Vol.2020-NL-246 No.7
Vol.2020-SLP-134 No.7
2020/12/2

マルチモーダル機械翻訳のための画像情報を考慮したデータ拡張

中村夏子[1,a)]　吉永直樹[2,b)]

概要：テキストに加えて画像を入力するマルチモーダル翻訳では，学習に用いる画像つき対訳データの構築コストが問題となる．本研究では，画像なし翻訳との問題設定の違いを考慮して，マルチモーダル翻訳に適した画像を考慮するデータ拡張手法を提案する．具体的に，画像付き目的言語テキストを元にしたマルチモーダル逆翻訳を用いたデータ拡張，さらに，より広範なドメインへの適用を意識して，画像のみを元にした画像キャプション生成を経由するデータ拡張手法を提案する．実験では，Flickr30k に基づく日英，仏英，独英翻訳データセットを用いて評価を行い，通常の逆翻訳に基づくデータ拡張との比較を通じて，提案手法の有効性を確認した．

1. はじめに

深層学習の導入により機械翻訳の性能が著しく向上した結果，文書の翻訳だけでなく，会話や映画の字幕など，実世界の様々な状況下で機械翻訳を運用する機運が高まっている．これらのより現実的な問題設定に応えるため，入力として原言語文に加えて画像を受け取るマルチモーダル機械翻訳が研究されている（2 節）．マルチモーダル機械翻訳では，翻訳時に入力テキストの内容と関係がある画像を参照することで，多義語や係り受け構造の曖昧性解消，また日英翻訳における省略された主語や，名詞の性と数の明示など言語特性の異なる言語への翻訳で必要となる情報の補完をすることができる（2 節）．マルチモーダル機械翻訳により，漫画や映画字幕，動画投稿サイトに投稿された動画，ビデオチャットでの発言，ニュース記事などに含まれる画像の説明文の翻訳など，非言語情報を伴うテキストの翻訳が改善すると期待されている．

マルチモーダル機械翻訳は通常のテキストのみを対象とする機械翻訳と比較して，入力テキストに付随する画像が必要となるため，学習データの開発コストが大きな問題となる．機械翻訳一般においてその翻訳精度は使用するモデルの他，学習データの大きさに強く依存することが知られており [1]，本研究でも 5 節で確認するようにマルチモーダル翻訳も例外ではない（図 2）からである．さらにマルチモーダル翻訳では，画像なし機械翻訳と比べて画像を追加するために入力の空間が大きくなり，パラメータ数を増やすことによる性能向上の余地も大きいと考えられるため，学習データの不足に対処することが重要となる．

1 東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
2 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo
a) n.nakamura@tkl.iis.u-tokyo.ac.jp
b) ynaga@iis.u-tokyo.ac.jp

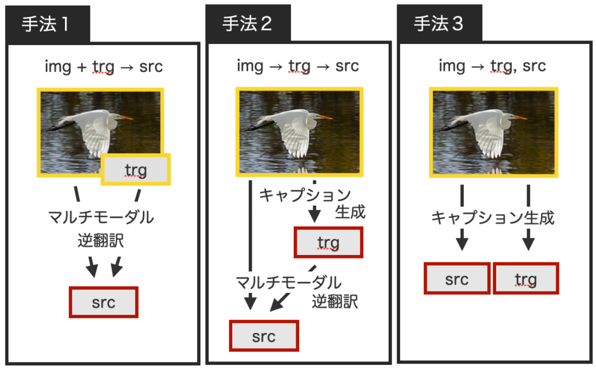

図 1 本研究で提案するデータ拡張手法における擬似教師データ生成方法．擬似教師データとなる画像と原言語文 (src)，目的言語文 (trg) のうち，黄色い枠で囲まれたものが活用する既存のデータ資源で，赤い枠で囲まれたものがシステムによって生成されたものである．

本研究では，この課題に対し機械翻訳でも研究されているデータ拡張によるアプローチを適用することで，学習データの不足を緩和することを目指す（図 1）．データ拡張では，逆翻訳など，既存の教師データに基づくモデルによって擬似的な教師データを生成し，その擬似教師データを元の教師データに加えることによってモデルの大規模学習を可能とする．そこで本研究ではまず，画像付きの目的言語テキストの存在を想定して，マルチモーダル逆翻訳に

