ニューラルネットに基づく機械翻訳

Graham Neubig
奈良先端科学技術大学院大学 (NAIST)
2015-9-15

I am giving a talk at Kyoto University

私 は 京都 大学 で 講演 を しています (終)

次の単語確率を推測

F = “I am giving a talk”

$P(e_1 = 私 \mid F) = 0.8$	$P(e_1 = 僕 \mid F) = 0.03$ $P(e_1 = 講演 \mid F) = 0.01$	...	$e_1 =$ 私
$P(e_2 = は \mid F, e_1) = 0.9$	$P(e_2 = が \mid F, e_1) = 0.09$	...	$e_2 =$ は
$P(e_3 = 講演 \mid F, e_{1,2}) = 0.4$	$P(e_3 = トーク \mid F, e_{1,2}) = 0.3$ $P(e_3 = 話 \mid F, e_{1,2}) = 0.03$	...	$e_3 =$ 講演
$P(e_4 = を \mid F, e_{1,3}) = 0.99$		...	$e_4 =$ を
$P(e_5 = しています \mid F, e_{1,4}) = 0.4$ $P(e_5 = 行っています \mid F, e_{1,4}) = 0.3$	$P(e_5 = している \mid F, e_{1,4}) = 0.15$ $P(e_5 = 行っている \mid F, e_{1,4}) = 0.1$	...	$e_5 =$ しています
$P(e_6 = (終) \mid F, e_{1,5}) = 0.8$	$P(e_6 = よ \mid F, e_{1,5}) = 0.1$	...	$e_6 =$ (終)

つまり、機械翻訳は

確率モデル

$$P(E|F)=\prod_{i=1}^{I+1} P(e_i|F, e_1^{i-1})$$

訳出過程

$i = 0$
while e_i is not equal to “(終)”:
    i ← i+1
    e_i ← argmax$_e$ P(e_i|F, $e_{1,i-1}$)

として定式化することができる

確率の推定法

翻訳モデル・言語モデル

翻訳モデル確率

$$P(E|F)=\prod_{i=1}^{I+1} P(e_i|F,e_1^{i-1})$$

↓ いったん入力を忘れて

言語モデル確率

$$P(E)=\prod_{i=1}^{I+1} P(e_i|e_1^{i-1})$$

問題：次の単語の確率 $P(e_i|e_1^{i-1})$ をどう計算する？

単語列の数え上げによる確率計算

$$P\left(e_i|e_1^{i-1}\right)=\frac{c\left(e_1^{i}\right)}{c\left(e_1^{i-1}\right)}$$

私 は 講演 を している </s>
私 の 勤め先 は 奈良 に ある </s>
奈良 は 大阪 に 近い </s>

P(は | <s> 私) = c(<s> 私 は)/c(<s> 私) = 1 / 2 = 0.5

P(の | <s> 私) = c(<s> 私 の)/c(<s> 私) = 1 / 2 = 0.5

数え上げの問題

- 頻度の低い現象に弱い：

学習：

私 は 講演 を している </s>
私 の 勤め先 は 奈良 に ある </s>
奈良 は 大阪 に 近い </s>

確率計算：

私 の 勤め先 は 大阪 に ある </s>

⇩

P(大阪 |<s> 私 の 勤め先 は) = 0/1 = 0

⇩

P(E= 私 の 勤め先 は 大阪 に ある </s>) = 0

履歴の制限： n-gram モデル

- **2-gram モデル：** 直前の 1 単語のみを利用

$$P(E) = \prod_{i=1}^{I+1} P(e_i | e_{i-1})$$

入力： 私 の 勤め先 は 大阪 に ある </s>

学習：

私 は 講演 を している </s>

私 の 勤め先 は 奈良 に ある </s>

奈良 は 大阪 に 近い </s>

- **3-gram, 4-gram, 5-gram** なども
  - + 精度が向上
  - - メモリ量、スパース性の問題が悪化

対数線形言語モデル [Chen+ 00] (1)

- より柔軟な確率計算法
- 履歴の単語に基づいて全単語のスコア **s** を計算

$$s\left(e_{i-n+1}^{i-1}\right) = \boldsymbol{b} + \sum_{k=1}^{n-1} \boldsymbol{w}_{k, e_{i-k}}$$

e_{i-2} = 勤め先
e_{i-1} = は

は
が
奈良
同僚
行う
…

$$\boldsymbol{b} = \begin{pmatrix} 3.0 \\ 2.5 \\ -0.2 \\ 0.1 \\ 1.2 \\ \cdots \end{pmatrix} \quad \boldsymbol{w}_{1,\text{は}} = \begin{pmatrix} -6.0 \\ -5.1 \\ 0.2 \\ 0.1 \\ 0.6 \\ \cdots \end{pmatrix} \quad \boldsymbol{w}_{2,\text{勤め先}} = \begin{pmatrix} -0.2 \\ -0.3 \\ 1.0 \\ 2.0 \\ 0.4 \\ \cdots \end{pmatrix} \quad \boldsymbol{s} = \begin{pmatrix} -3.2 \\ -2.9 \\ 1.0 \\ 2.2 \\ 2.2 \\ \cdots \end{pmatrix}$$

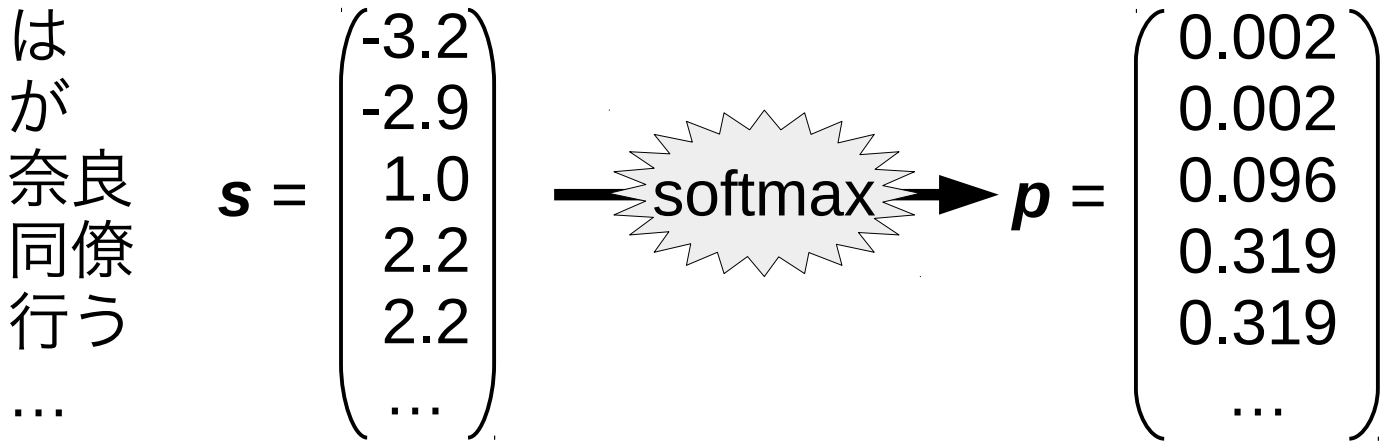
対数線形言語モデル [Chen+ 00] (2)

- 確率計算のため、スコアの指数を取り、正規化

$$p\left(e_i = x \middle| e_{i-n+1}^{i-1}\right) = \frac{e^{s\left(e_i = x \middle| e_{i-n+1}^{i-1}\right)}}{\sum_{\tilde{x}} e^{s\left(e_i = \tilde{x} \middle| e_{i-n+1}^{i-1}\right)}}$$

- ベクトルに対して行う際 softmax 関数とも言う

$$\boldsymbol{p}\left(e_i \middle| e_{i-n+1}^{i-1}\right) = \mathrm{softmax}\left(\boldsymbol{s}\left(e_i \middle| e_{i-n+1}^{i-1}\right)\right)$$



対数線形モデルの学習

- 確率的勾配降下法 (SGD) を利用することが多い
- 学習データの各単語 e_i に対してパラメータ w をどの方向に動かしたら正解の確率が良くなりそうかを計算

$$\delta = \frac{d}{d\,w} p\left(e_i \middle| e_{i-n+1}^{i-1}\right)$$ (尤度の勾配)

