1A1-OS-05a-3

画像とテキストの潜在的な意味情報を用いたニューラル翻訳モデルの提案

Neural Machine Translation with Latent Semantic of Image and Text

冨山　翔司 *1 Joji Toyama　味曽野雅史 *2 Masanori Misono　鈴木雅大 *3 Masahiro Suzuki　中山浩太郎 *4 Kotaro Nakayama　松尾豊 *5 Yutaka Matsuo

*1 東京大学大学院工学系研究科松尾研究室
The University of Tokyo, Matsuo Laboratory

Although attention-based Neural Machine Translation have achieved great success, attention-mechanism cannot capture the entire meaning of the source sentence because the attention mechanism generates a target word depending heavily on the relevant parts of the source sentence. The report of earlier studies has introduced a latent variable to capture the entire meaning of sentence and achieved improvement on attention-based Neural Machine Translation. We follow this approach and we believe that the capturing meaning of sentence benefits from image information because human beings understand the meaning of language not only from textual information but also from perceptual information such as that gained from vision. As described herein, we propose a neural machine translation model that introduces a continuous latent variable containing an underlying semantic extracted from texts and images. Our model can be trained end-to-end. Experiments conducted with an English–German translation task show that our model outperforms over the baseline.

1. はじめに

ニューラル翻訳 (NMT) は近年めざましい成功を収めている [Sutskever 14, Bahdanau 15]．NMT は，統計的機械翻訳と異なり，ルールやフレーズベースの規則をほとんど必要としないという長所を持つ．しかし，現在の標準的な NMT モデルであるアテンション機構を持った NMT[Bahdanau 15] は，ターゲットの言葉を生成する時にソースのある特定の部分に特に注目するため，ソース文全体の意味を把握することができていないという欠点が指摘されている [Tu 16]．この欠点に対処した研究として，ソースとターゲットが共通して持つ意味情報を，潜在変数として陽に含むモデル，Variational Neural Machine Translation(VNMT) がある [Zhang 16]．NMT はソース文全体の意味情報をモデル内に陰に含むのに対し，VNMT は意味情報をモデル内に潜在変数として陽に含む．提案モデルは VNMT と同じく，ソース文全体の意味情報をうまく捉えることによる翻訳精度向上を目標とする．提案モデルは潜在変数のモデリングに画像情報も用いることでより良い意味情報の獲得を試みる．

視覚情報は言語と結びついている．例えば，人間は言葉の意味を，周囲の環境から与えられる知覚的情報と結びつけることで理解する [Barsalou 99]．これを人間は自然に行っているが，計算機が異なるドメイン情報を統合的に理解するのは難しい．しかし，異なるドメインの情報の統合的理解は，自然言語処理の飛躍的な発展につながるかもしれない．近年では，画像からキャプションを生成した例 [Xu 15] やテキストから画像を生成した例 [Reed 16] があり，異なるドメイン情報を含めた自然言語の統合的理解への可能性が示唆されている．

本論文では，テキストと画像が持つ意味情報を，潜在変数として陽に含むニューラル翻訳モデルを提案する．本手法は Variational Autoencoder(VAE) [Kingma 14, Rezende 14] を用いることで，テキストと画像から潜在変数 $\mathbf{z}$ をモデリングする．本論文で提案するモデルは，翻訳する際，まず画像とソースから潜在変数 $\mathbf{z}$ を生成し，次に $\mathbf{z}$ を NMT のデコーダに組み入れ，最後にターゲットを出力する．本手法と VNMT との差分はテキストの情報に加えて画像情報を潜在変数のモデリングに用いたところにある．

実験では，Multi30k[Elliott 16] という，画像とそれに対応する英独の対訳コーパスを用い，NMT，VNMT，CMU(現在最高精度のマルチモーダル翻訳手法）の三つのベースラインに対して，提案モデルとの比較を行った．提案モデルは標準的な翻訳精度評価指標である METEOR [Denkowski 14] スコアにおいて全てのベースラインを上回った．また，提案モデルでどのように翻訳結果が良くなったのかを幾つかの例で示す．

