# 自然言語処理 レポート

**所属: 工学部 計数工学科**

**学生証番号: XX-XXXXXX**

**氏名: 佐藤 瞭**

**連絡先: XXXXXX**

# 選択した論文

Minh-Thang Luong, Hieu Pham, Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In EMNLP.

テーマは機械翻訳。

# 選んだ理由

論文を決め兼ねて機械翻訳の情報を漁っているときに、Sequence-to-Sequenceモデルに関する情報を見つけ、そのままでは長文の翻訳に弱いが、attentionメカニズムを利用することで精度を改善できる、という趣旨の記述を見かけた。また、GoogleがAttention is All You Need[2]という論文を発表して、attentionにより翻訳精度を大幅に改善させたという情報も得た。このattentionのしくみを理解したいと思い、（Attention is All You Needは指定カンファレンス外だったのもあるが）attentionを用いた機械翻訳の論文で昔のものを読めば基本的な部分を学べるのではないかと考え、Effective Approaches to Attention-based Neural Machine Translationを読むことにした。

# 要約

`(*)` がついている用語は、レポートの「用語」の章で解説してあるので、必要に応じて参照されたい。

## どんなものか

attentionメカニズムを用いてニューラル機械翻訳 (NMT, neural machine translation)`(1)` の精度を改善した。

## 先行研究と比べて優れているところ

- 精度の高さ: 2014年、2015年の、WMT(2)の英⇆独翻訳のタスクで最高得点を記録した。
    - 既知の手法を組み合わせた(、attention機構を使わない)モデルよりも、最大で 5.0 BLEU(3)高いスコアを記録した。
- attentionを利用したNMTの効率の良い構造を探した

# 背景

- NMTが(論文の時点で)最高の性能を発揮してきている
    - NMTの考え方はシンプルで、文の単語を一つずつ入力していき、文の終わりを示すシンボルが入力されたら訳文の単語を1つずつ出力していく、という流れになっている
- 一方で、(機械翻訳のタスクに限らず)ニューラルネットワークの学習でattentionの考え方が人気になってきている
    - 論文中の例では、画像とエージェントの行動の対応、音声の文字起こし、画像の説明文生成など
- NMTにもattentionを適用した研究はあるが、効率の良い構造を探す研究は(論文の時点では)行われていない(と主張している)
- attention mechanism: 系列を入力として系列を出力とするSequence-to-Sequenceモデルにattention層を加えて、Decode時に入力系列の情報を直接参照できるようになったニューラルネットワークの構造。Sequence-to-Sequenceモデルでは、入力系列をあるサイズのベクトルにまとめたものをDecoderが参照するため、入力系列が長いと情報の圧縮精度が落ち、翻訳精度も下がってしまう[3]。attention層を加えることで、Decode時のある地点で必要な入力の情報を直接参照できるようになるため、長文に対する翻訳精度を改善できる。

# 提案手法の要点

global attention(図1)、local attention(図2)の2つが提案されている。図の$\boldsymbol{a_t}$、$\boldsymbol{c_t}$を求めている部分からもわかるように、両者は注目する入力系列の部分が異なる。

## global attention

方針: 各単語の出力時に、入力単語すべてに注目する。

各出力時$t$に、その時の(出力側の)隠れ層$\boldsymbol{h_t}$をパラメータとして、入力系列側の隠れ層$\overline{\boldsymbol{h}}_s$全ての加重平均をとる($\boldsymbol{a_t}$が重み)。このときに得られたベクトル$\boldsymbol{c_t}$を「文脈ベクトル」と呼び、この時の出力の隠れ層$\boldsymbol{h_t}$と合わせて出力$\overline{\boldsymbol{h}}_t$を予測する。