Vol.2020-NL-246 No.7
Vol.2020-SLP-134 No.7
2020/12/2

より学習データを得る手法を提案する．さらに，より広範な状況下でのデータ拡張を実現するため，画像のみが存在する状況下でのデータ拡張も検討する．具体的には，画像からキャプション生成で目的言語文を生成した後にマルチモーダル逆翻訳を行う手法と，画像から原・目的言語へのキャプション生成を行なう手法を検討する（3 節）．

これらの提案手法の効果を検証するために，独英・仏英・日英マルチモーダル翻訳について評価実験を行った（4 節）．Multi30k（独英・仏英）[2] と Flickr30k entities JP（日英）[3] を用いて実験した結果，マルチモーダル逆翻訳とキャプション生成マルチモーダル逆翻訳の手法の有効性を確認するとともに，マルチモーダル逆翻訳では画像なし逆翻訳に基づくデータ拡張よりも大きな性能向上が得られることを確認した（5 節）．

2. 関連研究

マルチモーダル機械翻訳は 2016 年に行われた WMT16 shared task [4] で整備された Multi30k [2] を主に用いて様々な手法が提案されている．学習データセットの構築コストに対処した既存アプローチは大きく 2 つに分けることができる．一つは画像つき対訳データ以外の言語資源を活用したマルチモーダル機械翻訳，もう一つがデータ拡張である．

2.1 画像付き対訳データ以外の言語資源を活用したマルチモーダル機械翻訳

画像付き対訳データが不要なマルチモーダル翻訳モデルとしては，画像キャプショニングを用いたリランキング手法が提案されている [5][6][7]．これらのモデルでは，画像なし翻訳モデルから生成した複数の翻訳候補文を画像を用いてランキングし直し，トップになった文を出力する．このモデルでは，翻訳候補文生成の学習のための画像なし対訳データ，リランキングのための画像つき目的言語文のみを必要とし，画像つき対訳データセットを必要としない．

また，機械翻訳と画像ベクトル生成のマルチタスク学習に基づくマルチモーダル翻訳モデルも画像付き対訳データを必要としない [8][9][10]．これらのモデルでは，マルチモーダル機械翻訳を画像なし機械翻訳と画像ベクトル推定の 2 つのサブタスクに分け，それぞれのエンコーダを共有する形で同時に学習を行う．したがって，マルチタスク学習を行う際には画像つき対訳データセットが必要になるが，事前学習として各サブタスクを画像なし対訳データセット，画像つき目的言語文データセットを用いることができる．

また，以上のような外部データセットの利用に特化したモデル以外では，Calixto ら [11] が画像入力に関するパラメータを無視するという形で画像なし対訳データセットによる事前学習を行い，その有効性を確認している．このような事前学習は画像入力に関するパラメータを無視しても損失関数が計算できるモデルであれば可能だが，あらゆるマルチモーダル翻訳モデルに適用できるわけではない．

これらのアプローチに対し，本研究で提案するデータ拡張は学習データを直接増やすため，使用するモデルの制約がないという利点がある．以下では，マルチモーダル翻訳における既存のデータ拡張手法を紹介する．

2.2 マルチモーダル翻訳のためのデータ拡張

既存のマルチモーダル翻訳のためのデータ拡張としては，(画像なし）逆翻訳 [12] が試みられている [11][13]．これは，訓練済み画像なし翻訳モデルを用いて目的言語の画像つき単言語コーパスを原言語に翻訳して擬似対訳データを生成し，学習データに加えてデータ拡張をするものである．この画像なし逆翻訳に基づくデータ拡張は，以下に述べる 2 つの課題を抱えている．

1 つ目の課題は，擬似教師データとして生成される原言語文に画像情報が反映されておらず擬似教師データの質が低いことである．例えば，英語の目的言語文の単語 football に対し画像中にサッカーの図が含まれる場合，日本語の原言語文で対応する単語はサッカーであるはずであるが，この画像なし逆翻訳によってアメリカンフットボールという単語に置き換わってしまう可能性がある．このような場合，目的言語文と画像の対応を考えるようなモデル [5][11][14] への影響はさほどないと予想されるが，原言語文と画像の対応をとるモデル [8][15][16] では間違った対応づけがされてしまい，画像情報がうまく利用できない可能性がある．