連絡先: 冨山翔司，東京大学工学系研究科松尾研究室，08041381485, toyama.jouji@gmail．com

2. 関連研究

提案モデルは，VNMT の拡張である．また，提案モデルは，画像を使っているという点でマルチモーダル翻訳モデルの一種であると見ることもできる． 本章ではこれらについて簡単に説明する．

2.1 Variational Neural Machine Translation

VNMT は NMT に潜在変数を導入したニューラル翻訳モデルである．モデルの構造は図 1の，$\boldsymbol{\pi}$ からの矢印を除いたものと一致する． VNMT では $\mathbf{z}$ の推論に VAE を用いている．我々の提案モデルはテキストに加えて画像も用いて潜在変数 $\mathbf{z}$ を獲得する．

2.2 Multimodal Translation

マルチモーダル翻訳とは，対訳コーパスに加えて画像を用いて翻訳を行うタスクである．マルチモーダル翻訳は [Elliott 15] によって初めて提案された．数々の研究が為されたものの，翻訳精度を大きく改善したと言えるモデルは未だ無い [Caglayan 16]．

NMT ベースのマルチモーダル翻訳手法の代表的なものとして，[Huang 16] らの研究がある。[Huang 16] は，画像特徴量をソースの文章系列の先頭に入れることで，マルチモーダル翻訳を行った．画像特徴量は Region-based Convolutional Neural Networks (R-CNN) [Girshick 15] と VGG-19 [Simonyan 14] を用いて抽出した．本モデルは，Workshop of Machine Trans-

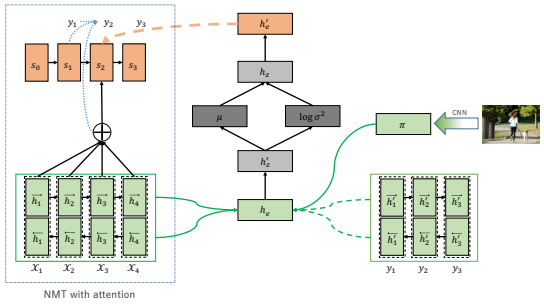

図 1: 提案モデルの構造
y からの緑破線は，訓練時にのみ用いられることを示す．

lation 2016 (WMT16[*1]) において NMT ベースのモデルの中で最も高い METEOR スコアを記録した．本研究ではこのモデルを CMU として，実験の比較対象として用いた．

3. 提案モデル

本章では，画像とテキストの意味情報を潜在変数として陽に含むニューラル翻訳モデルを提案する．

提案モデルはグラフィカルモデルを用いて図 2のように書くことができる．この変分下界は，

$$\mathcal{L} = -\mathrm{D}_{\mathrm{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})||p_\theta(\mathbf{z}|\mathbf{x},\,\boldsymbol{\pi})\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})}[\log p_\theta(\mathbf{y}|\mathbf{z},\,\mathbf{x})], \tag{1}$$

となる．ここで，$\mathbf{x}$, $\mathbf{y}$, $\boldsymbol{\pi}$, $\mathbf{z}$ はそれぞれ，ソース，ターゲット，画像，潜在変数を表し，p_θ と q_ϕ はそれぞれ，事前分布と事後分布を表す．$p(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})$ が真に求めたい分布であるが，この計算は手に負えないため，代わりに近似分布 $q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})$ をモデル化し，更に事前分布 $p_\theta(\mathbf{z}|\mathbf{x},\,\boldsymbol{\pi})$ を置くことで，テスト時にソースと画像から翻訳ができるようにする．

提案モデルでは式 (1) 内の分布をニューラルネットによってモデリングする．提案モデルは，1) エンコーダ， 2) 推論， 3) デコーダ の三つに分けることができる．提案モデルの全体を図 1に示す．

