解説

深層学習における アテンション技術の最新動向

Latest Trends of Attention Mechanisms in Deep Learning

西田京介　斉藤いつみ

Abstract

本稿では，ニューラルネットワークの基礎的な知識を有する読者を想定し，深層学習における重要な要素技術であるアテンションについて解説する．まず，アテンションとはどのような技術かについて概要を説明した後に，アテンション技術が注目される契機となった機械翻訳と画像キャプション生成のニューラルネットワークにおける導入例を説明する．そして，最新動向として，アテンションの階層化・並列化の例や，自然言語処理分野における最新のトピックであるセルフアテンションについて紹介を行う．

キーワード：深層学習，アテンション，ニューラルネットワーク

1. はじめに

アテンション技術の研究は深層学習が大きな注目を集める前から行われており，主に視覚的な注意，すなわち，画像処理において画像のどの部分に着目すればよいかについて取り組まれてきた[(1)]．その後，2014年にBahdanauらにより発表された機械翻訳におけるアテンション機構[(2)]がその高い精度と汎用性により大きな注目を集め，画像キャプション生成[(3)]，音声認識[(4)]，質問応答[(5)]，機械読解[(6)]をはじめとした様々なタスクで応用されるなど，深層学習の発展に大きく貢献している．

アテンションについて明確な定義は存在しないが，一般的には入力されたデータのうち，どの部分を重視するかを決定する手段の総称である．近年では，Bahdanauらにより導入されたメカニズムを指す場合が多い．具体的には，$\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T\}$をベクトル集合，$\boldsymbol{\alpha}=[\alpha_1, \alpha_2, \cdots, \alpha_T]^\top$を各ベクトルに対するアテンションの分布（重み；$\sum_i \alpha_i=1$）としたとき，式(1)のように$\boldsymbol{h}$の重み付総和であるコンテキストベクトル$\boldsymbol{c}$を出力する単純な計算処理である．

$$\boldsymbol{c}=\sum_{i=1}^{T}\alpha_i \boldsymbol{h}_i \qquad (1)$$

ここで，ベクトル$\boldsymbol{h}_i$に対するアテンションの重みα_iは，

$$\alpha_i=\frac{\exp(S(\boldsymbol{h}_i, \boldsymbol{s}))}{\sum_{k=1}^{T}\exp(S(\boldsymbol{h}_k, \boldsymbol{s}))} \qquad (2)$$

のように，ベクトル$\boldsymbol{h}_i$とクエリ$\boldsymbol{s}$を引数とするスコア関数Sの値に基づいて計算される．$\boldsymbol{s}$の表現やSの計算方法はタスクやネットワークにより異なるが，式(1)，(2)は多くの研究で共通である．

アテンションが大きな注目を集めたのは，ニューラルネットワークにおいて入出力間の関係性を学習するにあたり，入力データを一つの固定長のベクトル表現に変換することなく，可変長のデータ全体を保持したまま入力データの各要素と出力データとの関係性を学習することを可能にしたためである．アテンションはテキストをはじめとした系列データを扱う再帰的ニューラルネットワーク（RNN）への導入について盛んに研究が行われているが，系列データ以外にも利用可能な汎用的な技術である．

2. アテンション技術の利用例

本章では，アテンション技術が注目される契機となっ

西田京介　正員　日本電信電話株式会社NTTメディアインテリジェンス研究所
E-mail　nishida.kyosuke@lab.ntt.co.jp
斉藤いつみ　日本電信電話株式会社NTTメディアインテリジェンス研究所
E-mail　saito.itsumi@lab.ntt.co.jp
Kyosuke NISHIDA, Member and Itsumi SAITO, Nonmember (NTT Media Intelligence Laboratories, NIPPON TELEGRAPH AND TELEPHONE CORPORATION, Yokosuka-shi, 239-0847 Japan).
電子情報通信学会誌　Vol.101 No.6 pp.591-596 2018年6月