さらに 2 つ目の課題として，擬似教師データのもととなるデータとして画像あり目的言語文を用意するコストが無視できない点が挙げられる．以上を踏まえて，提案する新たなデータ手法について次節で説明する．

3. 画像情報を考慮したデータ拡張

本研究では，生成する擬似教師データの質と必要なデータ資源という 2 つの課題を考慮して，マルチモーダル翻訳における新たなデータ拡張手法を 3 つ提案する（図 1）．以下で各手法を紹介するとともに，各手法が応用に適する場面を必要なデータ資源（表 1）の点から考察する．

マルチモーダル逆翻訳

1 つ目の提案手法では学習済みマルチモーダル逆翻訳モデルを用いて画像なし逆翻訳よりマルチモーダル機械翻訳にとって良質の疑似教師データを得ることを目指す．本手法は，既存手法である画像なし逆翻訳と同様に画像付き目的言語文を元にしたデータ拡張を行うが，画像を参照しながら目的言語文を原言語文に逆翻訳するため，より質の高い擬似教師データが生成できることが期待される．漫画や映画，ビデオチャットにおける翻訳などでは画像つき目的言語文のデータセットは豊富にあるため，本手法が活用で

データ拡張手法	擬似教師データ生成モデルの学習データ	擬似教師データの生成に必要なデータ
画像なし逆翻訳	画像なし対訳データ	目的言語キャプション付き画像
マルチモーダル逆翻訳	画像つき対訳データ	目的言語キャプション付き画像
目的言語キャプション生成+マルチモーダル逆翻訳	画像つき対訳データ + 目的言語キャプション付き画像	画像
原・目的言語キャプション生成	目的言語と原言語のキャプション付き画像	画像

表 1　各提案手法で必要となるデータ

きる.

目的言語キャプション生成+マルチモーダル逆翻訳

次に擬似教師データ生成に必要なデータ資源に着目して，画像からキャプション生成モデルを用いて目的言語を生成し，画像と生成された目的言語テキストを用いてマルチモーダル逆翻訳する目的言語キャプション生成+マルチモーダル逆翻訳を提案する．この手法では、擬似教師データ画像のみから生成することができる．したがってこの手法を適用できる状況は画像なし逆翻訳やマルチモーダル逆翻訳よりも多く，特に画像データが収集しやすい写真を対象とする画像キャプション翻訳，また言語資源に乏しい目的言語への翻訳など，画像付きの目的言語文が得られない状況に適している.

原・目的言語キャプション生成

目的言語キャプション生成＋マルチモーダル逆翻訳と同様に画像のみからデータ拡張を行うもう一つの手法として，画像からキャプション生成モデルを用いて目的言語と原言語のテキストをそれぞれ独立に生成するデータ拡張を提案する．これは目的言語キャプション生成+マルチモーダル逆翻訳と同様に必要なデータ資源が少ないというメリットを持つのに加え，擬似教師データを生成するモデルの学習データに対訳データを必要としないというメリットがある (表 1).

したがって，データ拡張する前の教師データが少ないが，画像キャプション生成の学習データは豊富にある場合に適している．これは，原言語と目的言語が英語など言語資源豊かな言語である状況では非常に現実的な設定である.

4. 実験設定

提案手法の有効性を確認するため，既存のデータ拡張手法と本研究で提案するデータ拡張手法を用いてマルチモーダル翻訳モデルを学習し，データ拡張を用いずに学習したマルチモーダル翻訳モデルと翻訳性能の比較を行った．目的言語を英語，原言語はドイツ語，フランス語，日本語をとして，BLEU [17] を用いて翻訳性能の評価を行なった．以下で，詳細な実験設定を述べる.

