ATR日英・日中対話データベース

ATRMADE： ATR Machine-Aided Dialog Expressions

～音声翻訳システムを介した対話データ～

概要

ATRMADE(エイティーアールメイド)は、音声翻訳システムを介したリアルなシーンでの模擬対話データベースです。(株)国際電気通信基礎技術研究所(ATR)音声言語コミュニケーション研究所において、音声翻訳システム利用時のコミュニケーション行動、および言語表現の調査研究のために作成されました。旅行シーンでのタスクや役割を設定された話者による自由な対話音声を収集しています。日英対話と日中対話があり、自然発話音声と書き起こしテキスト、機械翻訳と翻訳(人手)による対訳データがあります。

特徴

◆自然発話音声/書き起こしデータ　⇒ 音声認識モデルの作成・評価に利用
自然発話の音声データと書き起こしデータがあります。

◆対話データ(モノリンガル/バイリンガル)　⇒ 対話システム研究に利用
旅行に関する様々なシーンを想定した模擬対話データです。書き起こしテキストと対訳データを使用することにより、日本語/英語/中国語のモノリンガル対話、および日英/日中のバイリンガル対話データとして利用できます。

◆対訳データ(日英/日中)　⇒ 翻訳モデルの作成・評価に利用
発話音声に対する書き起こしデータとその翻訳データが日英・日中の対訳データとして利用可能です。
発話音声の一部は複数パターンの翻訳を行っており、パラフレーズとしても利用可能です。

話者数/対話数/発話量

◆ 話者数

ATRMADEは、日英対話と日中対話があり、話者数は、日英対話：153名、日中対話30名、あわせて183名です。

	日本語男性	日本語女性	英語男性	英語女性	中国語男性	中国語女性	合計
日英対話	44	37	37	35	0	0	153
日中対話	8	8	0	0	7	7	30
男女別計	52	45	37	35	7	7	183
言語別計	97		72		14		183

◆ 対話数/発話時間/総発話数

	対話数	発話時間[1]	発話時間内訳		総発話数	総発話数内訳	
日英対話	1,367	25.2	日本語	12.3	24,265	日本語	11,985
			英語	12.9		英語	12,280
日中対話	287	4.8	日本語	2.3	5,348	日本語	2,387
			中国語	2.5		中国語	2,961
合計	1,654	30.0	30.0		29,613		29,613

[1] 発話前後の無音区間を含みます。また、日英対話のうち3,523発話については、音声ファイルフォーマットがμ-law形式となり、発話時間計算の対象外としています。

ATR-Promotions

2021.06.10

収録条件

- 収録環境：残響の少ない静音環境
- 収録機材：携帯電話とPDA、ヘッドフォン付き接話マイクと小型PC、ヘッドフォン付き接話マイクとPDA、PDAのいずれか(収録時期により異なる)

データ構成

1．**テーブルデータ**：TSV(タブ区切り)形式、文字コード UTF-8(BOM無し)、改行コード LF
- 対話テーブル(.tbl)：対話の一覧
- 話者テーブル(.tbl)：話者属性(言語、性別、年代、出身情報など)
- シーンテーブル(.tbl)：想定したシーンの一覧
- タスクテーブル(.tbl)：想定したタスクの一覧
- 役割テーブル(.tbl)：想定した役割の一覧
- 発話コメントテーブル(.tbl)：発話に対してタグ付けしたマークの一覧(繰り返し発話、認識誤りなど)
- 収録状況テーブル(.tbl)：収録状況(タイピスト有無など)の一覧

2．**対話単位のデータ**：文字コード UTF-8(BOM無し)、改行コード LF
- 対話参照用HTML(.html)：対話単位で一連の発話をブラウザ表示
- 発話コメントファイル(.cmt)：発話に対してタグ付けした内容(繰り返し発話、齟齬、など)
- 対話属性ファイル(.env)：対話の属性(シーン、タスク、役割、話者情報、発話数、収録状況など)

3．**音声データ**：WAV形式、16kHz、16bit、モノラル （但し、日英対話の内3,523ファイルについては8kHz μ-law形式）
- 発話音声データ(.wav)：発話単位の音声データ

4．**発話単位のテキストデータ**：文字コード UTF-8(BOM無し)、改行コード LF
- 発話データ(.utt)：書き起こしテキスト(言い誤り等含む)、発話テキスト
- 翻訳データ(.ht)：発話テキストから人手で翻訳を行ったデータ
- 機械翻訳データ(.mt)：収録時に翻訳システムが対話相手に伝えた内容(但し、一部のみ)
- カナデータ(.kana)：カタカナファイル(但し、日本語、且つ、16kHzの音声ファイルについてのみ)
- セグメントデータ(.si)：音声ファイルに対する発話区間の時間情報

※ (　)はファイルの拡張子を表す

納品物

- 構成：データ構成物、マニュアル
- 納品メディア：DVD (約4.1GB)

価格 (消費税抜)

商用利用：¥3,000,000　　研究利用：¥1,200,000

※　アカデミック価格についてはお問合せ下さい。
※　日中対話のみ、日英対話のみについてもご相談下さい。

ご利用方法

株式会社ATR-Promotionsより用途に応じた条件によるライセンス(利用許諾)をいたします。
ホームページまたは下記よりご連絡ください。

株式会社 ATR-Promotions
https://www.atr-p.com

〒619-0288　京都府相楽郡精華町光台2-2-2
TEL：0774-95-1300　　FAX：0774-95-1191
E-mail：voice@atr-p.com