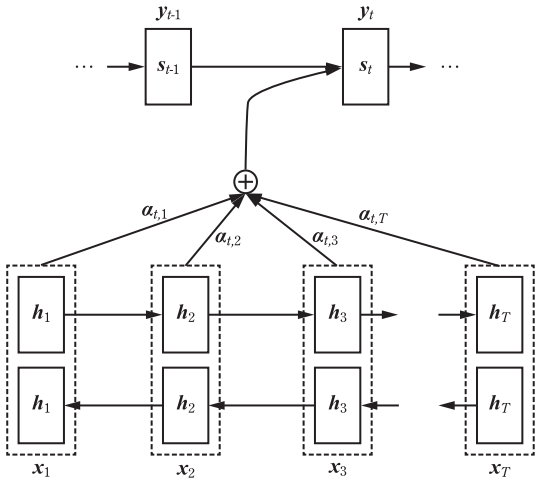


図1　Bahdanau らのニューラル翻訳モデル　元文 $\boldsymbol{x}$ を入力とした双方向 RNN（エンコーダ）の隠れ状態 $\boldsymbol{h}$ からアテンション $\boldsymbol{\alpha}_{ti}$ に基づいて翻訳文の単語 $\boldsymbol{y}_t$ を生成する．（出典：文献(2)Fig. 1）

た研究である Bahdanau らによる機械翻訳[2]と Xu らによる画像キャプション生成[3]について説明する．

2.1　機械翻訳

本節では，Bahdanau らによって提案されたアテンションを考慮するニューラル機械翻訳モデル[2]について説明する．アテンションを考慮するニューラル機械翻訳モデルは，2014 年に発表された従来のエンコーダ-デコーダ（符号器-復号器）モデル[7]を拡張するモデルとなっている．図 1 に概要を示す．

・従来モデル

入力系列 $\boldsymbol{x}$ を固定長ベクトルに符号化するエンコーダと，出力系列 $\boldsymbol{y}$ を予測するデコーダの二つの構成要素から成る．

まず，エンコーダは長さ T の入力ベクトル系列 $\boldsymbol{x}=\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$ を受け取り，RNN を用いて隠れ状態系列 $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}$ に変換する．そして，入力長分の隠れ状態ベクトルを，非線形変換 q により固定長ベクトル $\boldsymbol{c}$ に変換する．

$$\boldsymbol{h}_i=\mathrm{RNN}(\boldsymbol{x}_i, \boldsymbol{h}_{i-1}) \tag{3}$$
$$\boldsymbol{c}=q(\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}) \tag{4}$$

ここで，$q(\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\})=\boldsymbol{h}_T$ のように最終の隠れ状態を利用することが一般的である．

デコーダは，エンコーダが出力したコンテキストベクトル $\boldsymbol{c}$ と，$t-1$ 番目までの出力単語系列 $(y_1, \cdots, y_{t-1})$ を用いて t 番目の出力単語 y_t を予測する．各位置 t における出力単語 y_t の生成確率は下記の条件付確率で表すことができる．

$$p(y_t|y_1, \cdots, y_{t-1}, \boldsymbol{x})=g(y_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}) \tag{5}$$

ここで，$\boldsymbol{c}$ はエンコーダによって符号化された入力系列の固定長ベクトルを表し，$\boldsymbol{s}_t$ は翻訳文の位置 t における RNN の隠れ状態を表す．g は複数回の非線形変換を行うニューラルネットで表現される．

・アテンション考慮型モデル

次に，アテンションを考慮したエンコーダ-デコーダモデルについて説明する．RNN 型のエンコーダとデコーダから成るという点については従来のエンコーダ-デコーダモデルと同様の構成をしているが，入力側のコンテキストベクトルを一つの固定長ベクトルに変換するのではなく，翻訳文の単語位置に応じて異なるベクトルを計算する．具体的には，下記の式で表すことができる．

$$p(y_t|y_1, \cdots, y_{t-1}, \boldsymbol{x})=g(y_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}_t) \tag{6}$$

従来のエンコーダ-デコーダモデルとの違いは，式(5)におけるコンテキストベクトル $\boldsymbol{c}$ が，式(6)では翻訳文の位置 t に依存するベクトル $\boldsymbol{c}_t$ となった点である．$\boldsymbol{c}_t$ は下記の式を用いて計算する．

$$e_{ti}=S(\boldsymbol{h}_i, \boldsymbol{s}_{t-1}) \tag{7}$$
$$\alpha_{ti}=\frac{\exp(e_{ti})}{\sum_{k=1}^{T}\exp(e_{tk})} \tag{8}$$
$$\boldsymbol{c}_t=\sum_{i=1}^{T}\alpha_{ti}\boldsymbol{h}_i \tag{9}$$