- これを学習率 α にかけてパラメータを更新

$$w \leftarrow w + \alpha\,\delta$$

問題：変数の相互作用をうまく表現できていない

勤め先	は	奈良	→	○	勤め先	の	奈良	→ △
勤め先	は	同僚	→	△	勤め先	の	同僚	→ ○

- 単純と足し合わせるだけでは表現不可。解決策は？
  - 「勤め先 は」などの単語列もパラメータ化：

奈良
同僚
…

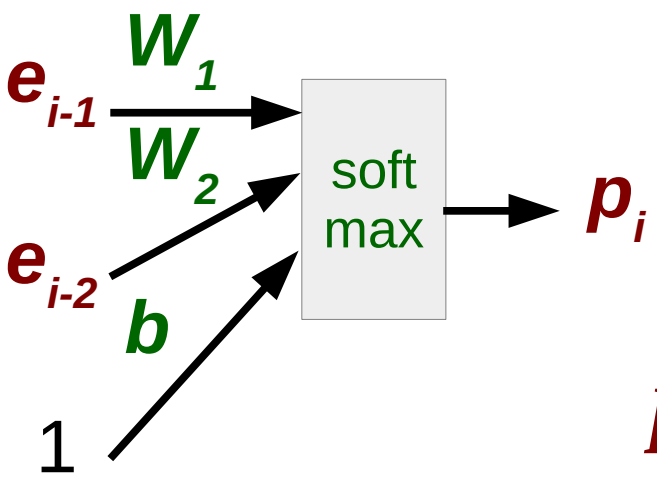
$$w_{2,1,\text{勤め先 , は}} = \begin{pmatrix} 2.0 \\ -2.1 \\ \cdots \end{pmatrix} \quad w_{2,1,\text{勤め先 , の}} = \begin{pmatrix} -1.2 \\ 2.9 \\ \cdots \end{pmatrix}$$

パラメータ数、メモリの爆発…

  - ニューラルネット！

ニューラルネット

対数線形モデルの概念図



$$p_i = \text{softmax}\left(b + \sum_{k=1}^{n-1} W_k e_{i-k}\right)$$

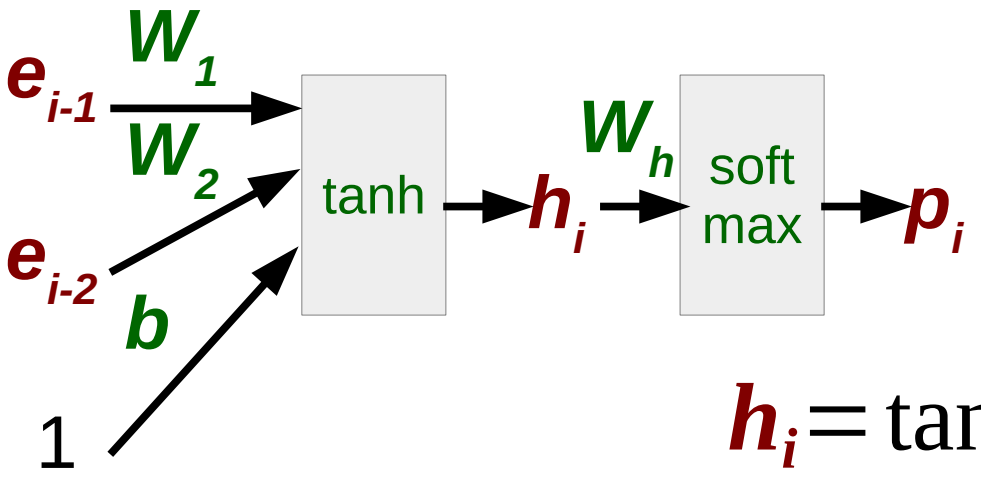
e_{i-1} と e_{i-2} は各単語に当たるだけが 1 の one-hot ベクトル

	は	が	奈良	同僚	勤め先	
e_{i-1} = {1,	0,	0,	0,	0,	...}
e_{i-2} = {0,	0,	0,	0,	1,	...}

W_1, W_2 は重み行列、 b は重みベクトル

ニューラルネット

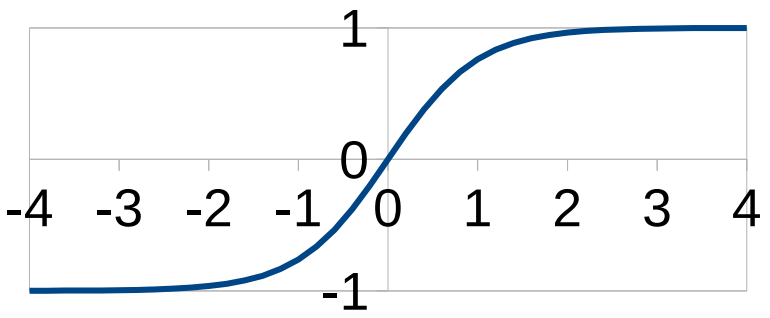
- 入力と出力の間に、非線形関数を計算する隠れ層を追加



$$h_i = \tanh\left(b + \sum_{k=1}^{n-1} W_k e_{i-k}\right)$$

$$p_i = \mathrm{softmax}\left(W_h h_i\right)$$

tanh →



ニューラルネットで何ができるか？

- 「特徴量」が学習可能
- 例： 話者本人が主語の文脈「{私, 僕, 俺}{は, が}」

は
が
私
僕
奈良
同僚
俺
…

$W_2[1]=\begin{pmatrix}-1\\-1\\1\\1\\-1\\-1\\1\\\cdots\end{pmatrix}$ $W_1[1]=\begin{pmatrix}1\\1\\-1\\-1\\-1\\-1\\-1\\\cdots\end{pmatrix}$ $b[1]=-1$

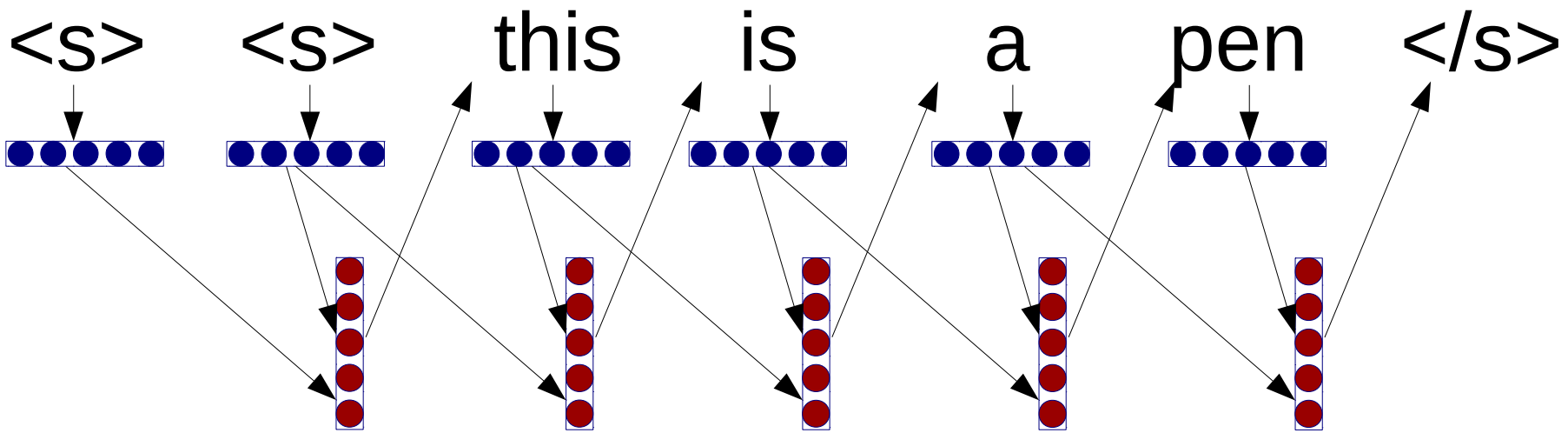
私 は → tanh(1)

彼 は → tanh(-1)

彼 の → tanh(-3)

