# BERT

浅川 伸一 asakawa@ieee.org

23/Mar/2020


- 1 GLUE leaderboard
- 2 GLUE 下位課題
- 3 SOTA モデルの特徴
- 4 BERT (2)
- 5 単語埋め込みモデルの問題点
- 6 2015 Google BERT (3)
- 7 従来モデルの問題点
- 8 モデル構成
- 9 BERT の入力表現
- 10 BERT の事前訓練: マスク化言語モデル
- 11 BERT の事前訓練: 次文予測課題
- 12 BERT: ファインチューニング
- 13 BERT モデルの詳細
- 14 BERT ファインチューニング手続き
- 15 BERT モデルサイズ比較
- 16 BERT モデル単方向，双方向モデル比較
- 17 BERT 事前訓練比較
- 18 文献
- 19 SOTA モデルの特徴
- 20 事前訓練とマルチ課題学習
- 21 Transformer: Attention is all you need
- 22 Transformer(2): Attention is all you need
- 23 Transformer(3): Attention is all you need
- 24 BERT, GPT, ELMo 事前訓練の違い
- 25 多言語対応
- 26 BERT の発展
- 27 埋め込みモデルによる構文解析
- 28 自然言語処理史略
- 29 要約
- 30 seq2seq model
- 31 Seq2seq (2)
- 32 Seq2seq (3)


## 1 GLUE leaderboard

- GLUE leaderboard
- SuperGlue leaderboard

## 2 GLUE 下位課題

- CoLA: 入力文が英語として正しいか否かを判定
- SST-2: スタンフォード大による映画レビューの極性判断
- MRPC: マイクロソフトの言い換えコーパス。2文 が等しいか否かを判定
- STS-B: ニュースの見出し文の類似度を5段階で評定
- QQP: 2 つの質問文の意味が等価かを判定
- MNLI: 2 入力文が意味的に含意，矛盾，中立を判定
- QNLI: Q and A
- RTE: MNLI に似た2つの入力文の含意を判定
- WNI: ウィノグラッド会話チャレンジ

## 3 SOTA モデルの特徴

- RoBERTa: BERT の訓練コーパスを巨大 (173GB) にし，ミニバッチサイズを大きした

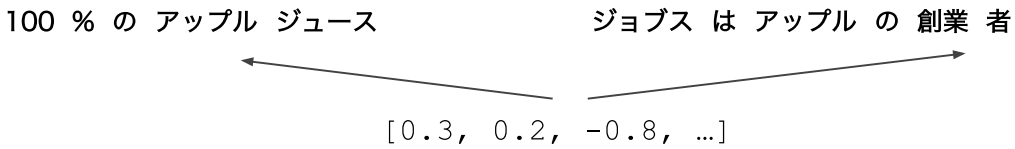
- XLNet: 順列言語モデル。2 ストリーム注意
- MT-DNN: BERT ベース の転移学習に重きをおいたモデル
- GPT-2: BERT に基づく。人間超えして 2019 年 2 月時点で炎上騒ぎ
- BERT: Transformer に基づく言語モデル。とに基づく，各下流課題を。事前訓練されたモデルは一般公開済。
- ELMo: 双方向 RNN による文埋め込み表現
- Transformer: 自己注意に基づく言語モデル。多頭注意，位置符号器.


