2026年03月13日 言語処理学会第32回年次大会 併設ワークショップ
日本語言語資源の構築と利用性の向上（JLR2026）

番組映像を活用したマルチモーダルデータセット開発の取り組み

NHK放送技術研究所
岡田拓也
遠藤伶, 中村純也, 田中大, 衣川和尭,
美野秀弥, 宮崎太郎, 河合吉彦

本取り組みの背景・目的

映像制作は労働集約型プロセスであり、その多くが制作現場の作業負荷に強く依存

AIを用いた映像制作作業の効率化

放送ドメインに特化した**マルチモーダル大規模言語モデル（MLLM）**の開発

そのためには大規模な日本語言語資源が必要
とくに映像・音声・テキストを含んだマルチモーダル資源の整備

放送済み映像アーカイブを**付加価値の高い言語資源**に変換

目的

- 指示学習用データセット
- ベンチマーク評価用データセット

の構築

内容

I. 放送済み番組映像データの整備

II. 指示学習データセットの構築

III. VQAベンチマークデータセットの構築

IV. モデル学習・評価

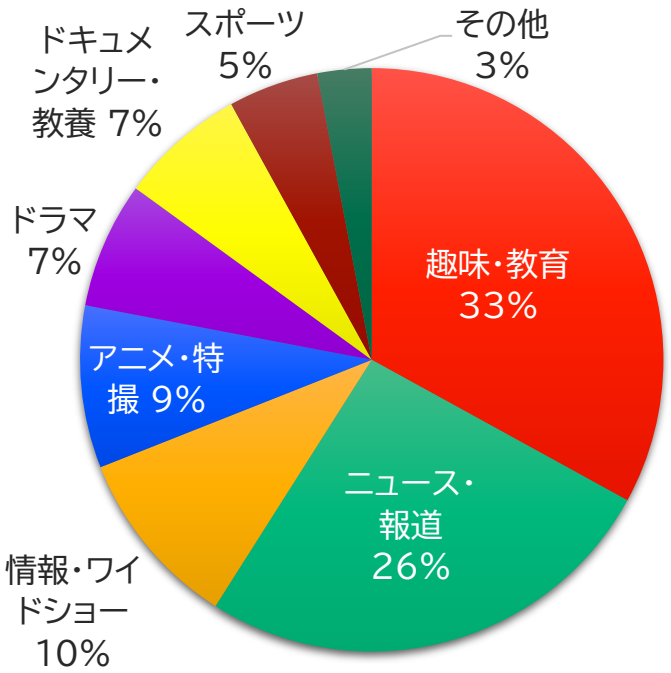
放送済み番組映像データの整備

番組映像データ整備の概要

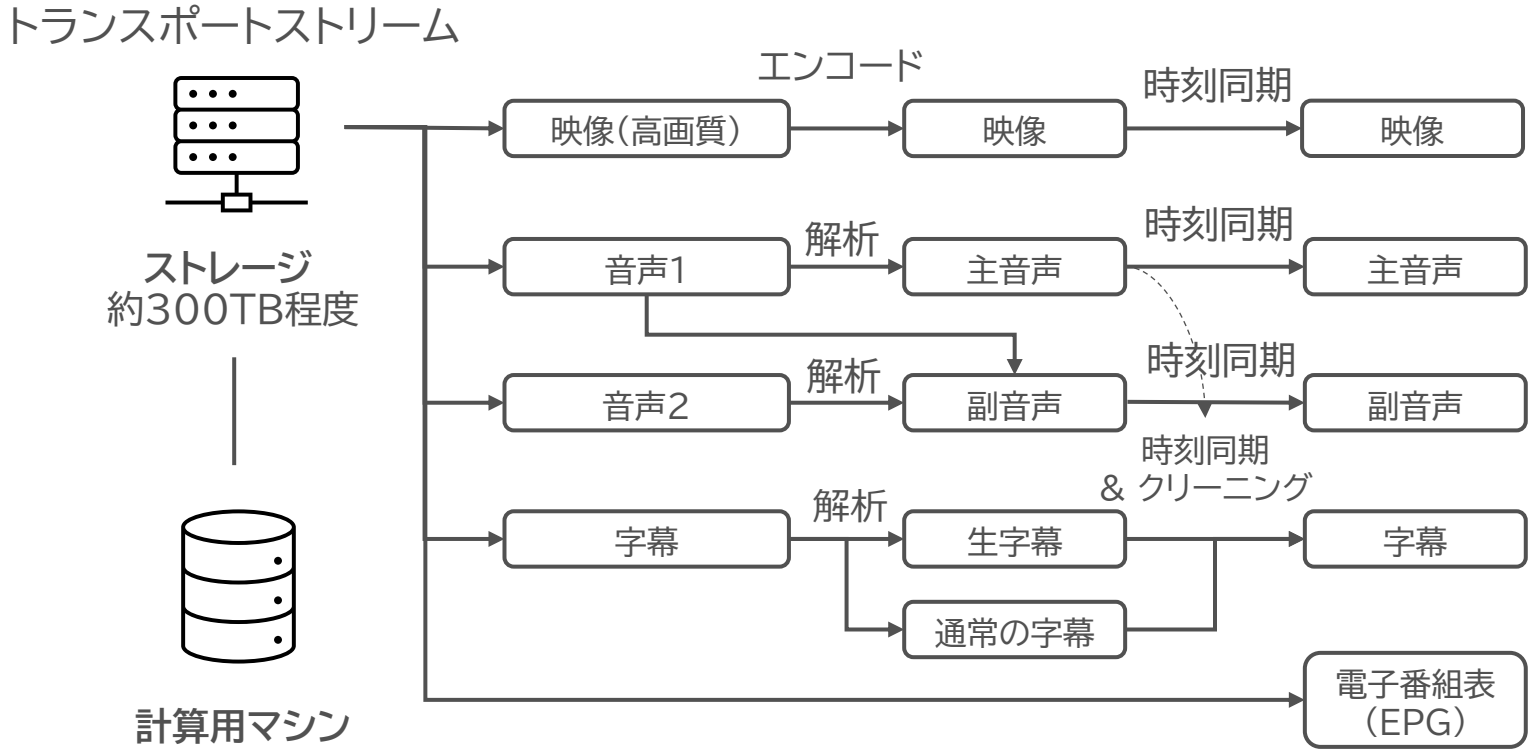
※ 2009年以降のNHK放送番組を対象
（15分以下の映像だけで総計2.3万時間）