- 両方が成り立てば、隠れ層の１ノード目は正の値
  そうでなければ、負の値
- 数え上げなら、全パターンを覚える必要あり！

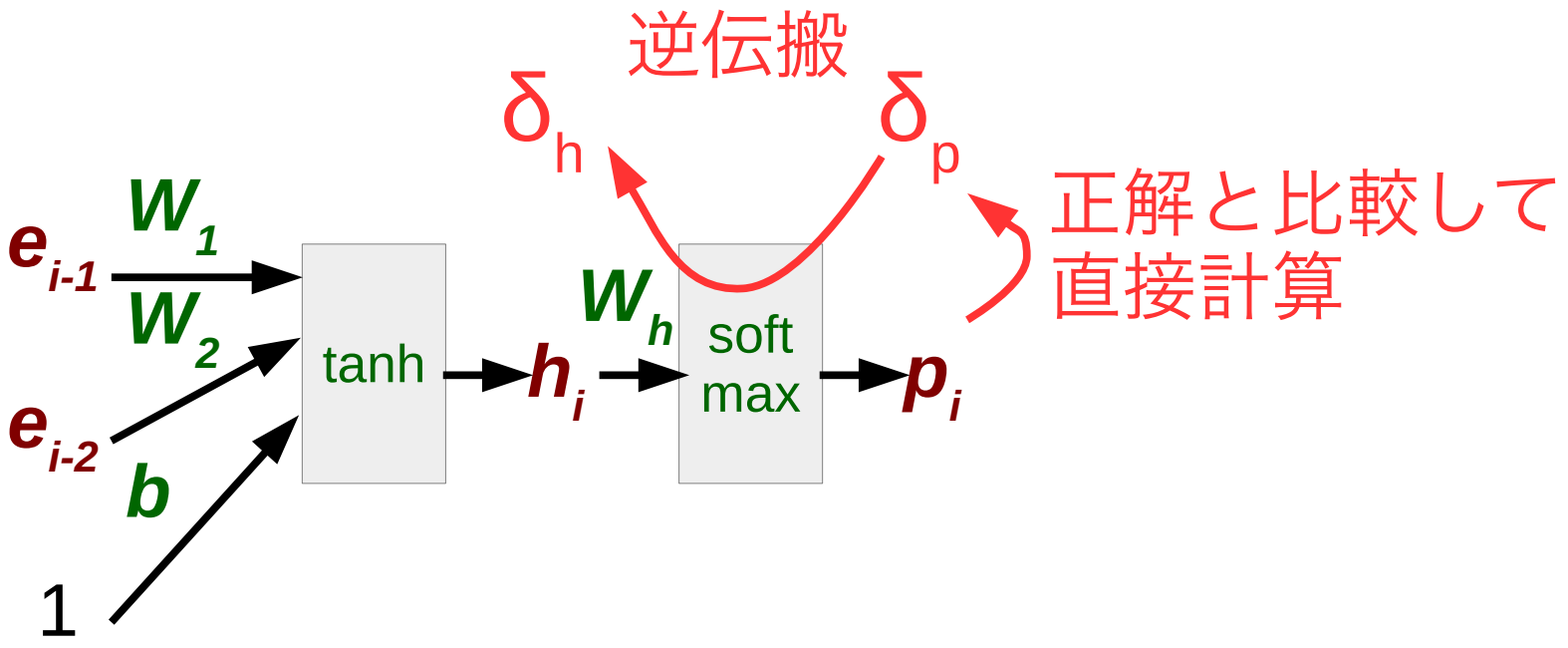
ニューラルネット言語モデル [Nakamura+ 90, Bengio+ 06]



- 低次元隠れ層で出力の類似性を考慮
- 単語表現で文脈の類似性を考慮
- 文脈のすべての単語を直接考慮するため、未知語を含めた文脈で壊れない

ニューラルネットの学習：逆伝搬

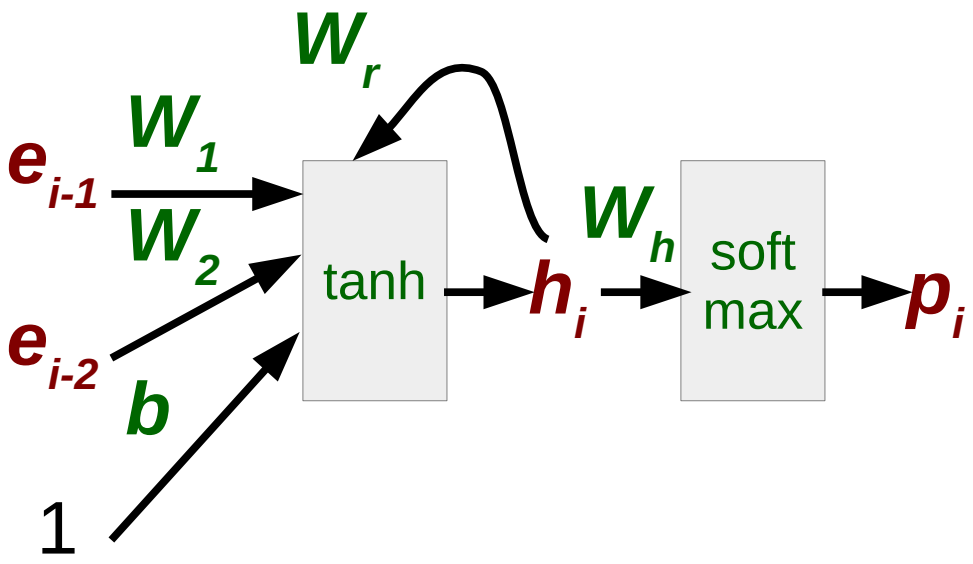
- 勾配を出力に近い方から逆順に伝搬



リカレントニューラルネット

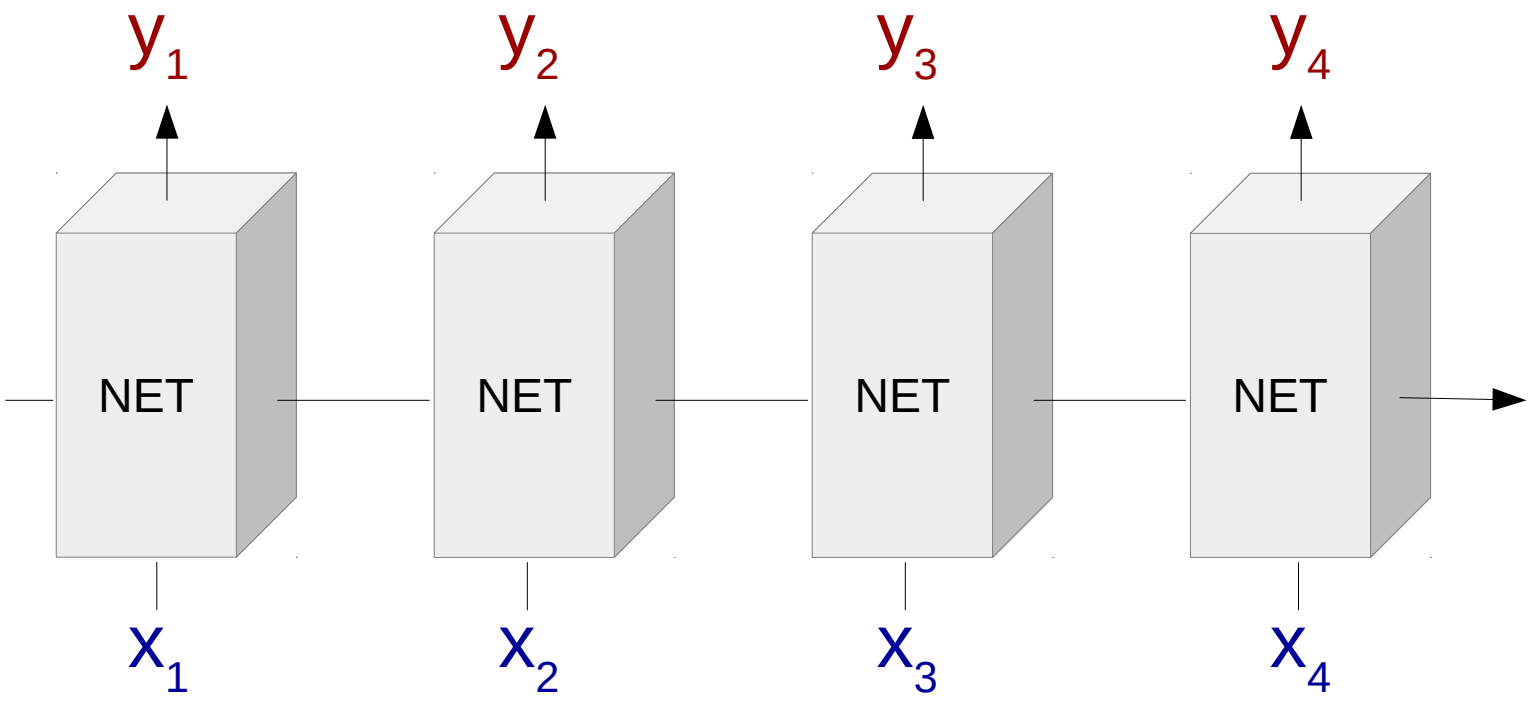
リカレントニューラルネット (RNN)