## 4 BERT (2)

- 単語埋め込み (word2vec など) 単語の共起情報 点相互情報量 PMI

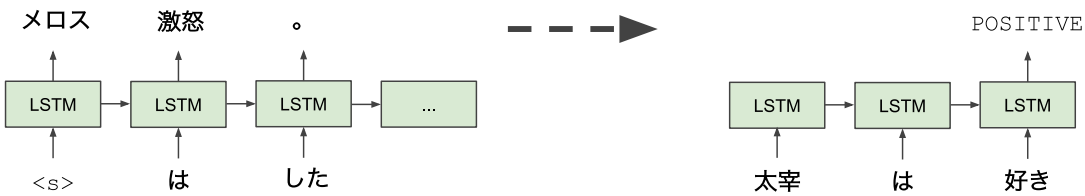

## 5 単語埋め込みモデルの問題点

- 文脈自由表現




- 文脈依存表現


## 6 2015 Google BERT (3)




## 7 従来モデルの問題点

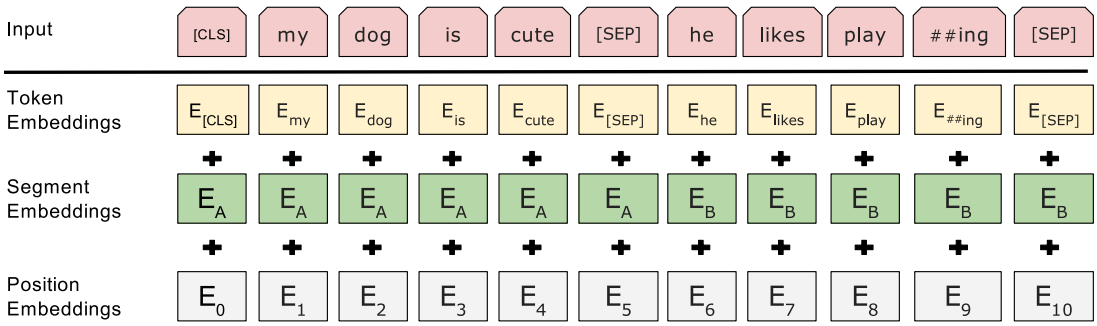
- Bahdanau, Luong らの注意は BiRNN を用いた双方向だが，他は RNN による単方向
- 単方向モデルと双方向モデルをどう扱ううか？

## 8 モデル構成

- 多頭自己注意 Multi-headed self attention
- フィードフォワードのみ採用
- 層正則化と残差コネクト
- 位置符号器

- トランスフォーマーと LSTM の相違
    1. 自己注意は局所依存性を持たない
- 長距離依存対策

# 9 BERT の入力表現




埋め込みトークンの総和，位置符号器，分離埋め込みの 3 者 From (Devlin et al. 2018) Fig. 2

# 10 BERT の事前訓練: マスク化言語モデル

全入力系列のうち 15% をランダムに [MASK] トークンで置き換える

- 入力はオリジナル系列を [MASK] トークンで置き換えた系列
- ラベル: オリジナル系列の [MASK] 部分にの正しいラベルを予測
- 80%: オリジナル入力系列を [MASK] で置換
- 10%: [MASK] の位置の単語をランダムな無関連語で置き換える
- 10%: オリジナル系列

# 11 BERT の事前訓練: 次文予測課題

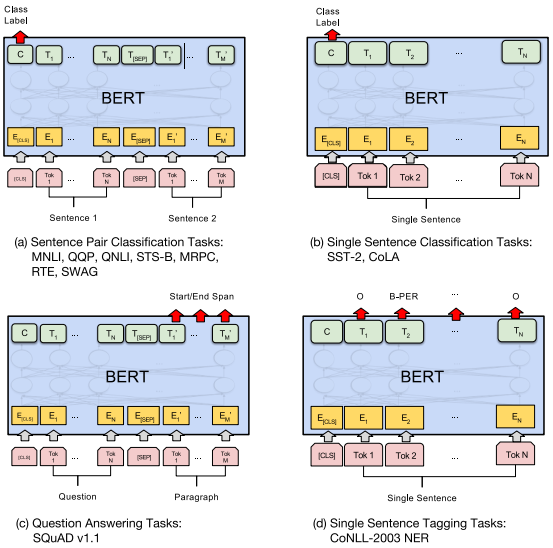
言語モデルの欠点を補完する目的，次の文を予測

[SEP] トークンで区切られた 2 文入力

- 入力: the man went to the store [SEP] he bought a gallon of milk.
- ラベル: IsNext
- 入力: the man went to the store [SEP] penguins are flightless birds.
- ラベル: NotNext