- 放送録画データ※を展開して映像・音声・字幕・電子番組表を抽出
- 各モダリティのデータに対して圧縮・整形（文字情報の正規化など）
- 音声と映像の同期ずれや字幕の時間情報を調整

放送録画データ内訳



データ整備パイプライン



字幕と映像の時刻同期

- 字幕の表示時刻と映像内発話に0-5秒程度の誤差
- 音声認識モデルWhisper(独自に追加学習済み)を用いて時刻アライン
- 字幕の時刻情報を音声認識時刻に置き換え

※ cost: S_i と T_j の文字列の不一致度合

$$argmin_{\{(i,j)\}} \sum_{(i,j)} cost(S_i, T_j)$$

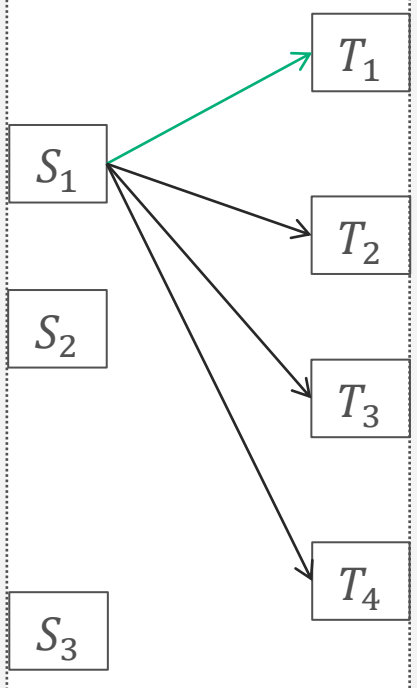
字幕の表示時刻
(正確なテキストだが時刻非同期)

00:53.440 --> 00:56.590
視聴者の方から質問が届きました。

00:56.590 --> 00:58.410
見ていきましょう、こちらです。

01:11.740 --> 01:13.460
後藤さん、いかがでしょうか。

各セグメントを編集距離でマッピング※



音声認識による書き起こし
(時刻同期だが不正確なテキスト)

00:51.440 --> 00:54.590
視聴者の方から質問が

00:54.590 --> 00:56.410
見ていきましょうこちら

00:57.100 --> 00.58.100
これは

01:10.740 --> 01:12.460
ごとうさん、いかがでしょうか

指示学習データセットの構築

開発中のデータセットファミリー

NHK

放送ドメインタスクを学習・評価するための日本語データセットファミリー

- タスク： 14種類（今後さらに改善・拡張予定）
- 学習サンプル：
  約20万QAペア（放送期間～2024/04/01）
- 評価サンプル：
  約4万QAペア（放送期間2024/04/01～）
- 作成手法： 自動／合成／人手
- 対象映像：
  15分以下の放送済み番組
  （整備済みデータの一部（重複あり））