- ノードの一部の出力が入力として戻ってくる

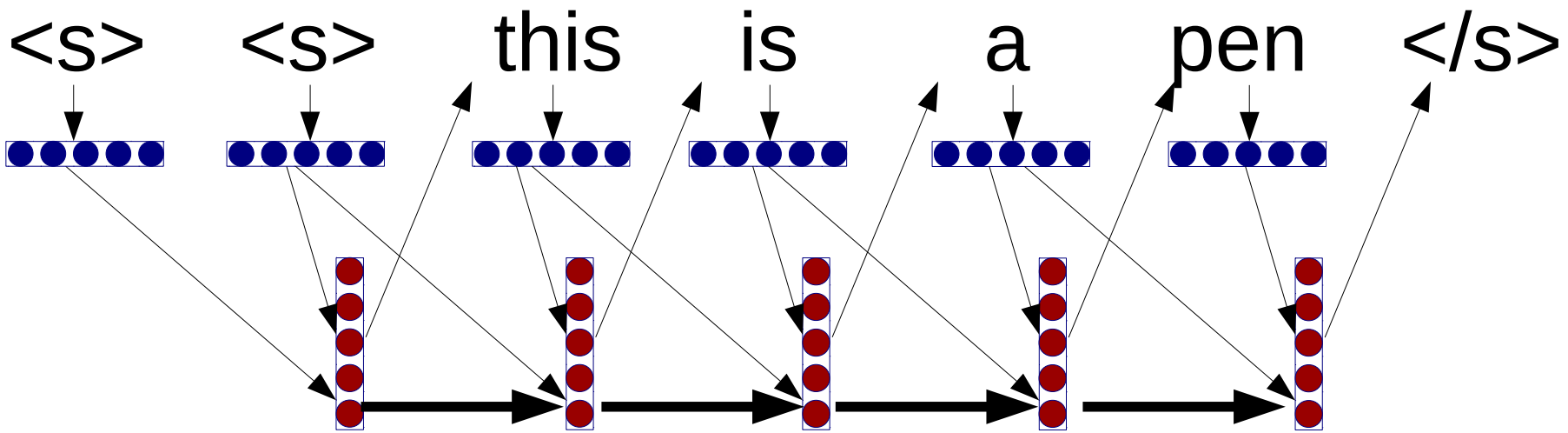


- 理由：長距離に渡る依存性の「記憶」が可能

系列モデルとしての RNN

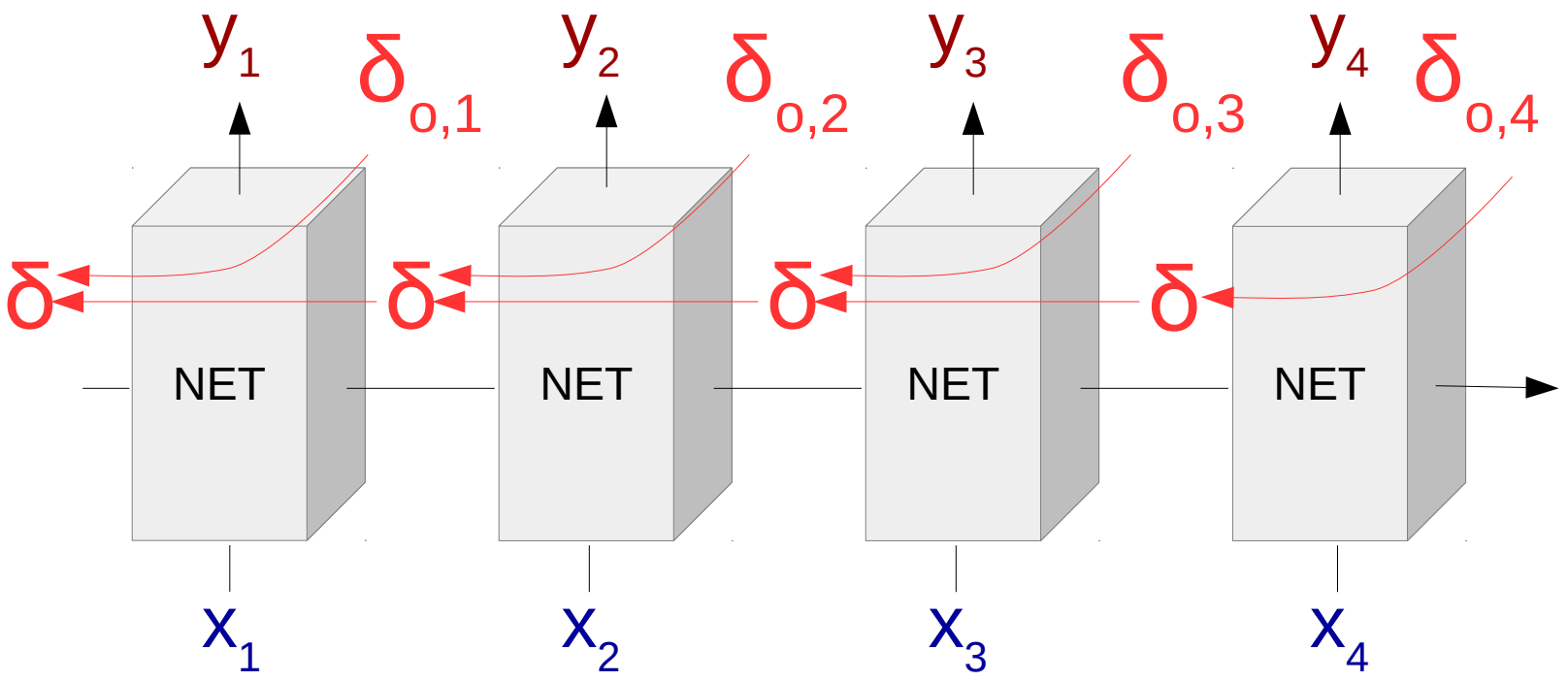


リカレントニューラルネット言語モデル [Mikolov+ 10]



