マルチモーダルニューラル機械翻訳のための教師付き視覚的注意の研究

Studies on Supervised Visual Attention for Multimodal Neural Machine Translation

愛媛大学大学院理工学研究科教授
二宮　崇

2001 年東京大学大学院理学系研究科情報科学専攻博士課程修了。博士（理学）。2017 年より愛媛大学大学院理工学研究科教授。自然言語処理の研究に従事。

愛媛大学大学院理工学研究科電子情報工学専攻
西原　哲郎

2020 年愛媛大学工学部情報工学科卒業。現在、愛媛大学大学院理工学研究科電子情報工学専攻博士前期課程在学中。自然言語処理の研究に従事。

同志社大学理工学部情報システムデザイン学科准教授
田村　晃裕

2013 年東京工業大学大学院総合理工学研究科博士課程修了。博士（工学）。2020 年より同志社大学理工学部情報システムデザイン学科准教授。自然言語処理の研究に従事。

愛媛大学大学院理工学研究科電子情報工学専攻
表　悠太郎

2019 年愛媛大学工学部情報工学科卒業。現在、愛媛大学大学院理工学研究科電子情報工学専攻博士前期課程在学中。自然言語処理の研究に従事。

東京大学大学院情報理工学系研究科准教授
中山　英樹

2011 年東京大学大学院情報理工学系研究科知能機械情報学専攻博士課程修了。博士（情報理工学）。2018 年より東京大学大学院情報理工学系研究科准教授。画像認識と自然言語処理の研究に従事。

1 はじめに

現在、ニューラルネットワークを用いた機械翻訳（ニューラル機械翻訳）が機械翻訳の主流となっている。注意機構を用いた再帰型ニューラルネットワーク（Recurrent Neural Network; RNN）に基づくニューラル機械翻訳モデルは初期のころから広く使用されてきたモデルである[1]。このモデルは、原言語文（翻訳元

言語の文）内の単語と目的言語文（翻訳先言語の文）内の単語間の関係を捉える言語間注意機構を用いることで、従来のRNNベースのニューラル機械翻訳よりも高い精度を実現した。また、近年、トランスフォーマーモデル[2]がRNNや畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を用いた手法と比べて高い精度を達成し、注目されている。トランスフォーマーモデルは、従来の言語間注意機構に加えて、同じ文中の単語間の関係を捉える自己注意機構を導入している。

ニューラル機械翻訳の性能を改善する手法については様々な研究がなされているが、その内の一つに、上述の言語間注意機構に制約を与える研究がある[3][4][5]。これらの研究では、アライメントツールを用いて原言語文と目的言語文間の単語の対応関係を予め取得し、その対応関係を教師データとして与えて言語間注意機構を学習させることで翻訳性能の向上を実現している。

機械翻訳手法の一つとして、原言語文に加えて画像を入力することで翻訳性能の改善を目指すマルチモーダルニューラル機械翻訳[6]がある。入力画像は、翻訳時の曖昧性解消や省略補完の手がかりとして役立つと考えられ、例えば、特許出願書類の明細書に付随する図を参照することでより質の高い特許翻訳が実現されることが期待される。マルチモーダルニューラル機械翻訳のモデルとして、Helcl ら[7]は、CNNによって抽出した画像の特徴量を翻訳に活用するために、文中の単語と画像の領域との対応関係を捉える視覚的注意機構をトランスフォーマーモデルのデコーダ内に導入したモデルを提案している。また、Delbrouck ら[8]は、RNNベースのマルチモーダルニューラル機械翻訳モデルのエンコーダに視覚的注意機構を導入したモデルを提案している。しかし、これらの視覚的注意機構は、マルチモーダルニューラル機械翻訳の訓練時に教師なしで自動的に学習が行われている。そのため、本来捉えるべき対応関係を常に捉えられるとは限らない。

本稿は、我々が行っているマルチモーダル機械翻訳の研究について紹介する。この研究では、マルチモーダルニューラル機械翻訳の性能改善のために、人手により与えられた文中の単語と画像領域との対応関係に基づいて教師付き学習を行う制約付き視覚的注意機構を提案している。従来の視覚的注意機構は、ある単語が注目すべき画像内の領域は教師として与えられず、マルチモーダルニューラル機械翻訳の学習を通じて自動的に学習されている。そこで、本研究では、原言語文中の単語と画像内のオブジェクトとの対応関係を示したデータを教師データとして用意し、その教師データを用いてトランスフォーマーモデルのエンコーダ内で視覚的注意機構を直接学習させる。