	タスク名	サンプル数		作成方法
		学習	評価	
1	電子番組表の紹介文生成	24,267	3,943	自動
2	次カット予測	11,792	3,178	合成
3	次カット視覚推薦	21,044	4,208	合成
4	字幕翻訳品質推定	546	869	人手
5	電子番組表誤り検知	17,672	4,873	自動
6	演者区間推定	43	9	人手
7	カメラワーク推定	2155	1879	人手
8	会話追跡	33,207	5,165	合成
9	トピック分割	28,543	4,343	合成
10	TV映像OCR	5,152	5,000	人手
11	TV顔検出	19041	1915	人手
12	人物相関図生成	70	70	人手
13	固有表現抽出	11,226	2,867	合成
14	ドラマVQA	0	110	人手

No. 9 トピック分割タスクの紹介

NHK

ニュース番組を対象としたトピック分割タスク
→ ニュース映像を入力とし、映像中に含まれるトピックを時系列に分類※
※トピック区分：社会、政治、経済、国際、災害・気象、スポーツ、文化・芸術、科学・技術、その他
（ユースケースとしては、配信サービスに映像を展開する際の検索性向上などを想定）

入力映像

19時のニュースです。

スタジオ映像

…

北アルプスで山岳遭難が相次いでおり、今日未明、救助活動が行われました。

ロケ映像

…

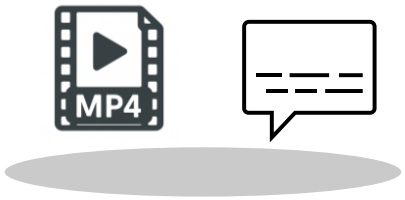
出力サンプル
（YAML形式）

```
topics:
- id: 1
  start: ”00:00:00”
  end: ”00:01:47”
  topic: 災害・気象
  title: 北アルプスで山岳遭難相次ぐ、2人死亡1人救助
- id: 2
  start: ”00:01:47”
  end: ”00:02:53”
  topic: 社会
  title: 母の日に向け青紫色カーネーションの出荷がピーク
```

トピック分割ラベルのデータ合成

正解ラベルを高速に作成するため映像と字幕を基にデータ合成

収録映像　時刻同期字幕



PySceneDetect を用いたシーン分割

カットの切り替わりを検出

各シーンの字幕テキストを抽出

シーンS_i
- 警察によりますと…

シーンS_{i+1}
- XX容疑者は調べに対し…

連続するS_i, S_{i+1}の字幕テキストから意味的に連続かを判定※

※判定指標

1. 意味的類似度
   Sentence Transformer
2. 時間的連続性
   （字幕間の時間ギャップが長ければペナルティ）

トピックT_i
- 警察によりますと…
- XX容疑者は調べに対し…

トピックT_{i+1}
- 関東甲信越は大気の状態が…

LLMを用いてトピックT_iを分類＆タイトル生成

トピックT_iの正解ラベル
- Start: 00:00:00
- End: 00:01:28
- 区分: 社会
- タイトル: 〇〇市で××

$$score(S_i, S_{i+1}) = wsim_{sem}(S_i, S_{i+1}) + (1-w)pen(\Delta t_i)$$

(S_i, S_{i+1})は連続 $\Leftrightarrow score(S_i, S_{i+1}) \geq \tau$

VQAベンチマークデータセットの構築

映像理解ベンチマーク

総合的な番組内容への理解度を評価するため2種類のベンチマークを構築
→ QAペアは人手で作成(1つの正解ラベルを含む4択問題)

01

連続テレビ小説『虎に翼』1,2話を対象としたドラマVQA
(全110問)

権利処理の問題で
掲載不可

02

ベンチマーク用独自制作映像を対象にした汎用番組VQA
(全190問)

ドキュメンタリー、バラエティ、教育、ドラマ


汎用番組VQAの例

映像内のアクションだけでなく、時系列構成、カメラワーク、テロップなど
映像制作に必要な情報に対する理解度を包括的に問う問題で構成

該当シーンのスクリーンショット

映像内で該当するテロップを見つける能力も必要

番組映像1本与えて以下の問題に解答させる想定

質問: テロップで登場する「デザイナー」の文字はどのような色設定(塗りと輪郭)になっていますか?

選択肢:
A. 白い塗りで黒い輪郭
B. 白い塗りで緑の輪郭
C. 赤い塗りで黒い輪郭
D. 黄色い塗りで紫の輪郭

正解: D

芸人さんのボケで存在しない釣り人のくだりに
(番組理解に必要なメタ認知能力?)

質問: 藍染川にいた釣り人は何人ですか?