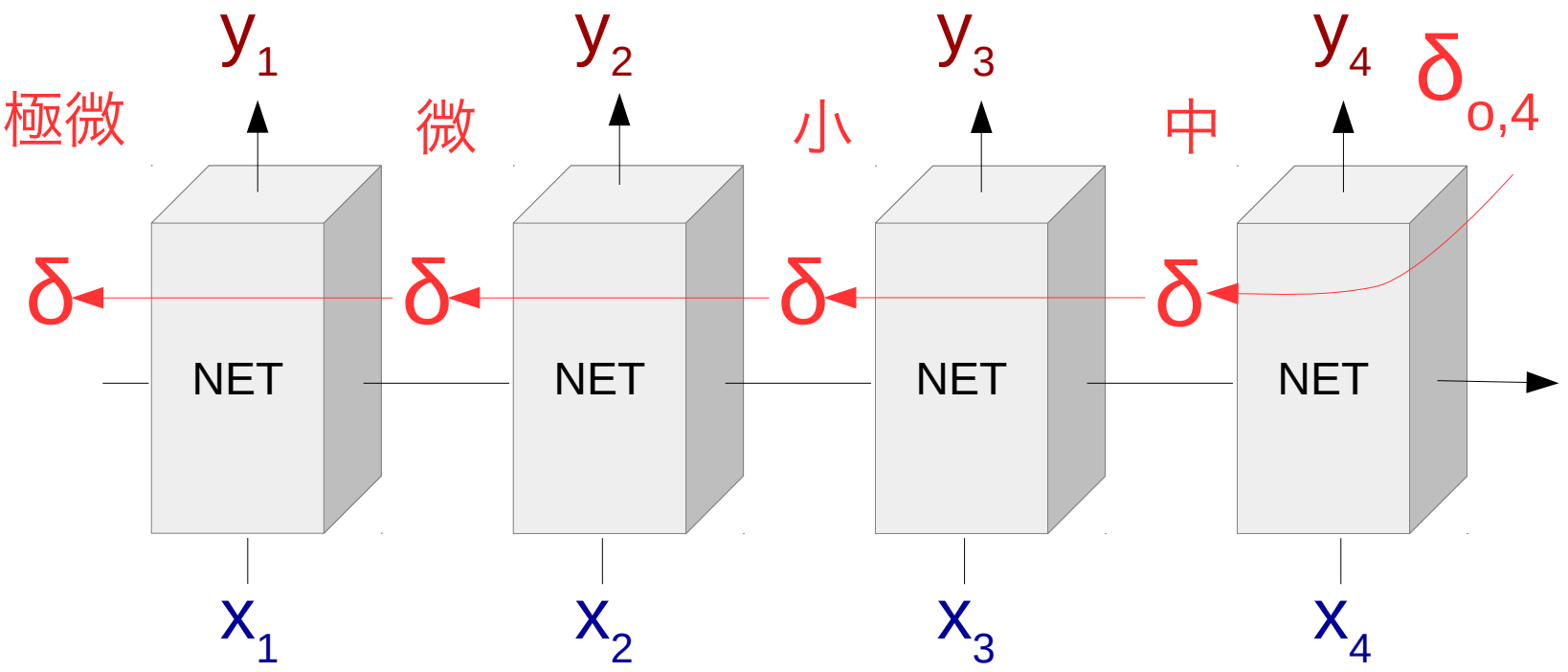
- 以前の単語を「記憶」する
- 機械翻訳、音声認識などで精度の向上を実現

RNN の勾配計算



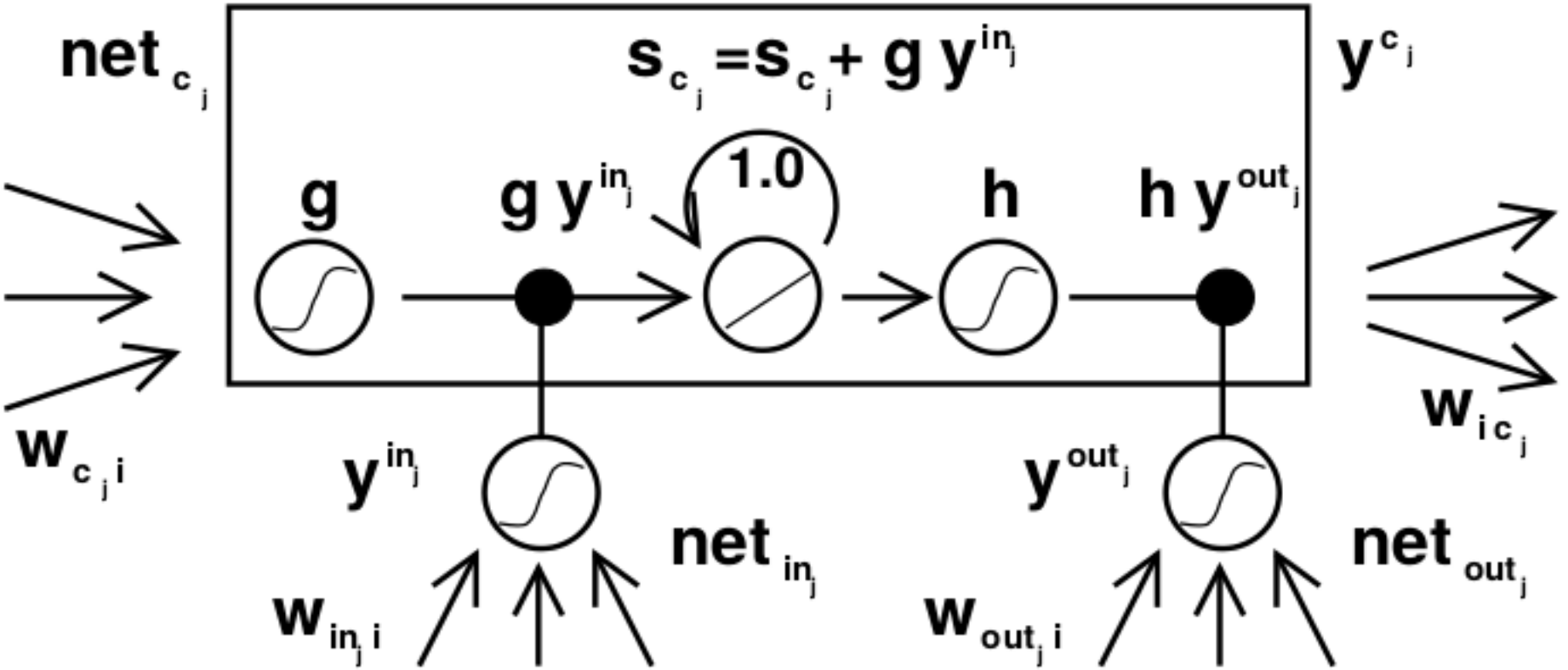
- まず系列のネット結果全体を計算
- 後ろからエラーを計算

ニューラルネットにおける消える勾配



Long Short-term Memory [Hochreiter+ 97]

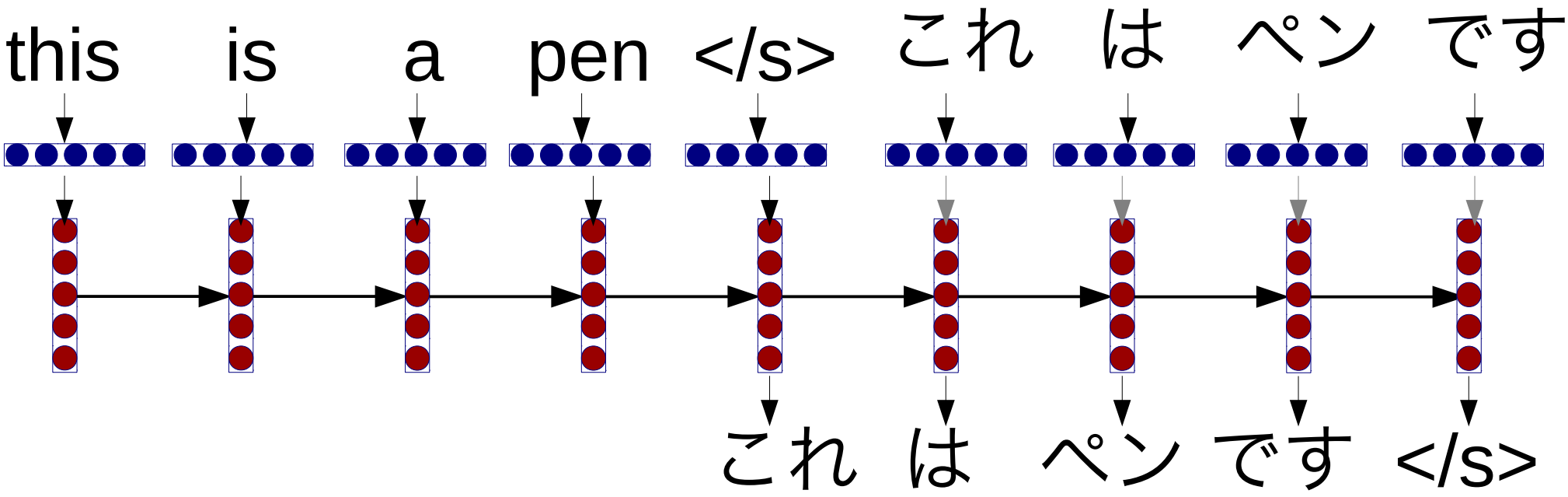
- 線形関数を使った隠れ状態＋ゲートで勾配をコントロール



Encoder-Decoder 翻訳モデル

[Kalchbrenner+ 13, Sutskever+ 14]