# 12 BERT: ファインチューニング

(a), (b) は文レベル課題，　(c),(d)はトークンレベル課題, E: 入力埋め込み表現, $T_i$: トークン $i$ の文脈表象。




From (Devlin et al. 2018) Fig.3

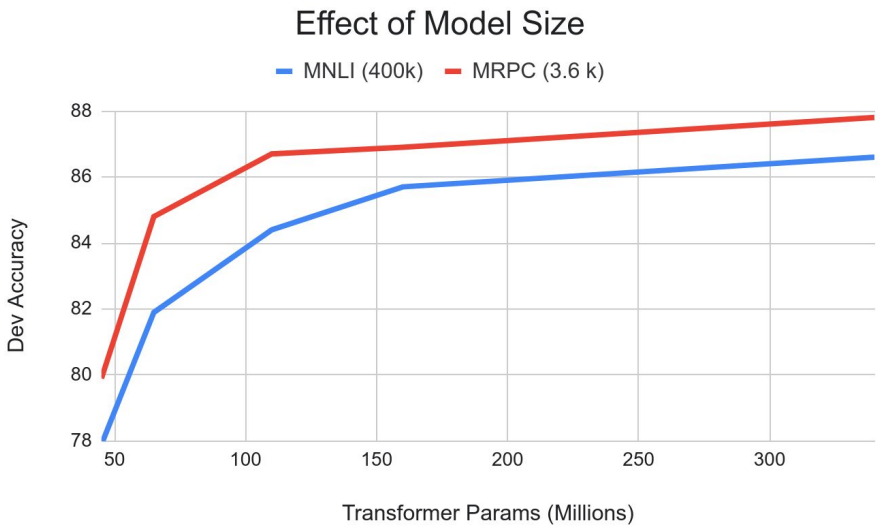
## 13 BERT モデルの詳細

- データ: Wikipedia (2.5B words) + BookCorpus (800M words)
- バッチサイズ: 131,072 words (1024 sequences * 128 length or 256 sequences * 512 length)
- 訓練時間: 1M steps (~40 epochs)
- 最適化アルゴリズム: AdamW, 1e-4 learning rate, linear decay
- BERT-Base: 12 層, 各層 768 ニューロン, 12 多頭注意
- BERT-Large: 24 層, 各層 1024 ニューロン, 16 多頭注意
- 4x4 / 8x8 TPU で 4 日間