2 背景

本章では、研究背景として最初にトランスフォーマーモデルの概要を述べる。次に、トランスフォーマーモデルにおける制約付き言語間注意機構について説明する。

2.1 トランスフォーマーモデル

トランスフォーマーモデルは、原言語文を受け取って中間表現に変換するエンコーダと、その中間表現を受け取って目的言語文を生成するデコーダから構成されている。エンコーダとデコーダはそれぞれエンコーダレイヤとデコーダレイヤを複数スタックした構成となっている。各エンコーダレイヤは自己注意機構と位置毎の全結合層の2つのサブレイヤを持っている。これらのサブレイヤ間では、残差接続[11]と層の正規化[12]が用いられる。

自己注意機構と言語間注意機構（Att）は以下の式で表される。

$$\mathrm{Att}(Q, K, V) = AV$$
$$A = \mathrm{softmax}(QK^{T} / d^{0.5})$$

ここで、Aは注意行列と呼ばれる。また、Q, K, Vはエンコーダ及びデコーダにおける隠れ状態を表し、dはQ, K, Vの次元数を表す。上式のQ, K, Vが前のサブレイヤから与えられた場合は自己注意機構となる。また、Qがデコーダ内の前のサブレイヤから、KとVがエンコーダの出力から与えられる場合は言語間注意機構となる。自己注意機構では同文中の単語間の関係を計算することができる。また、言語間注意機構では原言語文内の単語と目的言語文内の単語間の関係を計算することができる。

トランスフォーマーの特徴として、隠れ状態を部分空間に分割し、各部分空間において様々な情報を表現する

ことを可能とした複数ヘッドの注意機構がある。h 個のヘッドからなる複数ヘッドの注意機構（MHA）は以下のように表される。

$$MHA(Q,K,V)=[head_1;\cdots;head_h]\ W^O$$
$$head_i=Att(QW^Q, KW^K, VW^V)$$

ここで、[ ] はベクトルを結合することを表している。W^Q, W^K, W^V はそれぞれヘッド毎に定義されるパラメータ行列であり、d 次元ベクトルを線形変換により、d／h 次元に縮退させ、d／h 次元ベクトルを各ヘッドに渡す。W^O はパラメータ行列であり、結合された各ヘッドの出力に対し線形変換を行う。

位置毎の全結合層（FFN）は以下のように表される。

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2$$

ここで、W_1, W_2, b_1, b_2 はパラメータ行列である。

トランスフォーマーは隠れ状態に語順の情報を組み込むために位置エンコーディングが導入されている。この位置エンコーディングを単語の埋め込み表現に加えることで、単語の語順の情報を付与することができる。

2.2　制約付き言語間注意機構

Garg ら[5] は、トランスフォーマーモデルの言語間注意機構に原言語と目的言語間の単語の対応関係を教師として与えて学習を行う手法を提案している。アライメントツールを用いて言語間の対応関係を取得し、複数ヘッドの言語間注意機構のある 1 つのヘッドとの間で計算される誤差を最小化することによって注意機構の学習を行う。誤差の計算は以下のような交差エントロピーによって行われる。

$$L_a(A)=-M^{-1}\Sigma_{m=1,\cdots,M}\Sigma_{n=1,\cdots,N}\ G_{m,n}\log(A_{m,n})$$

ここで、M は目的言語文の文長、N は原言語文の文長、A は言語間注意機構により計算される注意行列、G は教師となる単語の対応関係を表した行列である。なお、n 番目の原言語文の単語と m 番目の目的言語文の単語が対応関係にある場合は、$G_{m,n}$ は 1 となり、それ以外は 0 となる。この機械翻訳モデルの目的関数 L は、上述の $L_a(A)$ を翻訳の誤差 L_t に加えて以下のように表される。

$$L=L_t+\lambda L_a(A)$$

ここで、λ はハイパーパラメータである。

3　マルチモーダル機械翻訳のための視覚的注意機構の教師付き学習

この章では、マルチモーダルニューラル機械翻訳の翻訳性能を向上させるための視覚的注意機構の教師付き学習について説明する。まず、トランスフォーマーベースのマルチモーダルニューラル機械翻訳モデルについて説明する。次に、我々の提案手法である視覚的注意の教師付き学習について述べる。