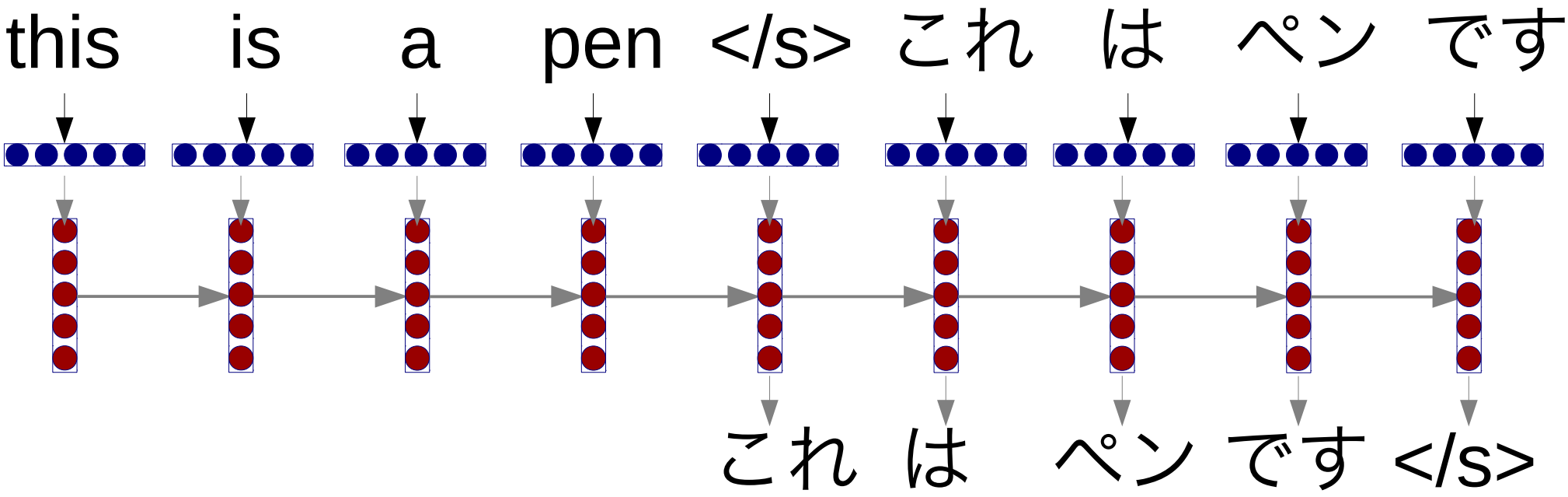
LSTM ニューラルネット翻訳モデル [Sutskever+ 14]



- つまり、入力言語で条件付けられた言語モデル

$$P\left(e_1^I|f_1^J\right)=\prod_{i=1}^{I+1}P\left(e_i|f_1^J,e_1^{i-1}\right)$$

訳文の生成

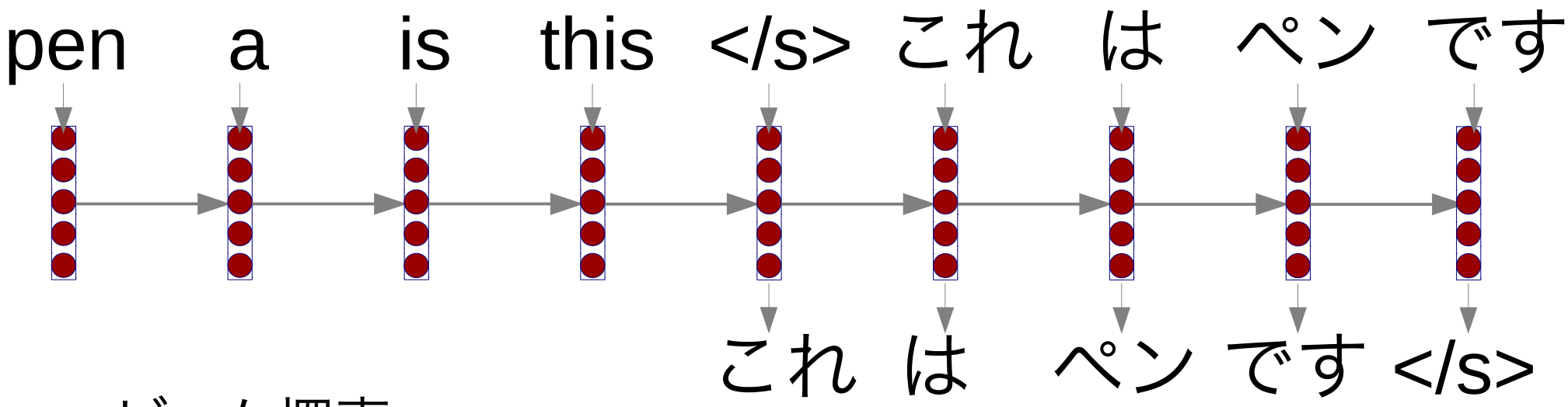


入力文をエンコード　一単語ずつ生成

$$\mathrm{argmax}_{e_i} P\left(e_i | f_1^J, e_1^{i-1}\right)$$

詳細

- 入力を逆順にする（学習が容易に）



- ビーム探索
- モデルのアンサンブル

疑問１：本当にそれだけで翻訳できるか？

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

日英における再現実験

- 旅行会話 11.6 万文で学習

	BLEU	RIBES
Moses PBMT	38.6	80.3
Encoder-Decoder	39.0	82.9

疑問２：
人手で評価しても通用するか？

再現実験：はい、ある程度は。

入力:　バスタブからお湯があふれてしまいました。
正解:　the hot water overflowed from the bathtub .
PBMT：the hot water up the bathtub .
EncDec:the bathtub has overflowed .

入力:　コーヒーのクリーム入りをください。
正解:　i 'll have some coffee with cream , please .
PBMT:　cream of coffee , please .
EncDec: i 'd like some coffee with cream .

ただし、問題はある

あきらめ：

入力: ギブス を し な けれ ば な り ま せ ん 。
正解: you 'll have to have a cast .
PBMT: i have a ギブス .
EncDec: you have to have a chance .

繰り返し：

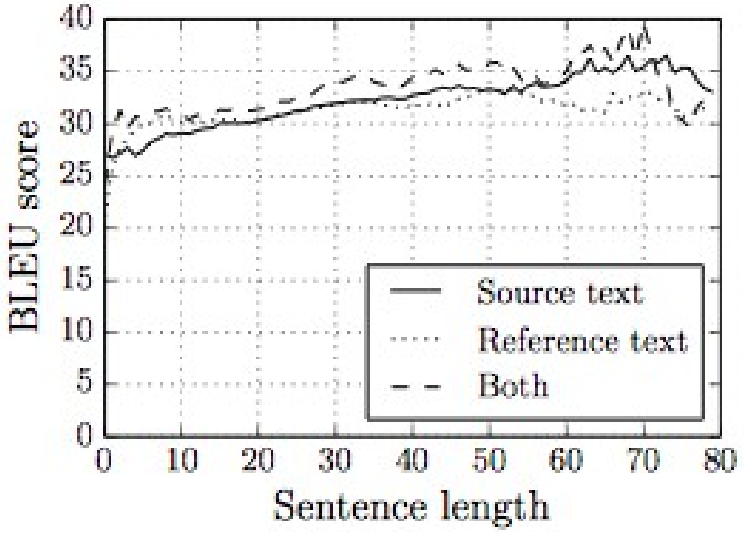
入力: どのファンデーションが私の肌の色に近いですか。
正解: which foundation comes close to my natural skin color ?
PBMT: which foundation near my natural skin color ?
EncDec: which foundation is my favorite foundation with a foundation ?