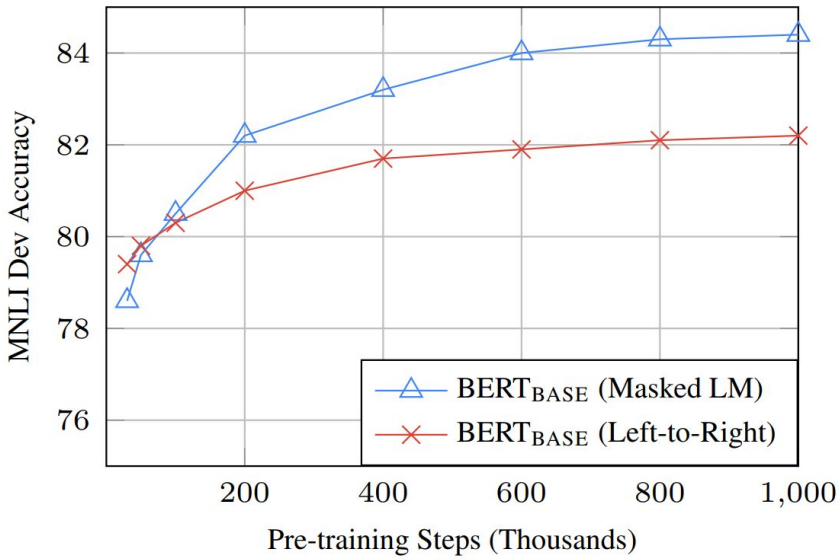
## 14 BERT ファインチューニング手続き

| Masking Rates | | | Dev Set Results | | |
|---|---|---|---|---|---|
| MASK | SAME | RND | MNLI | NER | |
| | | | Fine-tune | Fine-tune | Feature-based |
| 80% | 10% | 10% | 84.2 | 95.4 | 94.9 |
| 100% | 0% | 0% | 84.3 | 94.9 | 94.0 |
| 80% | 0% | 20% | 84.1 | 95.2 | 94.6 |
| 80% | 20% | 0% | 84.4 | 95.2 | 94.7 |
| 0% | 20% | 80% | 83.7 | 94.8 | 94.6 |
| 0% | 0% | 100% | 83.6 | 94.9 | 94.6 |

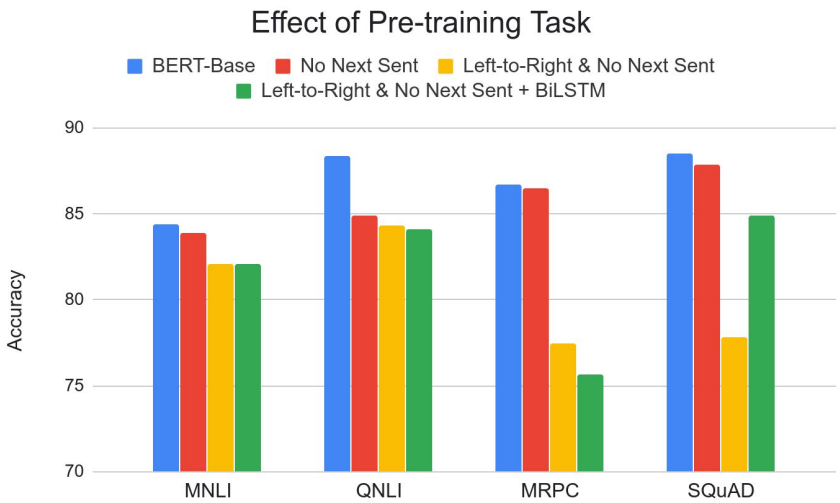
## 15 BERT モデルサイズ比較




## 16 BERT モデル単方向，双方向モデル比較




## 17 BERT 事前訓練比較



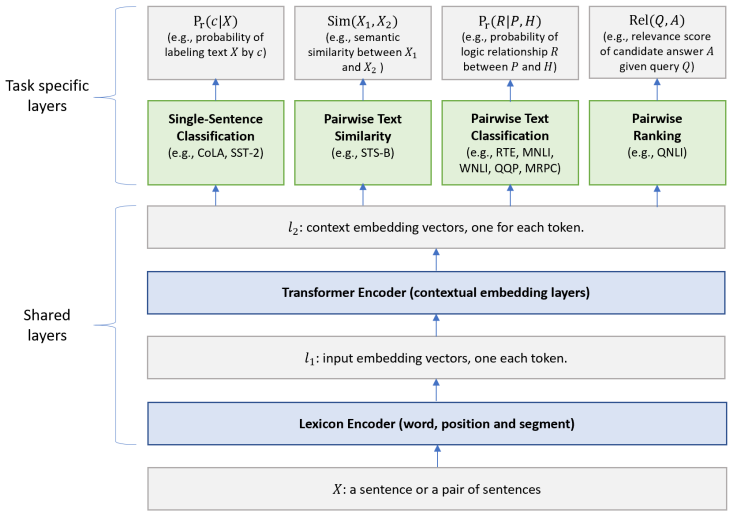

## 18 文献

1. BERT (Google) 論文 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding(Devlin et al. 2018)
2. GPT (OpenAI) 論文 Improving Language Understanding by Generative Pre-Training(Radford et al. 2018)
3. GPT-2 (OpenAI) ブログ Language Models are Unsupervised Multitask Learners
4. Transformer-XL (Google/CMU) 論文 Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context(Dai et al. 2019)
5. XLNet (Google/CMU) 論文 XLNet:Generalized Autoregressive Pretraining for Language Understanding(Yang et al. 2019)
6. XLM (Facebook) 論文 Cross-lingual Language Model Pretraining(Lample and Conneau 2019)
7. RoBERTa (Facebook), 論文 Robustly Optimized BERT Pretraining Approach(Y. Liu et al. 2019)
8. DistilBERT