3.1　トランスフォーマーベースのマルチモーダルニューラル機械翻訳モデル

図 1 にトランスフォーマーベースのマルチモーダルニューラル機械翻訳モデルの概要図を示す。本モデルは、原言語文エンコーダとデコーダに加えて、画像エンコーダを持つ。画像エンコーダでは、入力した画像に対して CNN を適用し、画像の特徴量を得る。次に、CNN の出力に対して自己注意機構を適用する。この自己注意機構によって画像の領域間の関係性を考慮する。最後に、

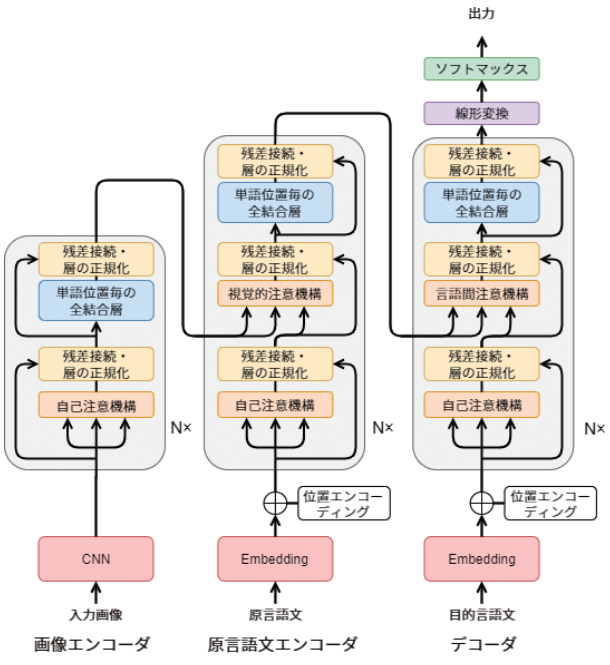


図 1　マルチモーダルニューラル機械翻訳モデル

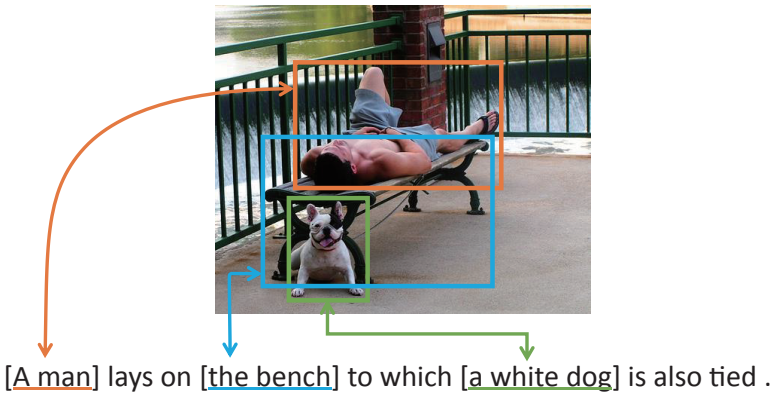


図２　Flickr30k entities の例

自己注意機構の出力に対して位置毎の全結合層を適用したものが画像エンコーダ全体の出力となる。

次に、原言語文エンコーダの内部で、原言語文に対する自己注意機構の出力と画像エンコーダの出力を用いて視覚的注意機構[13]を計算する。視覚的注意機構は注意機構の一種で、画像の領域と単語との関係性を計算する。前述の注意機構の式において、Q が原言語文エンコーダの自己注意機構の出力、K と V が画像エンコーダの出力となった際に視覚的注意機構となる。

単語 man と単語 lamp に対する視覚的注意の例をそれぞれ図５(c)と図６(c)に示す。図において、暗くなっている部分により強く注意が向けられていることを表している。入力された画像は、CNN によって高次元の目の粗い格子状の特徴量へと変換される。特徴量中の各領域は、元の画像の領域に対応しており、高次元の特徴量を持っている。例えば、図５(c)と図６(c)は７×７の領域に変換されていて、各領域が 2,048 次元の特徴量を保有している。視覚的注意機構では、各単語はこれらの画像特徴量の領域に対して注意が向けられる。このため、視覚的注意は、領域のヒートマップとして可視化することができる。

3.2　視覚的注意機構に対する制約

提案手法では、視覚的注意機構に対して原言語文の単語と対応する画像内のオブジェクトとの対応関係を人手で付けたものを教師データとして与えて制約を加えることによって学習を行う。具体的には、画像エンコーダの出力と原言語文エンコーダ内の自己注意機構の出力との間の視覚的注意機構に制約を与える。

