マルチモーダル情報の対応関係を捉えたモダリティ変換

濵園　侑美

人は自然言語を介して物事をより深く理解する．
近年，計算機の発達により，言葉や画像，数値情報などを扱う手法が発達し，その能力が向上している．さまざまな物事を計算機上で扱えるようになった一方，それぞれの物事はモダリティ毎に独立して扱われていたため，モダリティ間の関係性を捉えるに至っていない．そこで，本論文は言葉を中核に据えて，異なるモダリティの対応関係を捉え，生成する課題に取り組む．

一つ目は言語から非言語系列を生成する課題として，自然言語からロボット動作を生成する課題に取り組む．言語を計算機上で扱う手法として，意味の捉え方に文章中の単語の頻度情報や周辺単語による単語の類似度により自動的に意味を付与する分散意味表現を用いる．非言語系列として，本課題ではロボット動作を扱う．まず，人間の関節の動きとは異なるロボットに動作を生成させるためにロボットが行える基本動作を組み合わせることにより，人の動きを真似て行える枠組みを提案する．また，それらの基本動作を組み合わせることにより，複雑な動作を表現，生成することを可能とする．これらを達成したのち，言葉と動作の対応関係を学習する課題に取り組む．2 つの対応関係を捉える方法として，本課題ではニューラルネットワークを用いる．特に，多様なロボット動作と曖昧な表現との対応関係を学習できるようなニューラルネットワークの構造を提案し，検証する．

二つ目は非言語から言語を生成する課題として，一般ドメインの動画の実況生成課題に取り組む．本課題に則したデータセットは存在しないため，既存のビデオコーパスを用いて実況の収集を行い，データセットの構築を行う．その後，課題をタイミング推定と発話生成の 2 つのサブタスクに分割して，動画のみから実況を生成する課題の解決に取り組む．これは，実況は映像に合わせて出力されるテキストのため，実況を開始するタイミングや，実況テキストの長さを制御する必要があるためである．タイミング推定には，人の行動を推定できる枠組みを用いる．発話生成には，言語生成に適したニューラルネットワークモデル，特に Transformer を用いる．また，クローズドメインの実況生成を行なった先行研究と比較実験し，実況生成に必要な要素について検討する．

最後に，非言語から言語を生成する課題のうち，実世界で得られたデータとテキストを用いる場合に発生する問題について取り上げ，問題の原因を分析し，解決策を提案する．特に，時系列数値データの概況テキスト生成における問題点を取り上げる．既存の時系列数値データの概況テキスト生成に用いるデータセットには，時系列データを用いた data-to-text に特有の参照時刻の不整合問題と，実世界のデータとテキストを用いた場合に起こりうる，入力データから予測できない属性を含む出力を求められる問題が含まれている．データセットの作成手順や入出力データの特性を理解することで，既存の枠組みを大きく変えることなく，これらの問題を解決する方法を提案する．

以上三つの取り組みにより，言葉を中心としてモダリティーの方向性を変更する手法の提案と，方向性を示す．