言語	語彙数
英語	10,827
独語	18,885
仏語	11,838
日本語	13,222

表 2　Multi30k, Flickr 30k における各言語の語彙数

4.1 データセット

データセットは，Multi30k（独英・仏英）[2] と Flickr30k Entities JP（日英）[3] を用いた．どちらも，英語の画像キャプション生成データセット Flickr30k [18] を各言語に翻訳したものである．独英・仏英は各画像に対し対訳データが 1 つのみである一方で，日本語は各画像 2 つの対訳データがある．学習データセットの大きさを言語間で揃えるために日本語は各画像に対し 1 つ目のキャプションのみを対訳データとして扱った.

テキストの前処理として，英語，独語，仏語はすべて小文字化をしたのち Moses SMT toolkit v4.0[*1]を用いて，正規化とトークン化を行った．日本語は KyTea (ver. 0.4.7)[*2]を用いて単語分割を行った．これによって各言語の語彙は表 2 に示した通りとなり，翻訳とキャプション生成どちらもこの全ての語彙を用いて学習を行った.

画像の前処理として，pytorch (ver. 1.4.0)[*3]を用いて物体認識タスクで事前学習済みの ResNet-50 [19] の red4f レイヤーの活性化層から画像特徴抽出を行った.

4.2 モデル

画像なし翻訳モデルは, エンコーダを 2 層の双方向 GRU, デコーダを 2 層の Conditional GRU [20] とした．マルチモーダル翻訳モデルは，エンコーダを 2 層の双方向 GRU, デコーダを doubly-attentive decoder [11] とした．どちらも単語埋め込み層，モデルサイズは 500 次元とし，最適化手法は Adam [21] を用いた．バッチサイズは 40，学習率は 0.002，ドロップアウト率は 0.3，デコード時のビームサイズは 5 とした．実装は [11] の著者実装[*4]を用いた.

画像キャプション生成モデルは [22] を用いた．エンコーダとデコーダはそれぞれ 6 層の transformer [23] とし，単

*1 http://www.statmt.org/moses/
*2 http://www.phontron.com/kytea/index-ja.html
*3 https://pytorch.org
*4 https://github.com/iacercalixto/MultimodalNMT

データ拡張手法	学習データサイズ	de→en	fr→en	jp→en
データ拡張なし	14.5k	35.48	46.12	32.95
画像なし逆翻訳	14.5k+擬似 14.5k	**37.51**	48.85	35.09
マルチモーダル逆翻訳	14.5k+擬似 14.5k	37.41	**50.03**	**35.30**
目的言語キャプション生成+マルチモーダル逆翻訳	14.5k+擬似 14.5k	36.49	48.43	33.34
原・目的言語キャプション生成	14.5k+擬似 14.5k	32.61	42.87	20.38
データ拡張なし	29k	39.35	51.46	38.55

表 3 提案するデータ拡張手法と既存のデータ拡張手法，さらにデータ拡張をしない場合の翻訳の結果（BLEU）の比較

語埋め込み層は 512 次元，モデルサイズは 2048 次元とし，最適化手法は Adam [21] を用いた．バッチサイズは 15，学習率は 0.0005，ドロップアウト率は 0.5，デコード時のビームサイズは 1 とした．実装は [22] の著者実装*5を用いた．

4.3 データ拡張

データ拡張は，教師データの大きさが 14.5k(A) と 5k(B) の 2 つの設定で行い，どちらも 14.5k の擬似教師データを生成した．(A) では目的言語を英語，原言語を独語，仏語，日本語として実験を行い，(B) では目的言語を英語，原言語を独語として実験を行った．(A) では，逆翻訳モデルを教師データ 14.5k で学習し，キャプション生成モデルの学習は教師データの画像と同じ画像 14.5k の単言語キャプションつき画像で学習を行った．一方 (B) では，(A) よりも画像つき単言語キャプションデータが画像つき対訳データに比べて潤沢にある場合を想定して，逆翻訳モデルは教師データ 5k で学習し，キャプション生成モデルは教師データの画像を含む画像 14.5k 枚の単言語キャプションつき画像を用いて学習を行った．この状況は，逆翻訳モデルよりも画像キャプション生成モデルの学習データの方が大きいので，提案手法のうち原・目的言語キャプション生成に有利である．