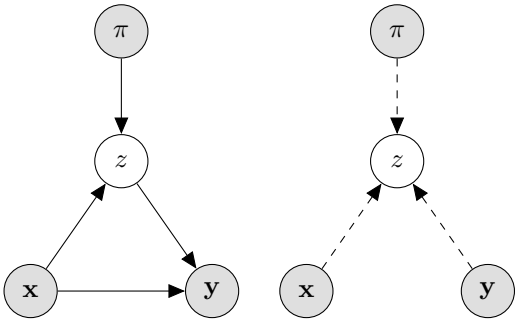

図 2: 提案モデル
提案モデルでは, 潜在変数 $\mathbf{z}$ はソース $\mathbf{x}$ と画像 $\boldsymbol{\pi}$ から生成される. その後, ターゲット $\mathbf{y}$ は $\mathbf{x}$ と $\mathbf{z}$ から生成される.$\mathbf{z}$ は $\mathbf{x}$,$\mathbf{y}$,$\boldsymbol{\pi}$ から推論される．

3.1 エンコーダ

エンコーダでは，意味特徴量 $\mathbf{h_e}$ をソース，ターゲット，画像から獲得する．この意味特徴量は推論で用いられる．この節は，図 1の緑部に対応する．

3.1.1 テキストエンコード

ソースとターゲットは [Bahdanau 15] と同じようにエンコードされる．双方向 RNN を用いて，ソースとターゲットはそれぞれ特徴量 $\mathbf{h}_f$ と $\mathbf{h}_g$ にエンコードされる．

3.1.2 画像エンコードと意味特徴量

提案モデルでは，画像特徴量を畳み込みネットワーク (CNN) から取得する．具体的には，画像を VGG-19 に入れ，fc7 層を取り出してそれを画像特徴量 $\boldsymbol{\pi}$ とする．その後，画像特徴量 $\boldsymbol{\pi}$ は，アフィン変換によって以下のようにエンコードされる．

$$\mathbf{h}_\pi = W_\pi \boldsymbol{\pi} + b_\pi \quad \text{where } W_\pi \in \mathbb{R}^{d_\pi \times d_{fc7}},\ b_\pi \in \mathbb{R}^{d_\pi}.$$

エンコードされた特徴量 $\mathbf{h}_f$， $\mathbf{h}_g$， $\mathbf{h}_\pi$ を全て結合することによって意味特徴量を，$\mathbf{h}_e = [\mathbf{h}_f; \mathbf{h}_g; \mathbf{h}_\pi]$ のように得る．ここで $\mathbf{h}_e \in \mathbb{R}^{d_e = 2\times d_h + d_\pi}$ である．

3.2 推論

提案モデルは事後分布 $q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})$ と事前分布 $p_\theta(\mathbf{z}|\mathbf{x},\pi)$ をニューラルネットによってモデリングする．この章は，図 1の黒と灰色の部分に対応する．

3.2.1 事後分布の近似

提案モデルでは，事後分布 $p_\theta(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})$ の代わりに近似事後分布 $q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})$ を，VAE を用いることでモデリングする．$q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})$ を次のような分布であると仮定する．

$$q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi}) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu}(\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi}),\,\boldsymbol{\sigma}(\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})^2\mathbf{I}).$$

平均 $\boldsymbol{\mu}$ と標準偏差 $\boldsymbol{\sigma}$ がニューラルネットの出力となる．

エンコーダによって生成された意味特徴量 $\mathbf{h}_e$ は次のように潜在空間に写像される

$$\mathbf{h'}_z = g(W_z^{(1)}\mathbf{h}_e + \mathbf{b}_z^{(1)}).$$

ここで，$g(\cdot)$ は要素に対する活性化関数であり，$\tanh(\cdot)$ とする．平均 $\boldsymbol{\mu}$ と標準偏差 $\boldsymbol{\sigma}$ は $\mathbf{h'}_z$ を用いて，

$$\boldsymbol{\mu} = W_\mu \mathbf{h'}_z + \mathbf{b}_\mu,\ \log\boldsymbol{\sigma}^2 = W_\sigma \mathbf{h'}_z + \mathbf{b}_\sigma,$$