# 19 SOTA モデルの特徴

- RoBERTa: BERT の訓練コーパスを巨大 (173GB) にし，ミニバッチサイズを大きした
- XLNet: 順列言語モデル。2 ストリーム注意
- MT-DNN: BERT ベース の転移学習に重きをおいたモデル
- GPT-2: BERT に基づく。人間超えして 2019 年 2 月時点で炎上騒ぎ
- BERT: Transformerに基づく言語モデル。**マスク化言語モデル** と **次文予測** に基づく 事前訓練，各下流課題をファインチューニング。事前訓練されたモデルは一般公開済。
- ELMo: 双方向 RNN による文埋め込み表現
- Transformer: 自己注意に基づく言語モデル。多頭注意，位置符号器.

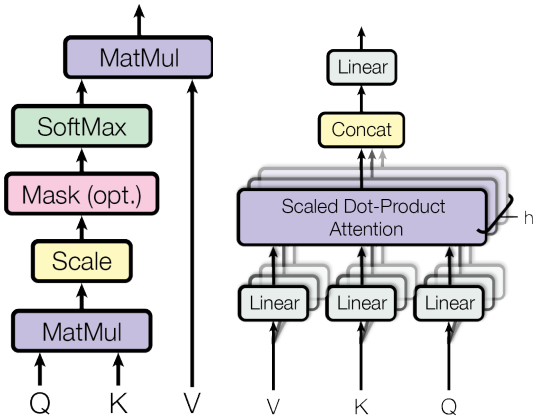
# 20 事前訓練とマルチ課題学習




From (X. Liu et al. 2019) Fig. 1

# 21 Transformer: Attention is all you need

$$attention(Q, K, V) = dropout\left(softmax\left(\frac{QK^\top}{\sqrt{d}}\right)\right)V \tag{1}$$




From (Vaswani et al. 2017) Fig. 2

# 22 Transformer(2): Attention is all you need

$$MultiHead(Q,K,V) = Concat\left(\underset{1}{head},\ldots,\underset{h}{head}\right)W^O \tag{2}$$

where, $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$

The projections are parameter matrices

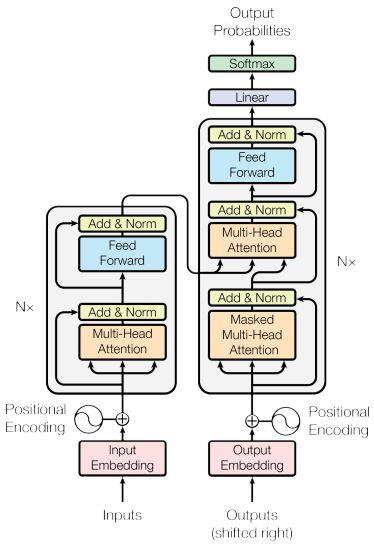
$W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$,
$W_i^K \in \mathbb{R}^{d_{model}\times d_k}$,
$W_i^V \in \mathbb{R}^{d_{model}\times d_v}$, and
$W^O \in \mathbb{R}^{hd_v\times d_{model}}$. $h = 8$,
$d_k = d_v = \frac{d_{model}}{h} = 64$

$$FFN(x) = \max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{3}$$

$$\underset{(pos,2i)}{PE} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{4}$$

$$\underset{(pos,2i+1)}{PE} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{5}$$