各画像に対して存在するキャプションの数が言語によって異なっており，英語・独語は 5 文，仏語は 1 文，日本語では 2 文である．4.1 節で述べたように対訳データセットは各画像につき 1 文としたが，画像キャプション生成は各言語で存在する全てのキャプションを用いて学習した．

5. 結果

各データ手法を適用した結果を表 3 に記す．表中一番上と下の行のデータ拡張なしは，学習データが全て教師データである場合の結果であり，一番下のデータ拡張なし (29k) は 14.5k の教師データに擬似教師データ 14.5k を加えてデータで学習したときの実質的な上限と解釈できる．表 3 から，マルチモーダル逆翻訳と目的言語キャプション生成+マルチモーダル翻訳はデータ拡張をしていない場合と比べて翻訳精度を向上できていることがわかる．さらに，マルチモーダル逆翻訳は画像なし逆翻訳と同等またはそれ以上の精度向上が達成されている．

*5 https://github.com/yahoo/object_relation_transformer

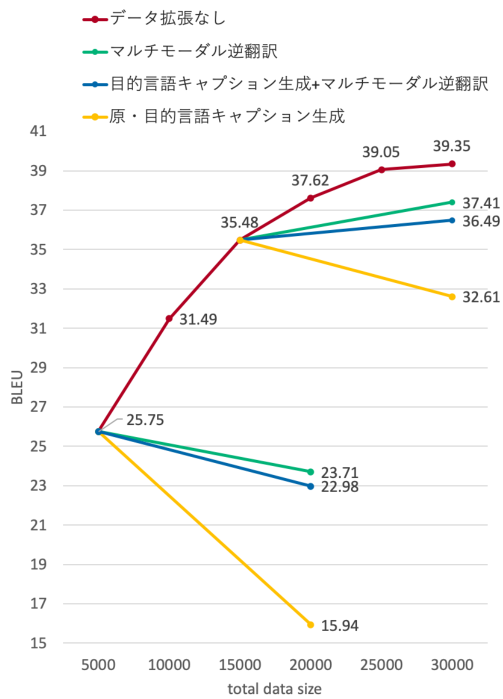

図 2 データ拡張なしの場合の学習曲線とデータ拡張をした翻訳精度（BLEU）の比較．

図 2 は独語から英語への翻訳について，データ拡張なしのマルチモーダル翻訳の学習曲線と，教師データが 5k と 14.5k の場合のデータ拡張の結果である．教師データが 15k の場合については，マルチモーダル逆翻訳の値とデータ拡張なしの学習データサイズ 20k のときの値がほぼ等しいことから，マルチモーダル逆翻訳により生成された擬似教師データ 14.5k は学習データの価値として教師データ 5k に匹敵することがわかる．

また教師データが 5k の場合は，どの提案手法も逆効果であった．マルチモーダル逆翻訳と目的言語キャプション生成+マルチモーダル逆翻訳については，擬似教師データを生成する逆翻訳モデルの学習データが少ないために擬似教師データの質が下がってしまったと解釈できる．原・目的言語キャプション生成も逆効果であった理由は次節で，追加実験を通して分析を行う．

