視覚情報に基づくマルチモーダル機械翻訳手法に関する研究

RESEARCH ON MULTIMODAL MACHINE TRANSLATION METHOD BASED ON VISUAL INFORMATION

呉　岩峰
指導教員　亀田　弘之

東京工科大学　大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻　亀田研究室

キーワード：自然言語処理、機械学習

１．初めに

視覚情報に基づくマルチモーダル機械翻訳とは、機械翻訳システムのパフォーマンスを向上させるために、テキスト機械翻訳をベースに、モデルの文脈理解を助ける補助として画像やビデオ情報を使用することである。 一般的な方法は、2 つの異なるモダリティの情報をエンコード時に融合することである。 本稿では、視覚情報に基づくマルチモーダル機械翻訳のサブタスクであるテキスト-画像機械翻訳について研究し、この分野における問題点を分析する。すなわち、画像情報にはテキストと無関係な内容が含まれており、冗長な画像情報は翻訳システムに影響を与える。 以上の問題点を解決するために、本稿では関連する研究を行う。

２．研究状況と分析

2017 年には Transformer が登場し、わずか数年で機械翻訳、さらには自然言語処理分野全体の研究方向をリードした。 代表的な作品は、マルチモーダル Transformer（MM Transformer）［1］と、カプセルネットワークを用いて視覚情報を抽出する DCCN ネットワーク［2］である。MM Transformer のエンコーダ入力は、アテンション・メカニズムによって得られた知覚テキストの画像表現であり、最終出力はエンコーダの出力をテキスト・ベクトルとスプライシングすることによって得られる。デコーダ部分は変更されず、ネイティブの Transformer デコーダ構造が維持される。 一方、DCCN ネットワークは、翻訳品質を向上させるために、ソース言語の文脈情報をグローバルおよびローカルの視覚情報と融合させ、マルチモーダルな文脈表現を得る。

画像の単語とテキストの単語には対応関係があるため、異なる言語の画像の同じ領域に対応する単語は同義であるべきであり、視覚情報をより有効に利用するために、ソース言語とターゲット言語の同じ意味論に基づく単語を画像の同じ領域に整列させる研究もある。 そこで、2 段階のデコーディング・モデルを提案する学者もいる［3］。プレーン・テキスト・モデルのデコーディング出力を得た後、画像情報を使って翻訳結果をさらに洗練させるのである。 これにより、テキスト情報をフルに活用できるだけでなく、視覚情報が必要なシーンでは、画像情報の助けを借りることもできる。

現在、テキスト-画像マルチモーダル機械翻訳に関する国内外の研究は非常に詳細かつ堅実である。本稿では、Transformer のモデルフレームワークに基づき、画像のテキストに依存しないノイズの扱いから始め、新規な手法を提案し、モデルの有効性をサポートするための理論的分析と実例分析を実施する。

しかし、これらの方法はいずれも、テキストにおいて補助的な役割を持つ視覚モダリティの情報を単純かつ効果的に抽出し、無駄な情報を破棄する

ことで、冗長な情報を導入することなく翻訳効果を向上させることはできない。 テキスト-画像マルチモーダル機械翻訳において、画像の役割は、テキストの文脈理解を支援することである。 しかし、画像によっては、テキストに依存しない情報が大量に存在し、画像の 3 分の 2 以上が画像の意味に影響を与えることなく削除され、さらにテキスト全体を完全に解釈するために一部しか残されていない画像もあり、これは画像によっては無駄なノイズ情報が大量に存在することを示している。 このようなテキストに依存しない画像ノイズの存在は、機械翻訳モデルに何らかの摂動を与える可能性があり、その結果、翻訳結果に影響を与えたり、誤った翻訳結果を生成するようにモデルを惑わしたりすることさえあります。

3. 研究方法

テキスト-画像機械翻訳タスクにおける通常のアプローチは、画像とテキストを融合してエンコーダに入力することである。しかし、画像にはテキストに依存しない情報が大量に存在するため、テキストに依存しないノイズ情報を除去し、翻訳を支援しながらモデルへの干渉を低減する手法を設計する必要がある。

この問題に対して提案する選択的注意は、冗長な情報が翻訳モデルに与える影響を最大限に回避するために、従来の特徴量の重み付き和の代わりに、画像特徴量の重要な部分を選択するために適用される。選択方法は Gumbel-Sigmoid の選択器を基に改良され、本稿で提案する選択的注意を形成する。この選択的注意機構は、画像特徴のうちテキストに関連する部分を自動的に選択する。

Transformer に基づく改良されたフレームワークは、視覚情報とテキスト情報を有機的に結合し、マルチモーダルな注意メカニズムを介して知覚テキストの視覚表現を生成する。 また、テキスト表現と知覚テキストの視覚表現は、独立したエンコーダを用いて符号化される。この方法がベースラインモデルより優れている点は、主にエンコーダにある。 視覚情報を合理的に利用するために、本稿では知覚テキストの視覚表現とマルチモーダルゲーティングネットワークを導入する。

知覚テキストの視覚表現は、視覚情報をテキスト情報に整列させ、テキスト情報と視覚情報の類似スコアを計算して注目行列を得、さらに注目行列に基づいて視覚表現の重みを再配分する。また、マルチモーダル機械翻訳において視覚情報をより合理的に適用させるためには、ゲーティングネットワークを設計する必要がある。 これは、マルチモーダル情報全体における視覚情報の割合を制御し、視覚情報の重みを動的に割り当てることで、特定のシーンではマルチモーダル情報をフルに活用し、一般的なシーンでは情報の冗長性を減らすという目的を達成することができる。

本稿で使用するデータセットは、Flickr30K[4] データセットを拡張した Multi30K である。Multi30K データセットには、29,000 組の学習データ、1014 組の検証データ、1,000 組のテストデータがある。また本稿では、テキスト画像のマルチモーダル機械翻訳における画像データに対して、VGG19 事前学習済みネットワークを用いて視覚的特徴を抽出する。

今後の研究内容は、Gumbel-Sigmoid 方法に基づくサンプリング処理により、画像特徴ノイズを低減する新しい手法とマルチモーダルゲーティングネットワークを設計である。

参考文献

[1] Yao S, Wan X. Multimodal transformer for multimodal machine translation[C]
[2] Lin H, Meng F, Dynamic Context-guided Capsule Network for Multimodal Machine Translation[C]
[3] Ive J, Madhyastha P S, Specia L. Distilling Translations with Visual Awareness[C]
[4] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]