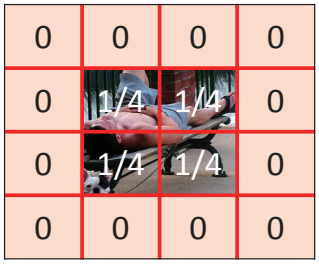
視覚的注意機構に対する制約は、教師となる対応関係を示した行列と注意行列との間の誤差が最小となるように適用される。誤差は以下のような交差エントロピーによって計算される。

$$L_{img\_src}(A) = -M^{-1} \Sigma_{m=1,\cdots,M} \Sigma_{n=1,\cdots,N} G_{m,n} \log(A_{m,n})$$

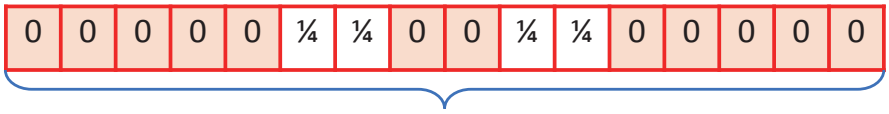
ここで、M は原言語文の文長を、N は CNN によって畳み込まれた画像の領域数を表す。また、A は注意行列を、G は教師となる対応関係を示した行列を表す。原言語文内の m 番目の単語が画像の n 番目の領域に対応しているとき、$G_{m,n}$ は 1 となり、それ以外の時は 0 となる。

本研究では、Flickr30k entities データセット[14]を用いて視覚的注意機構に対する制約を作成した。このデータセットは Flickr30k データセット[15]から作られたデータセットである。一つの画像に対して５つのキャプション文がつけられており、各キャプション文中の単語が画像内のオブジェクトと関係がある場合、図２のようにその単語が画像内のどの領域と関係があるか示されたデータセットとなっている。

今回はこのデータセットから原言語文内の単語と画像内のオブジェクトとの対応関係を抽出し、制約を作成する。まず、Flickr30k entities データセットに付与されている単語とオブジェクト間の対応関係を CNN で畳み込んだ際の領域にスケールさせる。例えば、画像エンコーダに用いる CNN によって画像を４×４に畳み込んだ場合、画像内の各オブジェクトと 16 つの領域との対応関係を求める。複数の領域に対応する場合は、各領域に等しく対応が張られるように値を平均化する。すなわち、値を「1/ 対応付いた領域数」とする（図３(a)）。その後、２次元の領域を１次元に線形化する（図３(b)）。この工程を原言語文のすべての単語に対して行う。



(a) 単語manに対する制約の例



(b) 線形化

図３　視覚的注意機構に対する制約の作成例

画像内のオブジェクトと対応がない単語については、Liu ら[16] や Mi ら[17] に倣い、特殊トークンを用いて処理を行った。Liu らや Mi らは言語間注意機構に対する制約を与える際に特殊トークンを導入している。本手法では、図３(b)に示した行列の先頭に特殊トークンを付与し、画像内のオブジェクトと対応関係にない単語をその特殊トークンに対して対応付けする。

視覚的注意機構に対する制約を与えたマルチモーダルニューラル機械翻訳モデルの目的関数 L は以下のように表される。

$$L = L_T + \lambda_1 L_{img\_src}$$

ここで、L_T はマルチモーダルニューラル機械翻訳モデルの損失関数を、L_{img_src} は視覚的注意機構における注意行列と制約行列との損失関数を表している。また、λ_1 は翻訳誤差 L_T と制約付き視覚的注意機構の誤差 L_{img_src} との間の重みを制御するためのハイパーパラメータである。

3.3　言語間注意機構に対する制約

本研究では、2.2 節で説明した制約付きの言語間注意機構をマルチモーダルニューラル機械翻訳モデルに導入し、翻訳性能の改善を図る。言語間注意機構に対して制約を与えるためには、原言語文内の単語と目的言語文内の単語との対応関係を取得する必要がある。今回はこれを取得するために Liu ら[16] や Garg ら[5] のように、アライメントツールを用いる。アライメントツールとしては、GIZA++[18] をマルチスレッドで動作させることを可能とした MGIZA[19] を用いた。