# 23 Transformer(3): Attention is all you need

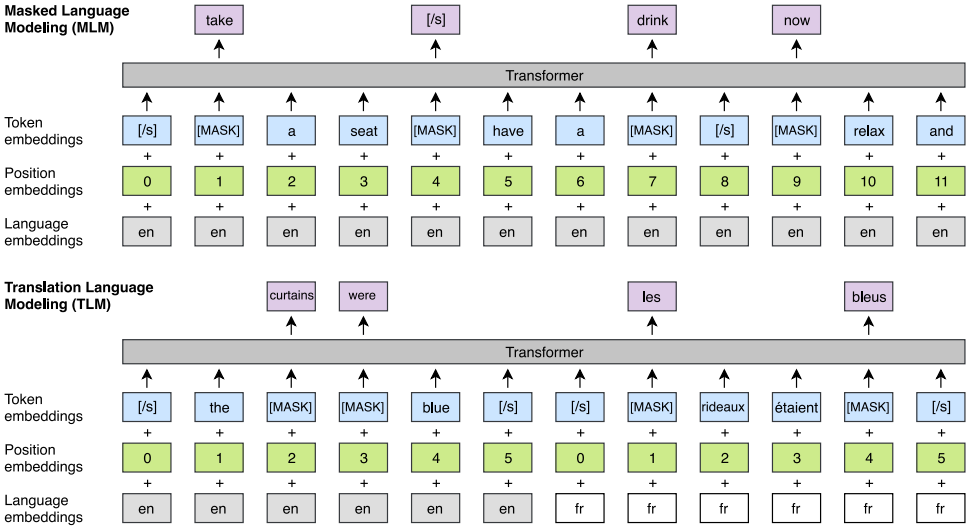



From (Vaswani et al. 2017) Fig. 1

# 24 BERT, GPT, ELMo 事前訓練の違い

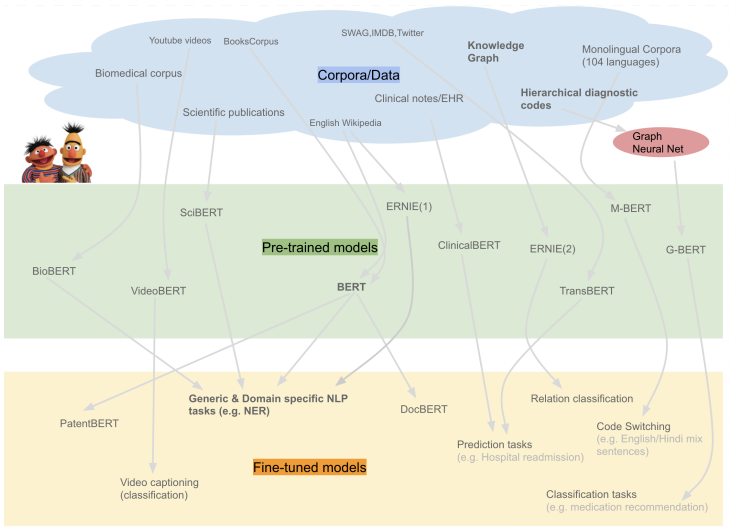
- BERT: トランスフォーマー，マスク化言語モデル，次文予測課題
- GPT: 順方向トランスフォーマー
- ELMo: 双方向 RNN による中間層の連結