ここで，α_{ti} はアテンションの重みを表しており，翻訳文の t 番目の要素を生成する際に，入力文に対してどのような重み付けを行うかを決定する役割を担っている．これは，t 番目の単語を生成する際に入力文のどの単語に着目すればよいかという入力文と翻訳文のソフトなアライメント情報も同時に学習していると捉えることができる．図 2 に，翻訳文の t 番目の単語を生成する際のアテンションのイメージ図を示す．図 2 では，α_{ti} の値がグレースケールで表されており，白色に近いほどアテンションスコアが高いことを表す．

なお，アテンションスコア関数 S については，$\boldsymbol{h}_i$ と $\boldsymbol{s}_{t-1}$ を入力として受け取って，スカラ e_{ti} を出力する多層パーセプトロンが利用される．スコア関数 S については，3.1 にて詳細を示す．

2.2　画像キャプション生成

本節では，Xu らにより提案されたアテンションを用いて画像のキャプションを生成するモデル[3]について説明する．このモデルでは機械翻訳と同様にエンコーダ-デコーダを用いており，機械翻訳からエンコーダが画像

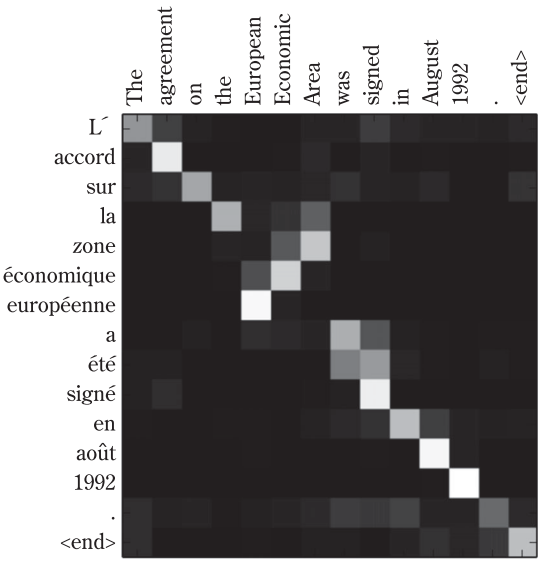


図 2　元文と翻訳文の単語間のアテンションの様子　横軸と縦軸はそれぞれ，元文（英語，エンコーダへの入力）と翻訳文（フランス語，デコーダの出力）の単語を表しており，各ピクセルは翻訳文の t 番目の単語を生成する際の元文の i 番目単語に対するアテンションの重み α_{ti} を表している．（出典：文献（2）Fig. 3(a)）

処理用に変更されている点を除けば本質的に同じモデルである．

具体的に，エンコーダは，学習済みの CNN を用いて，入力画像 $\boldsymbol{x}$ から T 個の部分領域ごとに D 次元の特徴量 $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}$ を抽出する．

次にデコーダは，キャプションの t 番目の単語 y_t の生成確率を，式(10)により決定する．g は複数回の非線形変換を行うニューラルネットで表現される．

$$p(y_t|y_1, \cdots, y_{t-1}, \boldsymbol{x}) = g(y_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}_t) \tag{10}$$

ここで，アテンションモデルは，Bahdanau らの機械翻訳モデル[(2)]と同様に，各部分領域 i に対して，次の単語を生成するために着目すべき領域の確率としてアテンションの重み α_{ti} を計算し，コンテキストベクトル $\boldsymbol{c}_t$ を求める．

$$e_{ti} = S(\boldsymbol{h}_i, \boldsymbol{s}_{t-1}) \tag{11}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{T} \exp(e_{tk})} \tag{12}$$

$$\boldsymbol{c}_t = \sum_{i=1}^{T} \alpha_{ti} \boldsymbol{h}_i \tag{13}$$

図 3　キャプション“woman is throwing a frisbee in a park.”生成時の各単語に対応するアテンションの様子　上段：入力画像，“A”，“woman”，“is”．中段：“throwing”，“a”，“frisbee”，“in”．下段：“a”，“park”，“.”．（出典：文献（3），Fig. 6(b)）