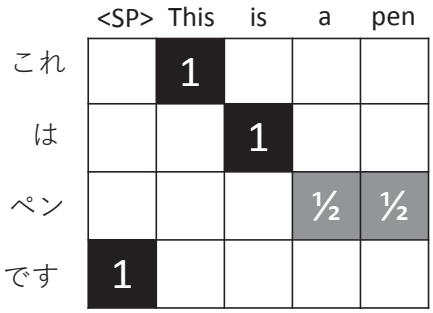


図４　言語間注意機構に対する制約の例

目的言語文内のある１単語が複数の原言語文内の単語と対応関係にある場合、等しく対応が張られるように値を平均化する。すなわち、値を「1/ 対応付いた単語数」とする。また、3.2 節のように、対応関係がない単語については、特殊トークンを用いて処理を行った。この方法では、図４に示すように原言語文の先頭に特殊トークンを設置し、対応関係を持たない目的言語文内の単語はこの特殊トークンに対応が張られるようにする。

視覚的注意機構と言語間注意機構の両方に制約を加えたマルチモーダルニューラル機械翻訳モデルの目的関数 L は次式のように表される。

$$L = L_T + \lambda_1 L_{img\_src} + \lambda_2 L_{src\_tgt}$$

ここで、L_{src_tgt} は言語間注意機構における注意行列と制約行列との損失関数を表している。また、λ_1 と λ_2 は翻訳誤差 L_T と視覚的注意機構による誤差 L_{img_src}、言語間注意機構による誤差 L_{src_tgt} との重みを制御するハイ

表1　実験結果。B と M はそれぞれ BLEU と METEOR を表す。

	英→独		独→英		英→日		日→英	
	B	M	B	M	B	M	B	M
NMT	38.76	57.59	42.58	39.19	43.69	59.27	44.21	40.03
MNMT	38.89	57.35	42.29	39.13	44.09	59.59	44.42	40.03
MNMT+SVA	39.91	58.11	42.52	38.86	44.51	60.03	44.76	40.40
MNMT+SVA+SCA	40.50	59.05	43.76	39.71	44.79	60.23	45.36	40.65

パーパラメータである。

4 実験

本研究では、英独翻訳、独英翻訳、英日翻訳、日英翻訳の 4 つの実験を行った。英独翻訳および独英翻訳では、Multi30k データセット[20] を用いた。このデータセットは、画像とその説明文が対になったもので、訓練データは 29,000 文対、開発データは 1,014 文対である。また、テストデータとしては test2016 を用いたため、1,000 文対である。

英日翻訳および日英翻訳では、我々が開発した Flickr30k Entities JP データセット[10] を用いた。Multi30k データセットには日本語文が含まれていなかったが、英文を人手で日本語に翻訳することで英日マルチモーダル対訳データを開発した。画像内オブジェクトと対応付けられた英語句に対して、対応する日本語句に対応付けを与えていることも特徴としている。Multi30k データセットのテキストの前処理に倣い、英文には小文字化、句読点の正規化、Moses のトーカナイザ[21] を施している。日本語文については KyTea[22] を用いて単語分割を行った。また、訓練データには日英共に 100 単語以下の対訳文のみを用いた。訓練データは 59,516 文対、開発データは 2,017 文対、テストデータは 2,000 文対である。

画像に対する前処理として、画像サイズを 256 × 256 になるようにリサイズした後、224 × 224 となるように中央部にクロップ処理を施した。画像エンコーダにおいて使用する CNN は ResNet50[11] を用いた。なお、ResNet50 から取得する画像特徴量は最終の畳み込み層の出力を用いた。したがって、抽出される画像特徴量のサイズは 7 × 7 × 2048 である。また、学習時に CNN の fine-tuning は行わない。画像エンコーダ、原言語文エンコーダおよびデコーダレイヤはそれぞれ 6 層から成る。複数ヘッドの注意機構におけるヘッド数は 8、埋め込み次元数は 512 とした。モデルの学習時にはミニバッチサイズを 128 とし、40 エポックの学習を行った。最適化手法には Adam[23] を用いた。英独および独英実験では BPE を適用した。推論時には目的言語文の生成を貪欲法により行った。