## local attention

方針: 各単語の出力時に、入力単語の一部のみに注目する。

各出力時$t$に、その出力の位置$t$とパラメータを用いて、入力のどの位置を中心に注目するかを決める(このパラメータは学習の対象)。求める位置$p_t$は、(入力系列の長さ)×($\mathrm{sigmoid}(\boldsymbol{h_t}$とパラメータの演算結果))と計算するので、0から入力系列の長さまでの値をとる。この位置$p_t$を中心に前後いくつかの入力系列の加重平均をとり、文脈ベクトルを計算する。このときの、「$p_t$を中心に前後いくつの幅で系列をみるか」というパラメータは経験的に選ばれる。

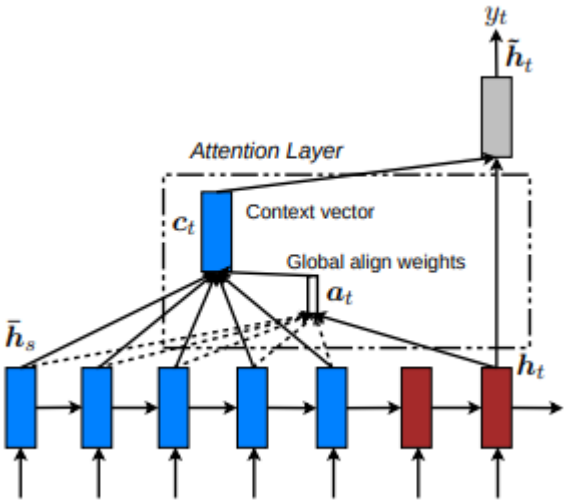


図1 global attention[1]

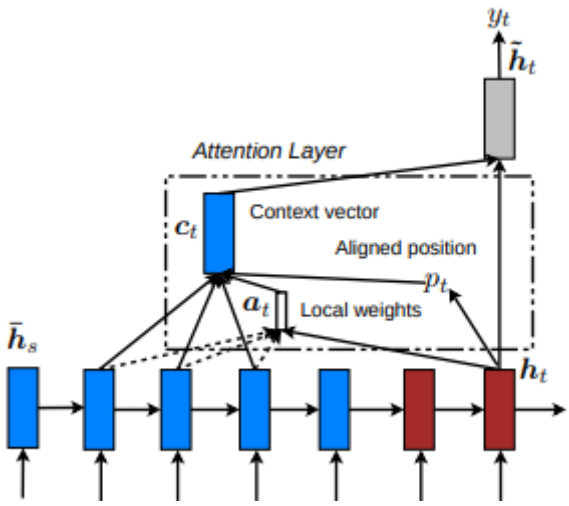


図2 local attention[1]

## 検証方法

2014年、2015年のWMTの英⇆独翻訳のタスクのデータセットを用いた。精度はBLEUで評価した。

既存のNMTの手法を組み合わせたモデルと、global attentionベースのMLT、local attentionベースのMLTで比較した。

2015年のWMTの独→英翻訳では（当時の）最高記録に及ばなかったが、2014年のWMTの英→独翻訳では（当時の）最高記録を出した。

## (論文内に記されている) 議論

- attentionalなモデルはnon-attentionalなモデルよりも学習が早い（学習曲線が下側にくる）
- attentionalなモデルは長文でもBLEUが悪化しない（non-attentionalなモデルは文が長くなるとBLEUが落ちてくる）
    - また、短い文章でもattentionalなモデルはnon-attentionalなモデルよりもBLEUが高い
- attentionalなモデルが正しく訳せて、non-attentionalなモデルが間違えた部分の例
    - 人名
    - 二重否定（"**not in**compatible"など）

→attentionalなモデルは、長文や人名などを訳す場合など、さまざまな場面でnon-attentionalなモデルよりも優れていると結論づけている

# 自らの意見、考察

- 翻訳する言語間の、語順の違いは精度にどの程度影響するのだろうか。例えば日本語は（主語→目的語→動詞）なのに対し英語は（主語→動詞→目的語）といった語順jの違いがある。文章が長くなると語順の差は複雑になっていくと思われるため、attentionalなモデルが語順の差に弱いとしたら、語順の違いが大きい言語間の翻訳は。文が長くなると落ちはじめるかもしれないのではないか。