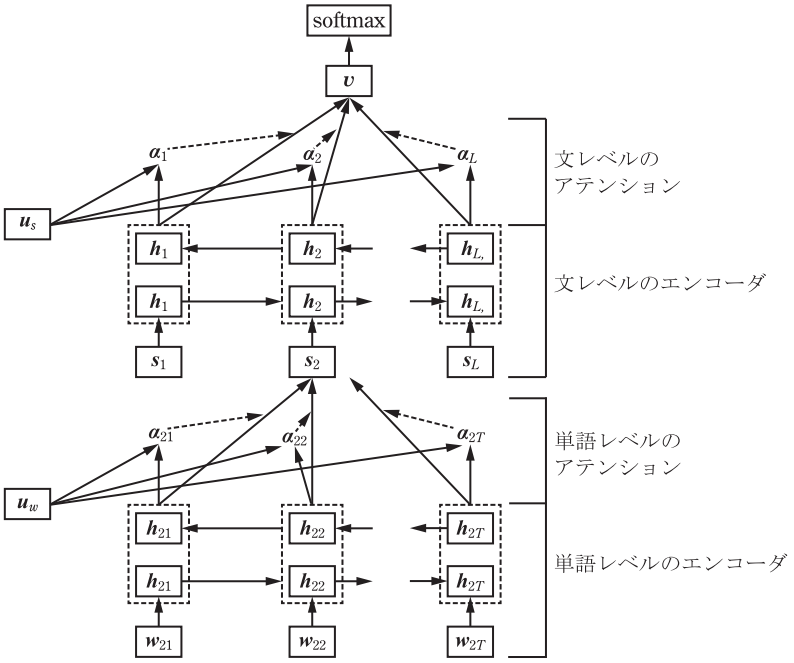


図4　文書分類における単語レベル，文レベルの階層アテンション構造　初めに単語レベルのアテンションに基づいて文ベクトル $\boldsymbol{s}_i$ を生成し，次に文レベルのアテンションに基づいて文書ベクトル $\boldsymbol{v}$ を生成する．（出典：文献(11)）

このようにアテンションを導入することによって，キャプション中の単語と画像の部分領域の関係性を直接捉えることで，精度良くキャプションを生成することができるようになる．

図3に，本モデルによってある画像（公園でフリスビーを投げている女性）を入力として生成したキャプション“A woman is throwing a frisbee in a park.”におけるアテンションの例を示す．6単語目の“frisbee”生成時はフリスビー，9単語目の“park”生成時は人物を除く背景部分に注意が向いていることが分かる．アテンション機構は画像キャプション生成においても大きな精度向上に貢献しており，アテンションの高い汎用性を示す結果となっている．

3. アテンション技術の最新動向

3.1　スコア関数

アテンションを行うメカニズムでは，式(1)，(2)に示すように，二つのベクトル $\boldsymbol{x}$，$\boldsymbol{y}$ を引数として受け取るスコア関数 $S(\boldsymbol{x}, \boldsymbol{y})$ を利用する．最もよく使われているスコア関数は下記に示す二つの関数の派生と見ることができる．

① 加法的関数[2]：$\boldsymbol{v}^\top \tanh(W_1\boldsymbol{x} + W_2\boldsymbol{y})$
② 乗法的関数[8]：$(W_1\boldsymbol{x})^\top(W_2\boldsymbol{y})$

ここで，$\boldsymbol{v}, W_1, W_2$ はニューラルネットワークの訓練されるパラメータである．また，tanh はベクトルの各要素ごとに適用される．

多層パーセプトロンは加法的関数と見ることができ，内積 $(\boldsymbol{x}^\top\boldsymbol{y})$ や bilinear $(\boldsymbol{x}^\top W\boldsymbol{y})$ は乗法的関数に含まれる．Britz らは機械翻訳タスクにおいて，この二つの方式について比較を行ったところ加法的関数の方が僅かに優れていることを実験的に示した[9]．また，$W_1\boldsymbol{x}$ の計算後の次元数 k については精度に大きな影響を与えないことを実験的に示した．一方で，Vaswani らの指摘のように，乗法的関数は高度に最適化された行列計算ライブラリが利用できるため，加法的関数に比べて高速に計算できるメリットがある[10]．精度についても，$\frac{1}{\sqrt{k}}(W_1\boldsymbol{x})^\top(W_2\boldsymbol{y})$ のように，係数 $\frac{1}{\sqrt{k}}$ を乗ずることで加法的関数と同程度に改善されるとの報告がある[10]．