翻訳性能は BLEU と METEOR を用いて評価した。テスト時には開発データに対する BLEU 値が最も高かったエポックのモデルを選択し、テストデータに対する性能を評価した。実験では、画像無しのトランスフォーマーモデル（NMT）、制約を与えていないマルチモーダルトランスフォーマーモデル（MNMT）、視覚的注意機構にのみ制約を与えたモデル（MNMT+SVA）、視覚的注意機構と言語間注意機構の両方に制約を加えたモデル（MNMT+SVA+SCA）を比較した。制約を加えた言語間注意機構は、2.2 節で説明した Garg らのものと同様である。また、6 層スタックされた原言語文エンコーダレイヤおよびデコーダレイヤの内、6 層目の注意機構に対して制約を与える。さらに、視覚的注意機構と言語間注意機構の一つのヘッドに対して制約を与える。なお、3.3 節で説明した目的関数 L におけるハイパーパラメータは Garg らに倣い、λ_1=0.05、λ_2=0.05 とした。また、視覚的注意機構にのみ制約を与えたモデルのハイパーパラメータは λ_1=0.05 とした。本実験で利用したデータセットには、単語と画像内のオブジェクトとの対応は英語の文章にのみ与えられている。英独翻訳と英日翻訳については、人手によって付けられた対応関係を直接利用して、視覚的注意機構に対する制約を作成した。独英翻訳と日英翻訳については、初めに英語の文に人手でつけられている対応関係を MGIZA によって得た言語間の対応関係を用いて独語および日本語との対応関係に変換し、視覚的注意機構に対する制約を作成した。

実験結果を表１に示す。すべての実験結果において、視覚的注意機構と言語間注意機構の両方に制約を加えた場合に、BLEU および METEOR が最大となることが実験的に確認できた。これらの結果は、提案した制約付き視覚的注意機構の有効性を示している。

5 考察

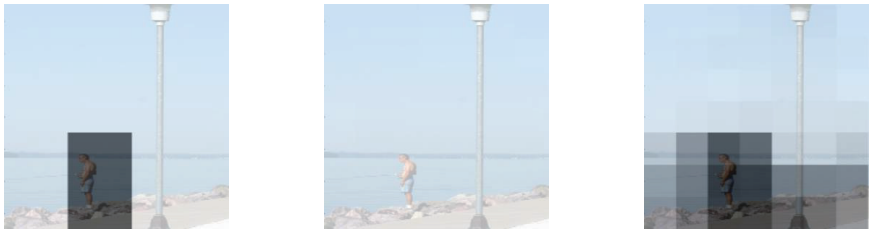
5.1 視覚的注意の例

図５と図６はそれぞれ英日翻訳におけるテストデータ中の単語 man と単語 lamp に対する視覚的注意を表している。これらの図において、より暗くなっている部分に強く注意が向けられていることを表している。図を見ると、視覚的注意機構に制約を与えなかった場合は画像全体に等しく注意が向けられている（図５(b)、図６(b)）。これに対し、制約を与えた場合は、２つの単語それぞれに対応する領域により注意が向けられていることが分かる。これらの結果は、視覚的注意機構に対する制約が、各単語が関連する領域へ注意を向ける効果があることを示している。

5.2 翻訳例

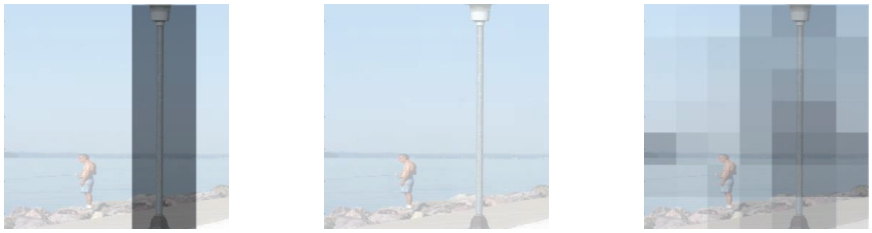
図７は日英翻訳と独英翻訳のテストデータに対する翻訳結果の例を表している。

図を見ると、制約なしのマルチモーダル機械翻訳モデルによって翻訳された文は、原言語文のいくつかの情報



(a) 正解データ　(b) 制約なし　(c) 制約あり

図５　単語 man に対する視覚的注意



(a) 正解データ　(b) 制約なし　(c) 制約あり

図６　単語 lamp に対する視覚的注意

Source: 赤いシャツを着た男の子が、黄色いシャベルで砂を掘っている。

MNMT: a boy in a red shirt is shoveling sand .

MNMT+SVA: a boy in a red shirt is digging in the sand with a yellow shovel .

Reference: a boy wearing a red shirt digs into the sand with a yellow shovel .

