整備に膨大なコストを要します。今後、データ駆動型社会がますます発展するにつれ、**データサイエンティストを目指す人は、人間の生の行動データから人間の行動意図や感情状態を推定するような高度で複雑な課題を解決することが求められます**。そのためには、**収集が容易なデータの扱い方だけではなく、生身の人間から得られる様々な行動情報の収集や整備・活用方法を身に付ける必要があります**。以下の二つの集中セミナーでは、音声言語領域の観点を中心に、人間の生体データの収集・整備・活用方法の知見の獲得を目指します。

音声言語データベース構築編

音声や会話のデータ利用に際しては、背景となる言語学などの体系的知識の活用やプライバシーへの配慮など、様々な点に留意することが求められます。データの収集から機械学習に活用できるデータベースの構築まで、一般的な課題や実際の技法を、汎用の音声言語データベース構築の実例に基づいて学習します。学習を通して、様々なノウハウと効率的な活用方法を獲得します。

特定利用生体データ収集編（仮）

※2019年度中開催予定（企画中）

ビジネスと研究の両観点において、解決すべき課題が明らかな場合におけるデータ収集は、要件定義などのデータ収集方針を明らかにすることからはじまり、収集したデータの品質・性質評価、効率的かつ安全な保管方法の検討、データベース格納・整備から利用方法の検討にいたるまで、多くの課題と知るべき手法が存在します。本セミナーでは、音声言語領域の幾つかの事例に基づき、課題解決のための一連の実作業を学習します。

データ収集・活用技法集中セミナーの全体構造

- D-DATaプログラムでは、データの分析やセキュリティ確保などの利用段階だけではなく、最近のトレンドでは見逃しがちなデータの準備の方法、すなわち収集や整備についても重視しています。
- 本集中セミナーは、人間の生体データのうち、音声言語データをいかに扱うかについて学びますが、Web上で収集できるログデータなど、デジタル世界で生じうるデータの扱い方についての集中セミナーも、別途企画しています。

D-DATaプログラム：データ収集・活用技法集中セミナー

生体データ編

- 音声言語データベース構築編
- 特定利用生体データ収集編（仮）

Webログデータ編

- Webログデータ収集：データ構造編（仮）
- Webログデータ収集：システム設計編（仮）

音声言語データベース構築編の概要

目標	➢ 生の行動データに基づくデジタルデータをより有効に活用できるようになる ➢ 自力で生の行動データを収集し、データ処理可能な段階まで整備するための知見を獲得する
受講要件	➢ 非情報系でも情報系でも、現時点の技能を問わず受講可能 ➢ 講義時にノートPC(Win/Mac)およびヘッドホン(イヤホン可)を持参可能であること ➢ 事前課題（特定ソフトのインストール程度）を実施可能であること
本講義のフォーカス	【本講義がカバーするもの】 ➢ 音声言語データの収集からデータベースを開発して検索などの利活用ができるようにするまでのプロセス・技術・環境 ➢ 言語学などの体系的知識を活用して汎用的に利用可能な音声言語データベース構築の体験 【本講義がカバーしないもの (ただし本講義の内容をもとに応用は可能】 ➢ 特定の目的（音声以外のメディア）に絞ったデータベース設計

開催スケジュール

開催日：2019/6/29(土)　時間：10:40-18:00
(昼休憩50分、休憩時間15分×2回)
場所：早稲田大学早稲田キャンパス付近（詳細は受講登録者に追ってお知らせします）
主催：早稲田大学D-DATaプログラム
共催：日本音声学会、オルトブリッジ・テクノロジー株式会社

時限	12月15日（土）
10:40-12:10	第１回：音声言語データベースの目的と価値
13:00-14:30	第２回：音声言語データベースの活用と構築
14:45-16:15	第３回：音声言語データベース構築の実習(1)
16:30-18:00	第４回：音声言語データベース構築の実習(2)

シラバス（第１回）　6/29 (土) 10:40-12:10

第１回　音声言語データベースの目的と価値

目標１：汎用の音声言語データベース（コーパス）の利用目的を理解する。

目標２：生の行動データを価値を高める形で利用するための基本的な考えを知る。

1.1 コーパスとは何か

学習項目： コーパスの基本的概念、構築の目的、アノテーションの重要性

- CSJ (Corpus of Spontaneous Japanese, 日本語話し言葉コーパス) の構成とサンプル
- コーパス構築の目的：音声認識研究、談話分析研究、言語処理研究、商用利用
- 高度なアノテーションによるコーパスの価値の高度化