現状では，どのスコア関数を用いるべきかについては結論が出ておらず，タスクやネットワークごとに実験的な評価を行って設定されることが多い．

3.2　階層化・並列化

アテンション技術の更なる発展として，アテンションの階層化や構造化などの手法が文書分類[11]や画像に対する質問応答[12,13]など様々なタスクにおいて提案されている．

文書分類におけるアテンションの階層化[11]では，文書を表現するベクトルを生成する際に，単語レベルのアテンションと文レベルのアテンションを階層的に用いる

手法を示した．図4に文献(11)で用いられている階層アテンション構造を示す．Yangら[11]は，まず単語レベルのアテンション α_{ij} と単語レベルの隠れ層 $\boldsymbol{h}_{ij}$ を用いて文ベクトル $\boldsymbol{s}_i=\sum_j \alpha_{ij}\boldsymbol{h}_{ij}$ を生成し，次に文レベルのアテンション α_i と文レベルの隠れ層 $\boldsymbol{h}_i$ を用いて文書レベルのベクトル $\boldsymbol{v}=\sum_i \alpha_i \boldsymbol{h}_i$ を生成する．このように，異なるレベルの情報に対して階層的なアテンション構造を用いることで文書という構造を持った長いテキストに対しても重要な要素を考慮しながらベクトル化を行うことができる．

画像に対する質問応答（VQA : Visual Quetion Answering）ではYangら[12]が多層のアテンション構造を用いるモデルを提案している．VQAでは，対象となる画像に対して画像中の要素に関して理解を問う問題が自然文で与えられ，正解を自然文で回答する．画像中の要素に関しての質問が与えられるため，自然文の質問と対象となる画像の双方を考慮しながら回答を生成する必要がある．Yangらの研究[12]では，アテンションを多層化することにより，多段階の推論に基づく質問応答を可能にしている．例えば“What are sitting in the basket on a bicycle?”という質問に対してまず初めのアテンション層では文中に出現する要素“basket”，“bicycle”など質問文に出現する物体を広く注目し，次のアテンション層で回答となる“dog”に注目することができる．

また，アテンションの並列化についても近年検討されている[10]．並列化においては，独立に学習を行うアテンション機構を H 個用意して，一つの入力ベクトル集合に対して複数個のコンテキストベクトル $\{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_H\}$ を獲得する．そして，式(14)のように，獲得したコンテキストベクトル集合を連結したベクトルを線形変換して最終的に一つのコンテキストベクトル $\boldsymbol{c}'$ を得る．

$$\boldsymbol{c}' = W[\boldsymbol{c}_1^\top, \cdots, \boldsymbol{c}_H^\top]^\top \tag{14}$$

ここで，$W \in \mathbb{R}^{d_{\text{model}} \times Hd_v}$ は学習パラメータ，d_v は各コンテキストベクトルのサイズ，d_{model} は最終的に得られたコンテキストベクトル $\boldsymbol{c}'$ のサイズである．このような並列化により，一つの入力データに対して複数の観点でアテンションを行うことが可能となる．

3.3 セルフアテンション

セルフアテンションは主に自然言語処理の研究において近年用いられる技術であり，機械読解[14),(15)]，含意認識[16]，機械翻訳[10]，文埋込[17]など様々な分野で用いられている．通常はアテンションの対象となる情報源に対して，別の情報をクエリとして用いるところを，セルフアテンションでは同一の情報源のみを使ってアテンションを行う．例えば，機械翻訳において，通常は原文に対して訳文の単語をクエリとして用いるところを，セルフアテンションでは原文に対して原文の単語をクエリとして用いる．これによって，同一文中で遠く離れた単語の関係性を理解しやすくなるメリットがある．

セルフアテンションを加法的関数を用いてベクトル集合 $\{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}$ に適用する場合，式(15)のように各トークン $\boldsymbol{x}_j$ をクエリとしてスコア関数を計算する．

$$S(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{v}^\top \tanh(W_1\boldsymbol{x}_i + W_2\boldsymbol{x}_j) \tag{15}$$