のように求めることができる．

3.2.2 事前分布

提案モデルでは，事前分布 $p_\theta(\mathbf{z}|\mathbf{x},\pi)$ を次のような分布であると仮定する．

$$p_\theta(\mathbf{z}|\mathbf{x},\,\boldsymbol{\pi}) \quad = \quad \mathcal{N}(\mathbf{z};\boldsymbol{\mu}'(\mathbf{x},\,\boldsymbol{\pi}),\,\boldsymbol{\sigma}'(\mathbf{x},\,\boldsymbol{\pi})^2\mathbf{I}).$$

$\boldsymbol{\mu}'$ と $\boldsymbol{\sigma}'$ は，3.2.1章と同じ方法で生成される．訓練時では，潜在変数 $\mathbf{z}$ は再パラメタ化トリックを用いて， $\mathbf{h}_z = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ のように獲得され，事前分布と事後分布の KL 情報距離 $\mathrm{D}_{\mathrm{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x},\,\mathbf{y},\,\boldsymbol{\pi})||p_\theta(\mathbf{z}|\mathbf{x},\,\boldsymbol{\pi})\right]$ を訓練の目的関数に加える．翻訳時は，$\mathbf{h}_z$ には $p_\theta(\mathbf{z}|\mathbf{x},\,\boldsymbol{\pi})$ の平均， つまり， $\boldsymbol{\mu}'(\mathbf{x},\,\boldsymbol{\pi})$ を用いる．

$\mathbf{h}_z$ はターゲット空間へ以下のように写像され，デコーダへ組み込まれる

$$\mathbf{h}'_e = g(W_z^{(2)}\mathbf{h}_z + \mathbf{b}_z^{(2)}) \quad \text{where } \mathbf{h}'_e \in \mathbb{R}^{d'_e}.$$

*1 http://www.statmt.org/wmt16/

3.3 デコーダ

本章は図 1のオレンジ部に対応する．ソース $\mathbf{x}$ と潜在変数 $\mathbf{z}$ が与えられた時，ターゲット $\mathbf{y}$ は以下のように生成される．

$$p(\mathbf{y}|\mathbf{z},\mathbf{x}) = \prod_{j=1}^{T} p(\mathbf{y}_j|\mathbf{y}_{<j},\mathbf{z},\mathbf{x}).$$

3.4 Model Training

式 (1) を求めるために，モンテカルロ法を用い，$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},\boldsymbol{\pi})} \approx \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{h}_z^{(l)})$ と近似する．ここで L はサンプリング数である．これによって訓練時の目的関数は

$$\mathcal{L}(\theta,\phi) = -\mathrm{D}_{\mathrm{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},\boldsymbol{\pi})||p_\theta(\mathbf{z}|\mathbf{x},\boldsymbol{\pi})\right] + \frac{1}{L}\sum_{l=1}^{L}\sum_{j=1}^{T}\log p_\theta(\mathbf{y}_j|\mathbf{y}_{<j},\mathbf{x},\mathbf{h}_z^{(l)})$$

となる．

4. 実験

4.1 実験設定

実験では Multi30k [Elliott 16] をデータセットとして用いる．訓練データとして 29,000 文，検証データとして 1,014 文，そしてテストデータとして 1,000 文を用意した。画像特徴量は VGG-19 CNN [Simonyan 14] の fc7 層を用いた．

訓練時，提案モデルは VNMT の学習と同様 [Zhang 16] に，NMT によってまず学習をした後，NMT と重みが共通の部分を残したまま，再学習した．毎エポック毎にデータセットをシャッフルした．翻訳時，ビームサーチを用い，ビーム幅は 12 とした．