注意型ニューラルネットに基づく翻訳

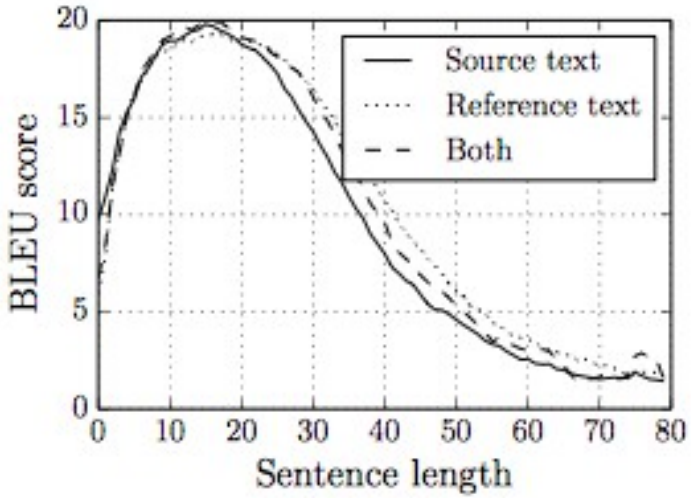
疑問３：
可変長の文を一定のベクトルで表せるか？

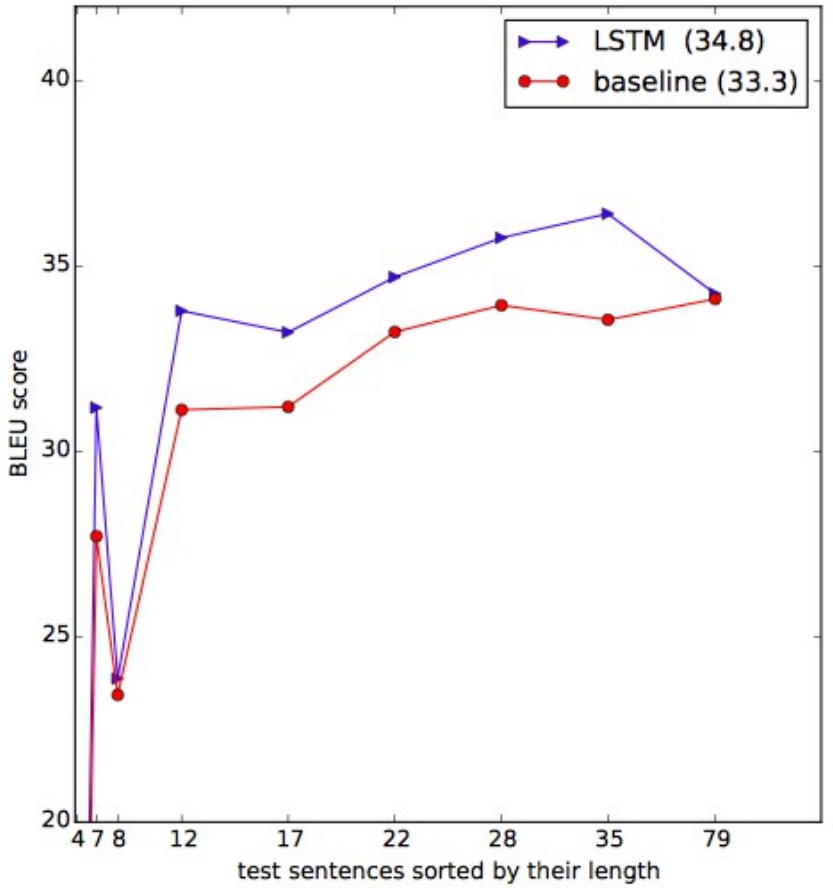
no? [Pouget-Abadie+ 2014] yes? [Sutskever+ 2014]

PBMT



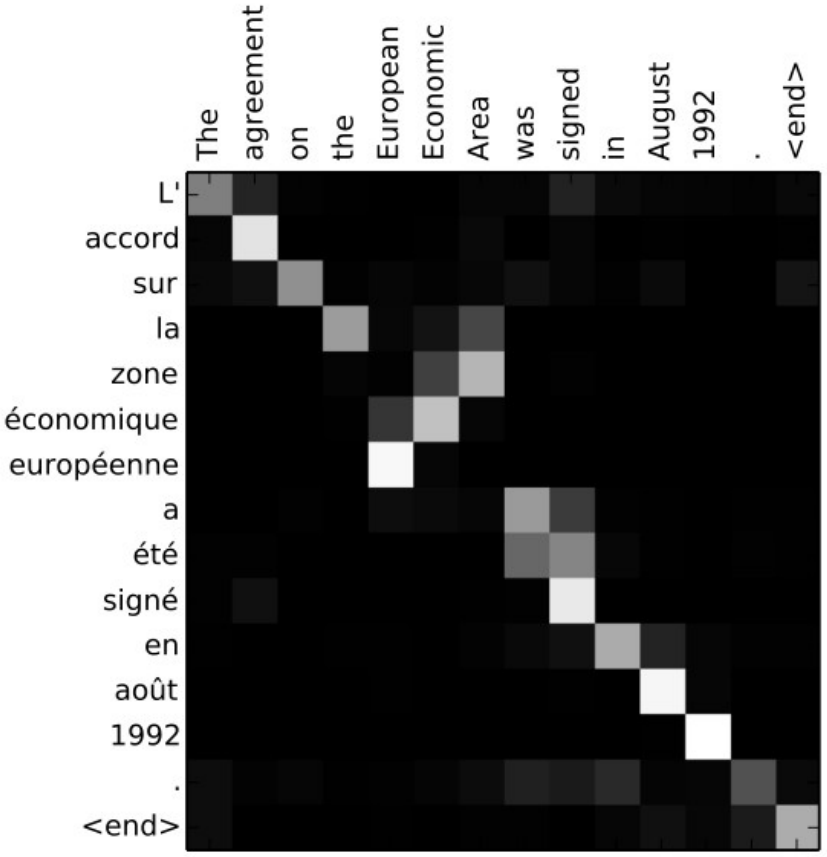
RNN





注意型ニューラル翻訳 [Bahdanau+ 15]