    - 語順の違いが大きくなるとlocal attentionの位置$p_t$の最適化は難しくなるのではないか。語順が似ていれば訳文の位置に近い値をとれば済みそうな印象があるが、語順の違いが大きいと、たとえば訳文の頭の方の単語を予測するときに原文の末尾の方を読まなければならない、といった複雑な位置推定が求められるのではないかと考えているためである。(例として、「私は課題をやるつもりでは**ない**。」→「I will **not** do my assignment.」という翻訳では、notの出力のタイミングで日本語の文の最後を読まないと否定かどうかわからない。)
- 語順の違いが大きい場合の他、まとまった文章を文脈に即して訳すタスクをlocal attentionのモデルで行う際も、位置の最適化は複雑になると思われる。ある部分を訳す際に文脈によって訳が異なる際にはそれ以前の文章から文脈を取り出す必要があり、文章の書き方や内容によって文脈を示す情報の位置は異なってくると考えたためである。
- 独⇆英以外の翻訳では精度は検証しなくてもいいのだろうか?上に挙げたように語順など、言語の組に依存する性質が無視できないのではないか。
    - 言語の組ごとにベストなモデルを見つければ実用上はいいのかもしれない
- 論文中で、global attentionよりもlocal attentionのほうが計算量が小さいことが述べられている。学習曲線をみるとlocal attentionのほうが下側にあったので、計算量・学習での誤差の低さや収束の速さのどの観点でもlocal attentionのほうが優れているといえる。これは実用上でも好ましい性質だと思った。
- global attentionよりもlocal attentionのほうが学習パラメータの数が多い。しかし、local attentionのほうが学習時の目的関数の収束が速い。つまり、パラメータを増やしたほうが学習時の収束が速くなる場合もあるということだろうか?これは数学的に説明できるんだろうか?
- 2019年現在、attentionは機械翻訳においてどれくらい有用なんだろうか?
- よくわからなかった部分:
    - hard attentionモデルは微分不可能だが、local attentionモデルは微分可能である、という記述
        - モデルの微分可能性がどういった意味をもつのかよくわからなかった
        - local attentionは、加重平均のウエイトを全体に割り振るsoft attention(今回はglobal attentionが対応)と系列の1箇所にだけ注目するhard attentionの考え方を混ぜたモデルである、というところまではわかった
        - 誤差逆伝播で嬉しいことがあるのだろうか?連鎖律がつかえる、ということだろうか??

# 用語

(1) neural machine translation: ニューラルネットワークを用いた翻訳。入力系列をベクトル化するEncoderと、ベクトルを出力系列に変換するDecoderのRNNからなる。RNN (recurrent neural network)は、時系列データを入力とするニューラルネットワーク。

(2) WMT: 機械翻訳のワークショップ/カンファレンス。例年、会議に先だって機械翻訳のコンペティションが開催されている[4]。

(3) BLEU[5]: 機械翻訳の評価指標の一つ。「(熟達した)人間が作成した訳文に近い訳文ほど機械翻訳の精度は高い」という方針で設計されている。人間の作業による評価を必要としないので、評価の作業コストを小さくできる。**bi**l**i**ngual **e**valuation **u**nderstudyを縮めてBLEUと呼んでいる。

---

# 参考文献

[1] Minh-Thang Luong, Hieu Pham, Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In EMNLP.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. 2017. Attention Is All You Need. In NIPS.

[3] 松村 雪桜, 佐藤 貴之, 小町 守. 2017. 逆翻訳によるニューラル機械翻訳の最適化. In 言語処理学会 第23回年次大会 発表論文集.

[4] https://japan.zdnet.com/article/35137576/ 2019/06/28 閲覧.

[5] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2012. BLEU: a Method for Automatic Evaluation of Machine Translation. In ACL.