# 25 多言語対応



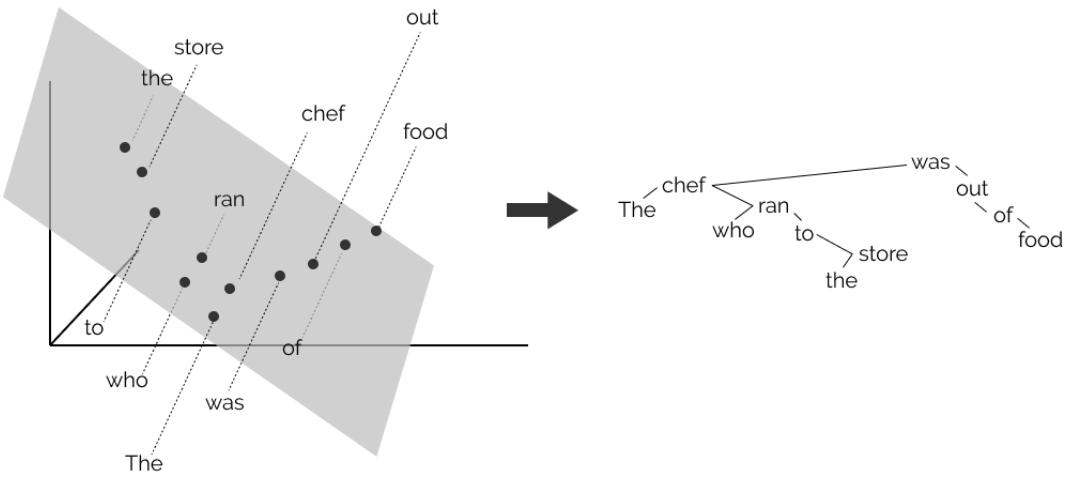

From (Lample and Conneau 2019) Fig. 1

# 26 BERT の発展




From https://towardsdatascience.com/a-review-of-bert-based-models-4ffdc0f15d58

# 27 埋め込みモデルによる構文解析




From https://github.com/john-hewitt/structural-probes

# 28 自然言語処理史略

- 第一世代: 1990年代まで
  1. 文法規則に基づく文の解析
  2. 構文解析アルゴリズム
  3. 文法の理論の発展

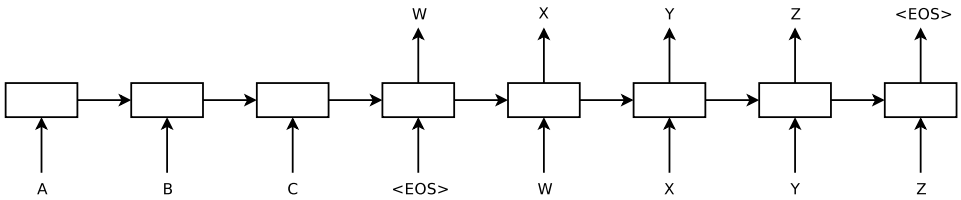
- 第2世代: 1990年以降
  1. 統計的自然言語処理
  2. 大規模データと機械学習

- 第3世代: 現在
  1. Deep Neural Networks を利用した研究

# 29 要約

- 表現学習 (Representation learning) と 埋め込み(Embedding)
  - 単語あるいは文をベクトルとして表現
  - 大規模コーパスから単語の意味表現の事前学習 distributional / distributed
- 再起型 ニューラルネット (Recurrent Neural Network)の利用
  - LSTM (Long Short-term Memory)，GRU (Gated Recurrent Unit)
- 注意 全体から，直接必要な情報を取得することが可能に
- End-to-end