そして，クエリ単語 $\boldsymbol{x}_j$ ごとにコンテキストベクトル $\boldsymbol{c}_j$ を下記のように求め，

$$\alpha_{ji} = \frac{\exp(S(\boldsymbol{x}_i, \boldsymbol{x}_j))}{\sum_k \exp(S(\boldsymbol{x}_k, \boldsymbol{x}_j))} \tag{16}$$

$$\boldsymbol{c}_j = \sum_i \alpha_{ji}\boldsymbol{h}_i \tag{17}$$

新しいベクトル集合 $\{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_T\}$ を獲得する．このセルフアテンション後のベクトル集合に対して，通常のアテンションを更に掛けるような処理も可能である．

また，式(18)のように，クエリを利用しないアテンション[17]も一種のセルフアテンションとみなすことができる．

$$S(\boldsymbol{x}_i) = \boldsymbol{v}^\top \tanh(W_1\boldsymbol{x}_i) \tag{18}$$

この場合の出力は

$$\alpha_i = \frac{\exp(S(\boldsymbol{x}_i))}{\sum_k \exp(S(\boldsymbol{x}_k))} \tag{19}$$

$$\boldsymbol{c} = \sum_i \alpha_i \boldsymbol{h}_i \tag{20}$$

のように，一つのベクトルとして得られる．文書分類のように，データが単一の情報源から得られるタスクではこちらの形式が利用される場合がある[11]．

4. おわりに

本稿では，アテンション技術について概要及び最新動向を示した．アテンションはニューラルネットワークの重要な要素技術として現在も創意工夫が続けられている．特に，系列データを扱う再帰的ニューラルネットワークとの組合せにおいては必要不可欠な技術に発展している．その一方で，スコア関数の選択やセルフアテンション・階層化の利用など，アテンション技術をどのようにニューラルネットワークに組み込むかについては実験的に決定されることが多く，今後の研究の進展が注目される．

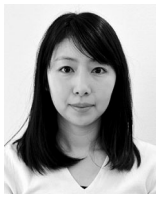
文　　献

（1） H. Larochelle and G.E. Hinton, “Learning to combine foveal glimpses with a third-order boltzmann machine,” NIPS, pp. 1243-1251, 2010.
（2） D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” ICLR, 2015.
（3） K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, and Y. Bengio, “Show, attend and tell : Neural image caption generation with visual attention,” ICML, pp. 2048-2057, 2015.
（4） J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” NIPS, pp. 577-585, 2015.
（5） S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, “End-to-end memory networks,” NIPS, pp. 2440-2448, 2015.
（6） K.M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to read and comprehend,” NIPS, pp. 1693-1701, 2015.
（7） K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” EMNLP, pp. 1724-1734, 2014.
（8） T. Luong, H. Pham, and C.D. Manning, “Effective approaches to attention-based neural machine translation,” EMNLP, pp. 1412-1421, 2015.
（9） D. Britz, A. Goldie, M. Luong, and Q.V. Le, “Massive exploration of neural machine translation architectures,” CoRR, abs/1703.03906, 2017.
（10） A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” NIPS, pp. 6000-6010, 2017.
（11） Z. Yang, D. Yang, C. Dyer, X. He, A.J. Smola, and E.H. Hovy, “Hierarchical attention networks for document classification,” NAACL-HLT, pp. 1480-1489, 2016.
（12） Z. Yang, X. He, J. Gao, L. Deng, and A.J. Smola, “Stacked attention networks for image question answering,” CVPR, pp. 21-29, 2016.
（13） J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” NIPS, pp. 289-297, 2016.
（14） J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” EMNLP, pp. 551-561, 2016.
（15） W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, “Gated self-matching networks for reading comprehension and question answering,” ACL, pp. 189-198, 2017.
（16） A.P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, “A decomposable attention model for natural language inference,” EMNLP, pp. 2249-2255, 2016.
（17） Z. Lin, M. Feng, C.N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” ICLR, 2017.

（平成 30 年 1 月 15 日受付　平成 30 年 1 月 24 日最終受付）

西田（にしだ）　京介（きょうすけ）（正員）

2008 北大大学院情報科学研究科博士課程了．博士（情報科学）．2006-2009 JSPS 特別研究員．2009 日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所主任研究員．2017 DBSJ 上林奨励賞受賞．AI・データマイニングに関する研究開発に従事．



斉藤（さいとう）　いつみ

2010 早大・理工・社会環境卒．2012 東大大学院工学系研究科博士前期課程了．同年，日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所研究員．自然言語処理に関する研究開発に従事．