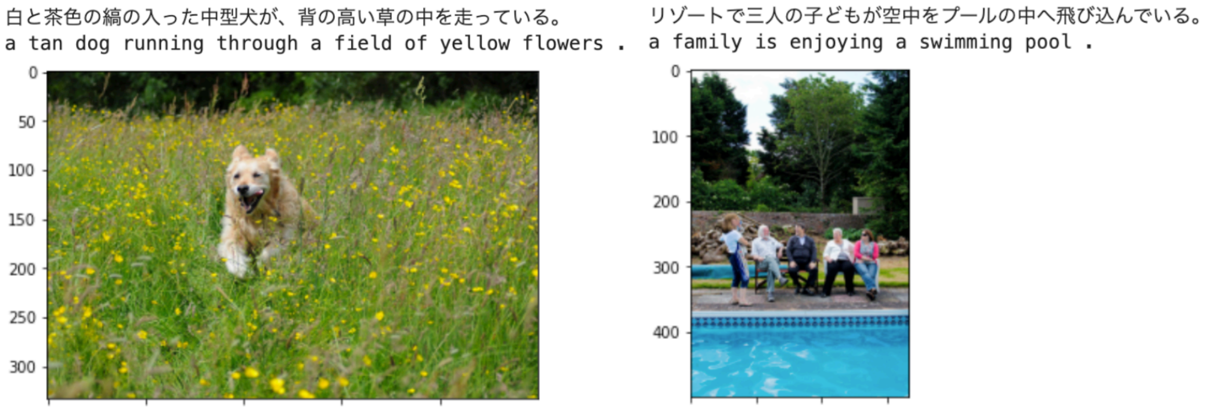

図 3 画像キャプション生成モデルによって出力された日本語と英語のキャプション．左は，各言語とも正しいキャプションが生成できているが，日本語文と英語文で含まれる情報が異なり，直接の対訳関係とはなっていない．右は日本語キャプションが画像に対して間違った文を生成している．

学習データサイズ	BLEU(de→en)
14.5k（データ拡張なし）	35.48
14.5k+擬似 14.5k	33.39
14.5k+擬似 130.5k	23.34

表 4 学習データとして，教師データに加えて人手で各言語独立に与えられたキャプションを擬似教師データとして加えた場合の翻訳結果．

6. 議論

5 節の実験結果において，原・目的言語キャプション生成によるデータ拡張で翻訳性能の性能改善が見られなかった原因としては以下の二点が考えられる．一つは画像キャプション生成により誤ったキャプションが生成されていること，もう一つが目的言語と原言語でそれぞれ正しいキャプションが生成されているが対訳関係になっていないことである（図 3）．我々は原・目的言語キャプション生成の改善に向けて，どちらの原因が優位なのかを調べるための追加実験を行った．

追加実験では，画像に対して人手でそれぞれ独立に与えられた原言語と目的言語のキャプション [2] を擬似教師データとして教師データに追加し，学習を行った．このデータ拡張は画像キャプション生成モデルの精度が理想的な状況である場合の原・目的言語キャプション生成であると捉えることができる．すなわち，このデータ拡張手法が有効であるならば，原・目的言語キャプション生成は画像キャプション生成モデルの学習データが豊富にある状況では有効に働く可能性があると推測できる．

追加実験の結果を表 5 に示す．この結果から，一つの画像に対して人手によって生成されたキャプションであっても，明示的な対訳関係がなければ教師データに加えることで翻訳の精度が下がってしまうことがわかった．このことから原・目的言語キャプション生成が逆効果になってしまうボトルネックは，目的言語と原言語のキャプションが直接の対訳関係にないことにあると考えることができる．

7. おわりに

本研究ではマルチモーダル翻訳のための画像情報を考慮したデータ拡張手法を 3 つ提案した．一つ目のマルチモーダル逆翻訳では既存手法である画像なし逆翻訳を上回る結果が得られた．また，2 つ目の画像から目的言語文を生成しさらにマルチモーダル逆翻訳をする手法は，データ拡張に使えるデータ資源が画像のみであるという不利な設定でありながらも有効な手法であることがわかった．3 つ目の画像から目的言語文と原言語文を独立に生成する手法は，生成される文対が直接対訳関係にないために逆効果になってしまうことがわかった．

今後の展望としては，各データ拡張において学習データに疑似学習データを混合する比率を変えた実験を行うとともに，原・目的言語キャプション生成の改善として，対訳関係をとるために 1 つのシステムから目的言語文と原言語文を同時に生成するマルチランゲージデコーダ [24] の利用を考えている．

謝辞 本研究は JST，CREST，JPMJCR19A4 の支援を受けたものである．

参考文献

[1] Koehn, P. and Knowles, R.: Six Challenges for Neural Machine Translation, *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, Association for Computational Linguistics, pp. 28–39 (online), DOI: 10.18653/v1/W17-3204 (2017).

[2] Elliott, D., Frank, S., Sima'an, K. and Specia, L.: Multi30K: Multilingual English-German Image Descriptions, *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany, Association for Computational Linguistics, pp. 70–74 (online), DOI:

10.18653/v1/W16-3210 (2016).

