進捗報告

1　進捗

なし

2　やること

パソコンセットアップしなおし. 今後の研究テーマについて相談. 昨年度岡田先生にプレゼンした創作支援のための動画を入力としたサウンドエフェクトの自動生成の話も少し実現可能性について話したいです.

3　興味

先週, 岡田先生から頂いた自然言語周りの直近の学会資料からいくつか読んで, 興味を持ったものに関して簡単にまとめる. bibtex が上手く動かなくて, 参考文献は後ろに直貼りしておきます.

3.1　動画キーフレーム物語生成手法の提案

入力動画データの中から指定した「キーフレーム」と, それに対応する「説明文」を用いて, 絵コンテのように動画を中身を瞬時に把握可能な要約を生成することを目標とした研究で, 動画キーフレーム物語生成タスクに対してベースラインとなる手法の提案として, 既存の動画要約データセットである ActivityNet Captions を拡張したデータセットを用いて, 教師あり学習に基づくベースラインを構築し, 性能評価及び分析を行っている. 図 1 にモデルの概略図を示す.

3.1.1　ActivityNet Captions

各動画は M 個の説明文とそれぞれの説明文に対応する K 個のキーフレームについてのアノテーションを持つ. キーフレームの最小単位は 0.5s となっている.

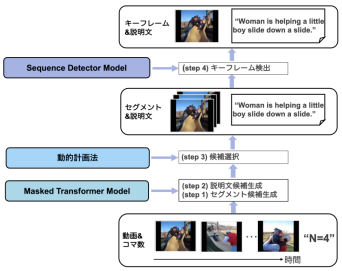

図 1: ベースラインとなるシステムの推論・生成モデル概略図

3.2　RINA: マルチモーダル情報を利用したキャラクターの感情推定

この研究では, ゲームのシナリオスクリプトを題材として, テキストを主体とした対象文非テキストの参照情報を結合することにより, キャラクターの感情推定の精度を向上させることを試み, その際複数の結合方法の効果について実験的に調査している. アーキテクチャに関しては, この研究で提案している Reference Information Normalize Adapter (以下 RINA という) モデルを用いて, テキスト情報の分散表現ベクトルに対し参照情報による補正を行うことで, 従来手法より推定精度の向上を確認している.

3.2.1　RINA

VilBERT や VLBERT などのモデルは, BERT の構造をそのまま流用し, 非テキストの画像情報を入力の追加トークンとして利用しているが, これらのモデルのタスクは本来テキスト情報のみ利用した BERT の事前学習タスクと大きく異なるため, 画像とテキストのマルチモーダルデータを用いて改めて事前学習が必要である. よってこの研究では, テキスト情報と非テキストの結合は Transformer モデル出力の後に行い (BERT など Transformer の内部構造に依存しない特性を重視), 非テキスト情報を考慮しテキストの分散表現ベクトルを補正する. 図 2 にモデル概略図を示す. 非テキスト情報としては, キャ

ラクターの名前, 性別などの属性や, 対象のセリフ文が発生するシーンの ID 情報を含め計 9 種類を用いている.

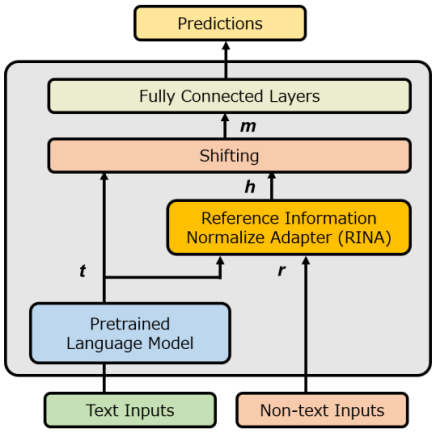

図 2: RINA モデル概略図

4 参考文献

@article{E5-3,
  title={動画キーフレーム物語生成手法の提案},
  author={佐藤俊, 佐藤汰亮, 鈴木潤, 清水伸幸, 東北大学, ヤフー株式会社},
  journal={言語処理学会 第 28 回年次大会},
  year={2022},
}

@article{A6-2,
  title={RINA: マルチモーダル情報を利用したキャラクターの感情推定},
  author={頼 展韜, 高橋 誠史, 株式会社バンダイナムコ研究所},
  journal={言語処理学会 第 28 回年次大会},
  year={2022},
}