1.2 音声言語研究とコーパスの歴史と発展

学習項目：音声言語研究の歴史、コーパス開発の歴史

- 統計的手法から深層学習へ
- 対話インタフェース開発・評価
- 音声言語コーパスを活用した最近の研究

シラバス（第２回）　6/29(土) 13:00-14:30

第２回　音声言語データベースの活用と構築

目標１：公開されているコーパスの入手方法を知り、代表的なコーパスの概観と各特長を理解する。

目標２：コーパスの内容・構築方法・利活用方法を知る。

目標３：生の行動データの価値を高める形で利用するための基本的な方法と留意点を知る。

2.1 代表的な音声言語データベースの公開状況

学習項目: 日本語の音声言語データベースの一覧、世界の音声言語データベースの一覧、企業におけるコーパスの整備状況など

- 国内・国外におけるコーパス公開状況
- コーパス検索・活用環境

2.2 代表的な音声言語データベース

学習項目：個々のコーパスの実例、音声言語データベースの検索・利活用方法

- データ収集方法／リッチなアノテーションの解説／アノテーションを統合したデータベースであるXML, RDBの紹介

2.3 音声言語データベースの構築と公開

学習項目：音声言語データの収集から音声データベースの構築までの方法、公開における個人情報や肖像権などへの留意

- 音声データ収集、音声言語データベース構築プロセス、ファイル／ディレクトリ構成、収録・公開に関する同意書

第３・４回　音声言語データベース構築の実習(1)(2)

目標：ミニワークショップ（実習）を通して、実際の構築を体験する。

実習の内容：

・音声ファイルからの転記

・形態素情報、音素情報、イントネーション情報、談話情報などの付与

・付与した情報の検索、横断的な整合性検証

・付与した情報の形式変換

・データベース化

教科書あるいは予習について

- 予習は必須ではありませんが、セミナー当日により深い理解を得たい方は、以下の教科書で事前学習することをお勧めします。
- 「話し言葉コーパス　設計と構築」小磯花絵(編)、朝倉書店、2015。

希望者には事前に教科書を貸し出します。
（冊数制限あり、先着順）

実習環境の準備について

- 実習を受けるためにご準備いただくこと
  - PCについて(Win/Mac)
    - **無線LAN接続に対応したPCをご持参ください。**
    - セキュリティの設定によっては演習環境が実行できない場合がありますので、ご自身でセキュリティ関係の設定を変更可能なPCの利用を推奨します。
    - 万が一の場合はPCや有線ＬＡＮアダプタを貸し出すこともできます。ご要望がある場合は事前にD-DATaプログラム事務局にご相談ください。
  - ヘッドホン(もしくはイヤホン)について
    - PCに接続可能なヘッドホン（イヤホン可）をご用意ください。タイムラグの発生を防ぐため、**有線接続式ヘッドホン**を用意することが望ましいです。

- インターネットへの接続方法
  - 学外の方には、早稲田大学の有線／無線ネットワークに接続するためのゲストアカウントをご用意します。当日までにお知らせいたします。

講師のご紹介

菊池　英明

博士(情報科学)、早稲田大学人間科学学術院　教授

音声言語、音声対話, ヒューマン・エージェント・インタラクションの研究に従事。
音声言語コーパスプロジェクトに多数参画し、オリジナルのコーパス設計や構築の経験も豊富。
さらに、大学での実習や学会チュートリアルを多数実施。
【略歴】
1991早大・理工・電気卒、1993同大大学院修士課程了。同年(株)日立製作所中央研究所入社。
早大理工総研助手、国立国語研究所非常勤研究員、早大人間科学部非常勤講師・専任講師・准教授を経て、
2012より早大人間科学学術院教授。国立国語研究所客員教授、理化学研究所客員研究員。博士(情報科学)。
【所属学会】
人工知能学会、日本音響学会、ヒューマンインタフェース学会、情報処理学会、電子情報通信学会等

沈　睿

博士(人間科学)、明海大学　総合教育センター　専任講師

コーパス言語学、自然言語処理、言語教育の研究に従事。
音声言語コーパスの開発とくに検索インタフェースのデザインが専門。
【略歴】
2005中国華東師範大学中国語教育卒、2009早大人間科学大学院修士課程了。
早大人間科学学術院助手、埼玉大学経済学部非常勤講師を経て、2015より明海大学専任講師、
早大人間科学部非常勤講師、早大eスクール非常勤講師、早大人総研招聘研究員。博士（人間科学）。

西川　賢哉

国立国語研究所 コーパス開発センター プロジェクト研究員

種々の音声コーパスの設計・構築に従事。音声言語コーパスプロジェクトに多数参画、コーパス構築環境の独自開発やチュートリアル講師を経験。
【略歴】
2000-2005 国立国語研究所 研究開発部門第2領域
2005-2009 慶應義塾大学大学院文学研究科博士後期課程（単位取得退学）
2009-2016 理化学研究所 脳科学総合研究センター 言語発達研究チーム
2016-　国立国語研究所 コーパス開発センター
【所属学会】
日本言語学会