選択肢:
A. 1人
B. 2人
C. 3人
D. 0人

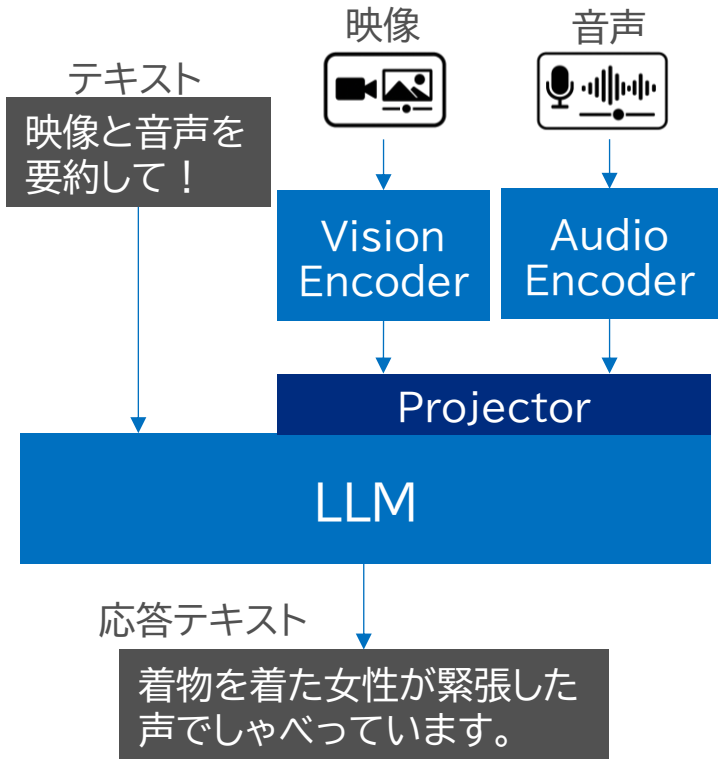
正解: D

マルチモーダルLLMの追加学習・評価

追加学習モデルの試作

構築したデータセットを用いてテキスト・映像・音声を統合的に処理できるマルチモーダルLLMを追加学習

モデル構造の概略図



学習条件

- ベースモデル: Qwen3-Omni-30B-A3B
- タスク: SFT(Full FT+LoRA rank=8 alpha=32)
- フレームワーク: ms-swift + Megatron-LM
- 計算環境: A100 GPU×16枚
- 精度: bfloat16
- バッチサイズ: MBS=2, GBS=64, accum=2
- 系列長: 32768
- 最大ピクセル数: 17920000
- LoRA対象: 全線形層
- 並列設定: EP=4, TP=4
- バックエンド: NCCL, FlashAttention

初期実験の結果

スコア：0～1 で1が最大、 背景あり:学習済タスク、 青字 👑 :最高スコア

タスク名	Qwen3-Omni（ベースモデル）	独自モデルv1（5タスク学習）	独自モデルv2（9タスク学習）
電子番組表紹介文生成	0.01	**0.40** 👑	0.36
次カット予測	0.08	0.08	**0.19** 👑
次カット視覚推薦	0.09	0.07	**0.22** 👑
字幕翻訳品質推定	**0.10** 👑	0.08	0.07
電子番組表誤り検知	0.44	0.75	**0.77** 👑
演者区間推定	0.00	0.02	**0.36** 👑
会話追跡	0.05	0.57	**0.63** 👑
トピック分割	0.06	0.04	**0.35** 👑
TV映像OCR	0.31	0.22	**0.69** 👑
人物相関図生成	**0.27** 👑	0.23	0.20
ドラマVQA	0.58	0.63	**0.70** 👑

基本的には学習によって性能向上

→ 一部タスクでは性能低下

- 字幕翻訳品質推定
- 人物相関図生成

データ品質・難易度設定に一部課題

学習していないタスクでも性能向上

- 演者区間推定
- ドラマVQA

汎用的な映像理解能力を獲得

まとめ

- 放送済み映像アーカイブを整備してマルチモーダルデータセットを構築
  - 指示学習データセット
    - 動画編集支援・ガードレールを想定した放送ドメインタスク
  - VQAベンチマークデータセット
    - 連続エピソードドラマ（朝ドラ）を対象としたVQA
    - ベンチマーク用独自制作映像を対象にした汎用番組VQA
- 追加学習モデルの試作・評価
  - 学習タスクに対する性能はほとんどのデータで向上
  - 一部データセットの品質・難易度設定については改善が必要
  - 汎用的な映像理解能力の向上を確認

<u>データセットファミリーは今後さらに改良・拡張していく予定</u>