提案モデルと比較モデルの実装は *dl4mt*[*2] をベースに行った． *dl4mt* は基本的には [Bahdanau 15] と同一であるが，デコーダ部分に GRU の代わりに conditional GRU を用いている．結果の評価には `MultEval` [*3] によって算出した METEOR スコアと BLEU スコアを用いた．

提案モデルの画像特徴量の代わりに，ガウシアンノイズのみを加えたものを比較対象に加えた．このモデルは画像特徴量を一切用いていない．この実験は，我々の提案モデルの精度向上がランダム値を与えられたことによる過学習の緩和ではなく，画像の情報によってもたらされていることを示すために行った．表 1では N と表記する．

4.2 結果

表 1: Multi30k を用いた実験結果．括弧付きの結果は ‘-norm’ パラメータの時であることを示す． CMU のスコアは [Huang 16] を参照した．

	METEOR ↑		BLEU ↑	
	val	test	val	test
NMT	50.06 (53.01)	49.67 (53.72)	33.1	33.9
VNMT	50.29 (53.09)	49.66 (53.57)	33.4	33.7
CMU	- (-)	50.8 (54.1)	-	**35.1**
N	50.50 (53.70)	50.44 (54.43)	34.5	34.6
Our Model	**51.26 (54.93)**	**51.56 (55.60)**	**35.1**	**35.1**

∗2 https://github.com/nyu-dl/dl4mt-tutorial

∗3 https://github.com/jhclark/multeval, `MultEval` のデフォルトの評価指標である meteor1.4 では無く，meteor1.5 を用いている．

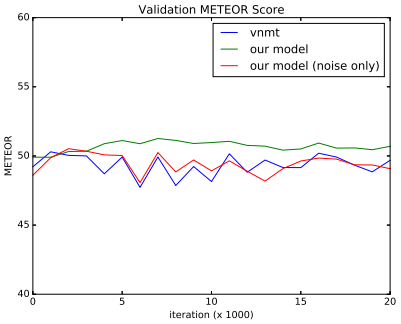

図 3: 検証用データにおける 1000 イテレーション毎の METEOR スコアの推移

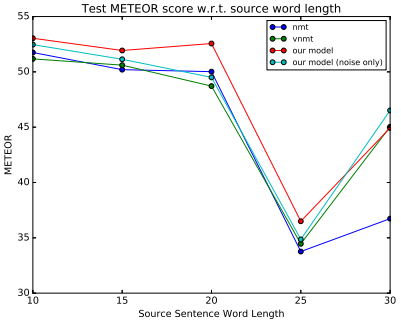

図 4: ソース文の長さでグループ化した時の平均 METEOR スコア

表 1に示した実験結果から，提案モデルは METEOR スコアでは他のベースラインを上回り，BLEU では CMU と同等のスコアを記録していることがわかる．図 3は検証用データにおける METEOR スコアの推移を示し，図 4はソース文の長さでグループ化した時の平均 METEOR スコアを示す．

4.3 定量的分析

表 1は，提案モデルが METEOR スコアで他のベースラインを上回り，BLEU では CMU と同等のスコアを記録していることがわかる．VNMT は NMT とほとんど変わらないスコアだった．これは今回用いた Multi30k がほとんど短い文で構成されており，特に長文に対して良い翻訳ができる [Zhang 16]VNMT の利点が生かされなかったのかもしれない．

表 3では，検証用データにおいて，提案モデルが VNMT, noise only に比べて 1000 イテレーション毎の METEOR スコアで上回っていることがわかる．表 3はまた，提案モデルが VNMT， noise only に比べて安定して METEOR スコアを記録していることがわかる．

表 4は，提案モデルが 30 単語以下のソース文に対して，VNMT， NMT， noise only より良い METEOR スコアを記録していることがわかる． 30 単語以上のソース文に対しては NMT が他手法に比べて大きく METEOR スコアが落ちたが，これは VNMT が NMT に比べて長文の翻訳がうまくいくという主張 [Zhang 16] に一致する．