[3] Nakayama, H., Tamura, A. and Ninomiya, T.: A Visually-Grounded Parallel Corpus with Phrase-to-Region Linking, *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, pp. 4204–4210 (online), available from ⟨https://www.aclweb.org/anthology/2020.lrec-1.518⟩ (2020).

[4] Specia, L., Frank, S., Sima'an, K. and Elliott, D.: A Shared Task on Multimodal Machine Translation and Crosslingual Image Description, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany, Association for Computational Linguistics, pp. 543–553 (online), DOI: 10.18653/v1/W16-2346 (2016).

[5] Hitschler, J., Schamoni, S. and Riezler, S.: Multimodal Pivots for Image Caption Translation, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 2399–2409 (online), DOI: 10.18653/v1/P16-1227 (2016).

[6] Shah, K., Wang, J. and Specia, L.: SHEF-Multimodal: Grounding Machine Translation on Images, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany, Association for Computational Linguistics, pp. 660–665 (online), DOI: 10.18653/v1/W16-2363 (2016).

[7] Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L. and van de Weijer, J.: Does Multimodality Help Human and Machine for Translation and Image Captioning?, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, Berlin, Germany, Association for Computational Linguistics, pp. 627–633 (online), DOI: 10.18653/v1/W16-2358 (2016).

[8] Elliott, D. and Kádár, Á.: Imagination Improves Multimodal Translation, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei, Taiwan, Asian Federation of Natural Language Processing, pp. 130–141 (online), available from ⟨https://www.aclweb.org/anthology/I17-1014⟩ (2017).

[9] Zhou, M., Cheng, R., Lee, Y. J. and Yu, Z.: A Visual Attention Grounding Neural Model for Multimodal Machine Translation, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 3643–3653 (online), DOI: 10.18653/v1/D18-1400 (2018).

[10] Helcl, J., Libovický, J. and Variš, D.: CUNI System for the WMT18 Multimodal Translation Task, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, Association for Computational Linguistics, pp. 616–623 (online), DOI: 10.18653/v1/W18-6441 (2018).

[11] Calixto, I., Liu, Q. and Campbell, N.: Doubly-Attentive Decoder for Multi-modal Neural Machine Translation, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, Association for Computational Linguistics, pp. 1913–1924 (online), DOI: 10.18653/v1/P17-1175 (2017).

[12] Sennrich, R., Haddow, B. and Birch, A.: Improving Neural Machine Translation Models with Monolingual Data, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics, pp. 86–96 (online), DOI: 10.18653/v1/P16-1009 (2016).

[13] Calixto, I. and Liu, Q.: Incorporating Global Visual Features into Attention-based Neuralachine Translation., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 992–1003 (online), DOI: 10.18653/v1/D17-1105 (2017).

[14] Ive, J., Madhyastha, P. and Specia, L.: Distilling Translations with Visual Awareness, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 6525–6538 (online), DOI: 10.18653/v1/P19-1653 (2019).

[15] Ma, M., Li, D., Zhao, K. and Huang, L.: OSU Multimodal Machine Translation System Report, *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, Association for Computational Linguistics, pp. 465–469 (online), DOI: 10.18653/v1/W17-4751 (2017).

[16] Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J. and Luo, J.: A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, pp. 3025–3035 (online), DOI: 10.18653/v1/2020.acl-main.273 (2020).

[17] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 311–318 (online), DOI: 10.3115/1073083.1073135 (2002).

[18] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 67–78 (online), DOI: 10.1162/tacl_a_00166 (2014).

[19] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition (2015).

[20] Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A. V. M., Mokry, J. and Nădejde, M.: Nematus: a Toolkit for Neural Machine Translation (2017).

[21] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization (2017).

[22] Herdade, S., Kappeler, A., Boakye, K. and Soares, J.: Image Captioning: Transforming Objects into Words (2020).

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention Is All You Need (2017).

[24] Wang, Y., Zhang, J., Zhai, F., Xu, J. and Zong, C.: Three Strategies to Improve One-to-Many Multilingual Translation, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Association for Computational Linguistics, pp. 2955–2960 (online), DOI: 10.18653/v1/D18-1326 (2018).