# 30 seq2seq model

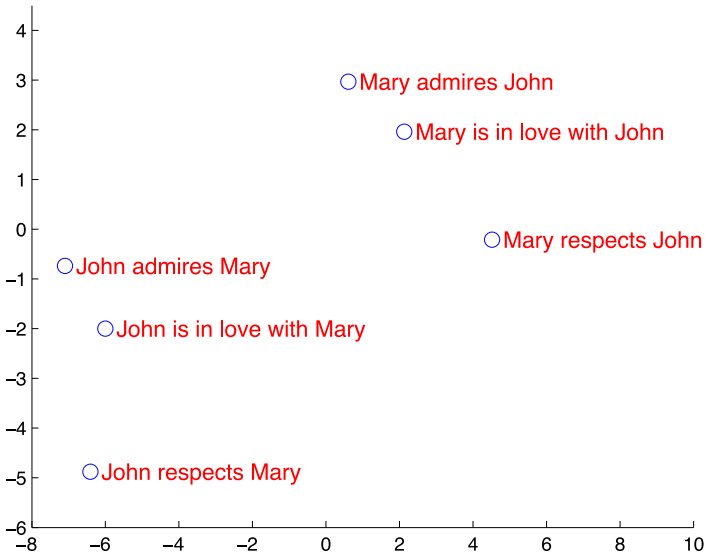



From (Sutskever, Vinyals, and Le 2014) Fig. 1 翻訳モデル seq2seq の概念図

“<eos>” は文末を表す。中央の “<eos>” の前がソース言語であり，中央の “<eos>” の後はターゲット言語の言語モデルである SRN の中間層への入力として用いる。

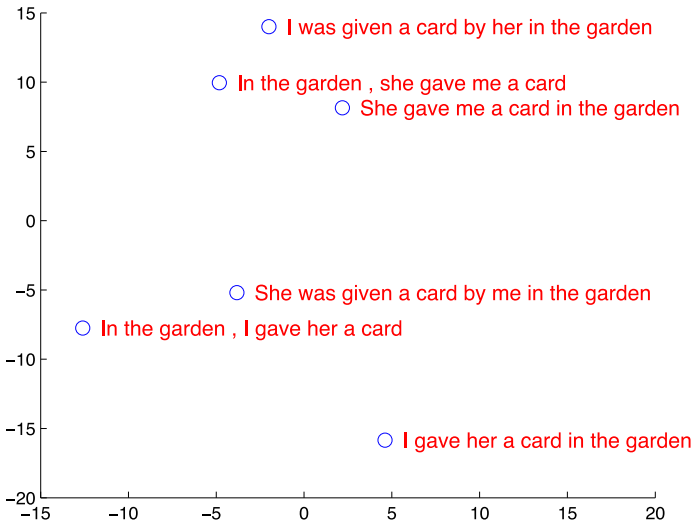
注意すべきは，ソース言語の文終了時の中間層状態のみをターゲット言語の最初の中間層 の入力に用いることであり，それ以外の時刻ではソース言語とターゲット言語は関係が ない。逆に言えば最終時刻の中間層状態がソース文の情報全てを含んでいると みなしうる。 この点を改善することを目指すことが 2014 年以降盛んに行われてきた。 顕著な例が後述する **双方向 RNN**，**LSTM** を採用したり，**注意** 機構を導入することであった。

# 31 Seq2seq (2)




From (Sutskever, Vinyals, and Le 2014) Fig. 2

# 32 Seq2seq (3)




From (Sutskever, Vinyals, and Le 2014) Fig. 2

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. “Transformer-Xl: Attentive Language Models Beyond a Fixed-Length Context.” *ArXiv Preprint* 1901.02860v3 [cs.LG].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *arXiv Preprint*.

Lample, Guillaume, and Alexis Conneau. 2019. “Cross-Lingual Language Model Pretraining.” *ArXiv Preprint* 1901.07291v1 [cs.CL].

Liu, Xiaodong, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. “Multi-Task Deep Neural Networks for Natural Language Understanding.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4487–96. Florence, Italy: Association for Computational Linguistics.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. “RoBERTa: A Robustly Optimized Bert Pretraining Approach.” *ArXiv Preprint*.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. “Improving Language Understanding by Generative Pre-Training.” *Https://S3-Us-West-2.amazonaws.com/Openai-Assets/Research-Covers/Language-Unsupervised/Language_understanding_paper.pdf*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. “Sequence to Sequence Learning with Neural Networks.” In *Advances in Neural Information Processing Systems (NIPS)*, edited by Zoubin Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, 27:3104–12. Montreal, BC, Canada.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. 2017. “Attention Is All You Need.” *arXiv Preprint*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. “XLNet: Generalized Autoregressive Pretraining for Language Understanding.” *ArXiv Preprint*, 1906.08237.