4.4 定性的分析

定性的分析として，提案モデルが VNMT に比べて特に METEOR スコアが大きく改善した翻訳結果と，大きく改悪した翻訳結果を，それぞれ上から１０文ずつをネイティヴのドイツ語話者に見せ，コメントをもらった．

頂いたコメントによると，提案モデルが改善を見せた文のうち 6 文は，提案モデルが VNMT に比べて名詞の訳漏れが少ないからであった．一方，提案モデルが改悪を見せた文のうち，VNMT が提案モデルに比べて名詞の訳漏れが少なかった文は１文だけであった．提案モデルが改悪を見せた理由は前置詞の誤りや，文法的な問題によるものがほとんどであった．

これは提案モデルが単語の訳漏れを抑制していることを示唆している．訳漏れとは，ソース文の単語を丸々訳し損ねてしまう現象であり，これは attention 機構が文の全体の意味を踏まえないことによって引き起こされるとで指摘されている [Tu 16]．提案モデルは画像を用いることによって，単語，特に画像内で表現されることの多い名詞の訳漏れを防いでいるのではないかと考えられる．図 5 がその一例であり，‘trampoline’ の訳漏れを抑制している．

ソース	a woman does a somersault on a trampoline on the beach.
正解	eine frau macht einen salto auf einem trampolin am strand.
VNMT	eine frau macht einen salto am strand.
提案モデル	eine frau macht einen salto auf einem trampolin am strand.

図 5: 提案モデルによって訳漏れが抑制されている例

5. 結論

本論文では，画像とテキストの意味情報を，潜在変数として陽に含むニューラル翻訳モデルを提案した．提案モデルは METEOR において他のベースラインを上回り，BLEU では他のベースラインと同等の翻訳精度を持つ．実験結果から，提案モデルが VNMT に比べて名詞の訳漏れを抑制していることが考えられる．

謝辞

本研究は JSPS 科研費 JP25700032, JP15H05327, JP16H06562 の助成を受けたものである．

参考文献

[Bahdanau 15] Bahdanau, D., Cho, K., and Bengio, Y.: Neural machine translation by jointly learning to align and translate, in *ICLR* (2015)

[Barsalou 99] Barsalou, L. W.: Perceptual symbol Systems, *Behavioral and Brain Sciences*, Vol. 22, pp. 577–609 (1999)

[Caglayan 16] Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and Weijer, van de J.: Does Multimodality Help Human and Machine for Translation and Image Captioning?, in *WMT* (2016)

[Denkowski 14] Denkowski, M. and Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language, in *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation* (2014)

[Elliott 15] Elliott, D., Frank, S., and Hasler, E.: Multilingual Image Description with Neural Sequence Models, *ArXiv e-prints* (2015)

[Elliott 16] Elliott, D., Frank, S., Sima'an, K., and Specia, L.: Multi30K: Multilingual English-German Image Descriptions, *CoRR*, Vol. abs/1605.00459, (2016)

[Girshick 15] Girshick, R.: Fast R-CNN, in *ICCV* (2015)

[Huang 16] Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C.: Attention-based Multimodal Neural Machine Translation, in *WMT* (2016)

[Kingma 14] Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M.: Semi-supervised Learning with Deep Generative Models, in *NIPS* (2014)

[Reed 16] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H.: Generative Adversarial Text to Image Synthesis, in *ICML* (2016)

[Rezende 14] Rezende, D. J., Mohamed, S., and Wierstra, D.: Stochastic Backpropagation and Approximate Inference in Deep Generative Models, in *ICML* (2014)

[Simonyan 14] Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *CoRR*, Vol. abs/1409.1556, (2014)

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, in *NIPS* (2014)

[Tu 16] Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H.: Modeling Coverage for Neural Machine Translation, in *ACL* (2016)

[Xu 15] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, in *CVPR* (2015)

[Zhang 16] Zhang, B., Xiong, D., and Su, J.: Variational Neural Machine Translation, in *EMNLP* (2016)