(a) 日英翻訳の例

Source: ein rothaariger mann mit dreadlocks sitzt und spielt auf einer akustischen gitarre .

MNMT: a man with red-hair sits on an acoustic guitar .

MNMT+SVA: a red-haired man with dreadlocks is sitting and playing an acoustic guitar .

Reference: a red-haired man with dreadlocks is sitting playing and acoustic guitar .

(b) 独英翻訳の例

図７　翻訳例

が抜け落ちていることが分かる。例えば、図7(a)では、「a yellow shovel」という情報が、図7(b)では画像内の男性の特徴である「dreadlocks」という情報が抜け落ちている。これに対し、制約を加えたモデルではこれらの抜け落ちていた情報を正しく翻訳出来ている。これは、制約付き視覚的注意機構により、原言語文の各単語と画像内の関連する領域が対応付けられ、原言語文がより適切に符号化されたからと考えられる。

5.3 人手での単語の対応関係を用いた実験

Flickr30k entities JP データセットには、原言語文内の単語と目的言語文内の単語との間に人手で対応関係がつけられている。そこで、提案手法であるモデルの、MNMT+SVA と MNMT+SVA+SCA の場合について、この人手での対応関係を用いた実験を行った。表2に実験結果を示す。

表2　人手での単語の対応関係を用いた実験結果。B と M はそれぞれ BLEU と METEOR を表す。

	英→日		日→英	
	B	M	B	M
NMT	43.69	59.27	44.21	40.03
MNMT	44.09	59.59	44.42	40.03
MNMT+SVA	44.51	60.03	44.76	40.40
MNMT+SVA+SCA	44.85	60.35	44.90	40.41

表を見ると、MNMT+SVA と MNMT+SVA+SCA の両方が MNMT より良い結果となっている。また、MNMT+SVA+SCA が英日翻訳と日英翻訳の両方において MNMT+SVA よりも良い結果となった。この結果より、人手による単語の対応関係を用いて注意機構に制約を与えることに効果があることが確認できる。

6 関連研究

ニューラル機械翻訳は、原言語の単語と目的言語の単語間での自動もしくは人手による対応関係に基づいて言語間注意機構を訓練することによって、その性能が改善されてきた。Liu ら[3]や Mi ら[4]は RNN ベースのニューラル機械翻訳モデルに、Garg ら[5] はトランスフォーマーベースのニューラル機械翻訳モデルにおいて制約付き言語間注意機構を提案している。

マルチモーダルニューラル機械翻訳のモデルは様々な種類のものが提案されている。初期のころは、RNN ベースのニューラル機械翻訳を拡張させた、RNN ベースのマルチモーダルニューラル機械翻訳[8][24][25] が主流であった。近年は、トランスフォーマーベースのマルチモーダルニューラル機械翻訳の研究が盛んに行われている[7][13][26][27][28]。ほとんどのマルチモーダルニューラル機械翻訳モデルでは、視覚的注意機構によって画像の特徴を組み込んでいる。原言語文の単語と画像領域との関係を捉えるために視覚的注意機構を利用している研究[8][28] や、目的言語文の単語と画像領域を捉えるために視覚的注意機構を利用している研究[7][13][24][27][29] がある。なお、これらの研究で利用されている視覚的注意機構は訓練時に自動的に学習が行われており、視覚的注意機構に制約を加えた手法ではない。

7 まとめ

本稿では、視覚的注意機構に対する制約を用いたマルチモーダルニューラル機械翻訳モデルを紹介した。提案手法では、ある画像とその説明文中の単語との間に人手で付けられている対応関係を用いて視覚的注意機構の教師データを作成し、その教師データによってマルチモーダルニューラル機械翻訳モデルのエンコーダ内で視覚的注意機構に制約を与えた。実験では、Multi30k データセットを用いた英独翻訳および独英翻訳と Flickr30k Entities JP データセットを用いた英日翻訳および日英翻訳を行い、提案手法によってトランスフォーマーベースのマルチモーダルニューラル機械翻訳モデルの性能が改善できることを確認した。

今後は、本実験で用いたトランスフォーマーベースのマルチモーダルニューラル機械翻訳モデル以外のモデルに対しても、提案手法である制約を与えた視覚的注意機構が有効であるかどうかを検証していきたい。

謝辞

本研究成果は、国立研究開発法人情報通信研究機構の委託研究により得られたものである。ここに謝意を表する。

参考文献

[1] T. Luong, H. Pham, and C. D. Manning. (2015) Effective approaches to attention-

based neural machine translation. In Proc. of EMNLP 2015, pp. 1412–1421.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin. (2017) Attention is all you need. Advances in Neural Information Processing Systems 30, pp. 5998–6008.

[3] L. Liu, M. Utiyama, A. Finch, and E. Sumita. 2016. Neural machine translation with supervised attention. In Proc. of COLING 2016, pages 3093–3102.

[4] H. Mi, Z. Wang, and A. Ittycheriah. 2016. Supervised attentions for neural machine translation. In Proc. of EMNLP 2016, pages 2283–2288.

[5] S. Garg, S. Peitz, U. Nallasamy, and M. Paulik. 2019. Jointly learning to align and translate with transformer models. In Proc. of EMNLP-IJCNLP 2019, pages 4452–4461.

[6] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, and S. Frank. 2018. Findings of the third shared task on multimodal machine translation. In Proc. of the Third Conference on Machine Translation: Shared Task Papers, pages 304–323.

[7] J. Helcl, J. Libovický, and D. Variš. 2018. CUNI system for the WMT18 multimodal translation task. In Proc. of the Third Conference on Machine Translation: Shared Task Papers, pages 616–623.

[8] J. B. Delbrouck and S. Dupont. 2017. Modulating and attending the source image during encoding improves multimodal translation. CoRR, abs/1712.03449.

[9] D. Elliott, S. Frank, K. Sima' an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In Proc. of the 5th Workshop on Vision and Language, pages 70–74.

[10] H. Nakayama, A. Tamura, and T. Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In Proc. of LREC 2020, pages 4204–4210.

[11] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In Proc. of CVPR 2016, pages 770–778.

[12] J. Ba, J. R. Kiros, and G. E. Hinton. 2016. Layer normalization. ArXiv, abs/1607.06450.

[13] J. Libovický, J. Helcl, and D. Mareček. 2018. Input combination strategies for multi-source transformer decoder. In Proc. of the Third Conference on Machine Translation: Research Papers, pages253–260.

[14] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. International Journal of Computer Vision, 123 (1): 74–93.

[15] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2: 67–78.

[16] L. Liu, M. Utiyama, A. Finch, and E. Sumita. 2016. Neural machine translation with supervised attention. In Proc. of COLING 2016, pages 3093–3102.

[17] H. Mi, Z. Wang, and A. Ittycheriah. 2016. Supervised attentions for neural machine translation. In Proc. of EMNLP 2016, pages 2283–2288.

[18] F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, 29 (1) :19–51.

[19] Q. Gao and S. Vogel. 2008. Parallel implementations of word alignment tool. In Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pages 49–57.

[20] D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In Proc. of the 5th Workshop on Vision and Language, pages 70–74.

[21] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. (2007) Moses: open source toolkit for statistical machine translation, In Proc. of ACL on Interactive Poster and Demonstration Sessions, pp. 177–180.

[22] G. Neubig, Y. Nakata, and S. Mori. (2011) Pointwise prediction for robust, adaptable Japanese morphological analysis. In Proc. of ACL 2011, pp. 529–533.

[23] D. P. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. CoRR, abs/1412.6980.

[24] I. Calixto, Q. Liu, and N. Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In Proc. of ACL 2017, pages 1913–1924.

[25] O. Caglayan, W. Aransa, A. Bardet, M. García-Martínez, F. Bougares, L. Barrault, M. Masana, L. Herranz, and J. van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In Proc. of the Second Conference on Machine Translation, pages 432–439.

[26] S.-A. Grönroos, B. Huet, M. Kurimo, J. Laaksonen, B. Merialdo, P. Pham, M. Sjöberg, U. Sulubacak, J. Tiedemann, R. Troncy, and R. Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In Proc. of the Third Conference on Machine Translation: Shared Task Papers, pages 603–611.

[27] J. Ive, P. Madhyastha, and L. Specia. 2019. Distilling translations with visual awareness. In Proc. of ACL 2019, pages 6525–6538.

[28] Z. Zhang, K. Chen, R. Wang, M. Utiyama, E. Sumita, Z. Li, and H. Zhao. 2020. Neural machine translation with universal visual representation. In Proc. of ICLR 2020.

[29] H. Takushima, A. Tamura, T. Ninomiya, and H. Nakayama. 2019. Multimodal neural machine translation using cnn and transformer encoder. In Proc. CICLING 2019.