- 対象の文をエンコーディングし、文のどこに注意するかを決定しながら翻訳



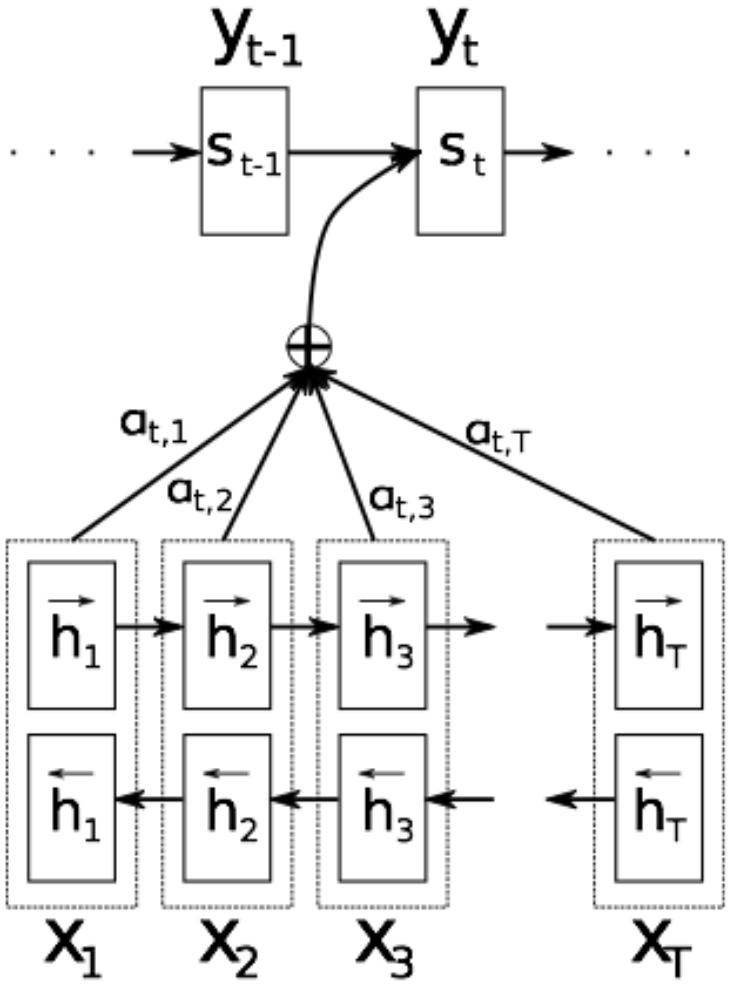
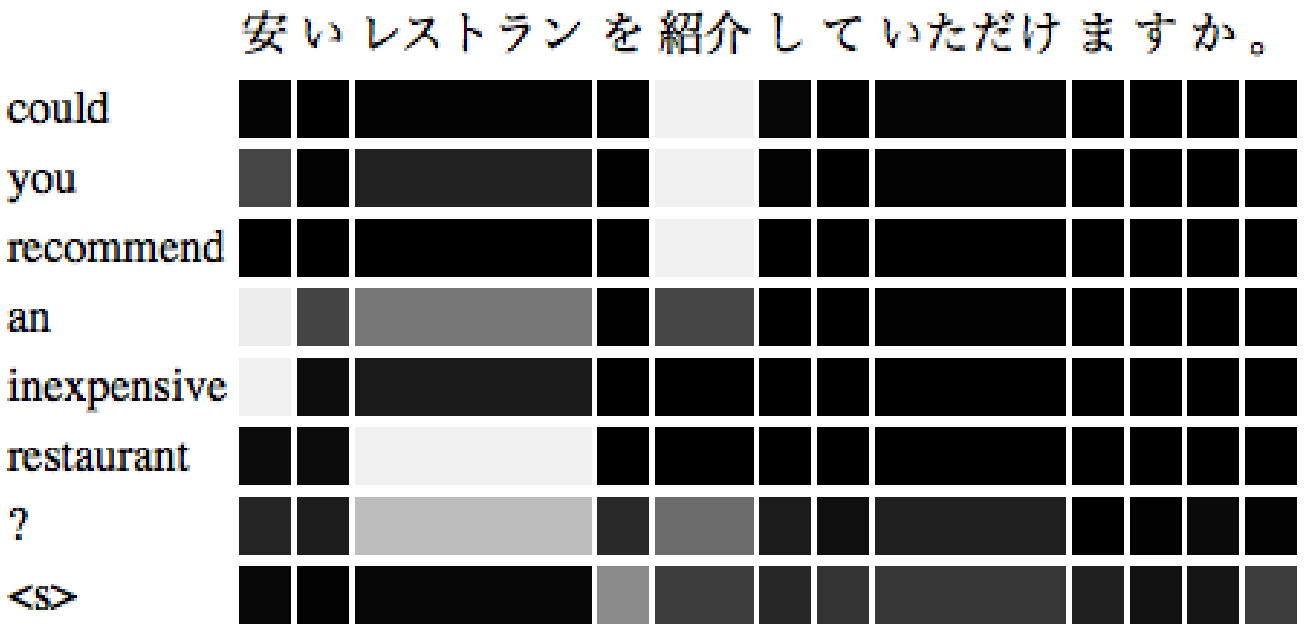


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

再現実験

- 日英旅行対話 11.6 万文で学習

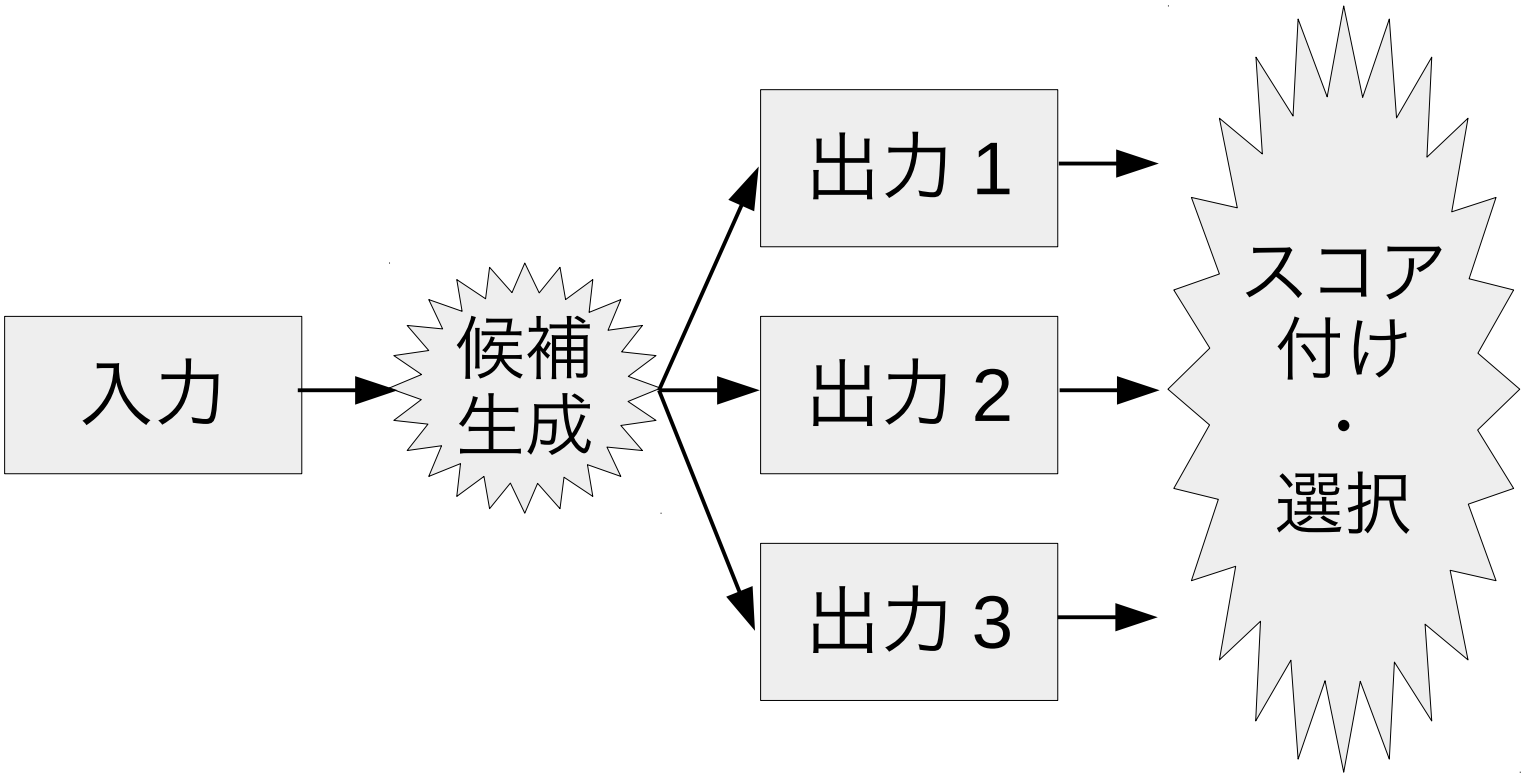
	BLEU	RIBES
Moses PBMT	38.6	80.3
Encoder-Decoder	39.0	82.9
Attentional	40.8	84.0



従来法との組み合わせ：リランキング

リランキング

- 従来のシステムから結果を出し、ニューラル翻訳のスコアを使いながら選択



[Sutskever+ 2014] の結果

- 英語→フランス語
- Workshop on Machine Translation 2014

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

Workshop on Asian Translation における日本語を用いた実験

- ベースラインは構文情報を用いる強いシステム
- すべての言語、自動・人手評価で一貫して大きな性能向上

		BLEU	RIBES	HUMAN
en-ja	Baseline	36.6	79.6	49.8
	Reranking	**38.2**	**81.4**	**62.3**
ja-en	Baseline	22.6	72.3	11.8
	Reranking	**25.4**	**75.0**	**35.5**
zh-ja	Baseline	40.5	83.4	25.8
	Reranking	**43.0**	**84.8**	**35.8**
ja-zh	Baseline	30.1	81.5	2.8
	Reranking	**31.6**	**83.3**	**7.0**

実例

入力： 另外，各国也进行了本国销售的食品的实态调查。
正解： また，各国でも，自国で販売している食品の実態調査が行われた。
Base: また，各国は自国販売の食品の実態調査を行った。
Rerank: また，各国でも本邦で販売される食品の実態調査を行った

入力: 在此，以研究教育现场的“风险交流”的实情为前提，整理了如下项目。

正解:
ここでは教育現場における「リスクコミュニケーション」のあり方を検討するための前提を以下の項目に分けて整理した。

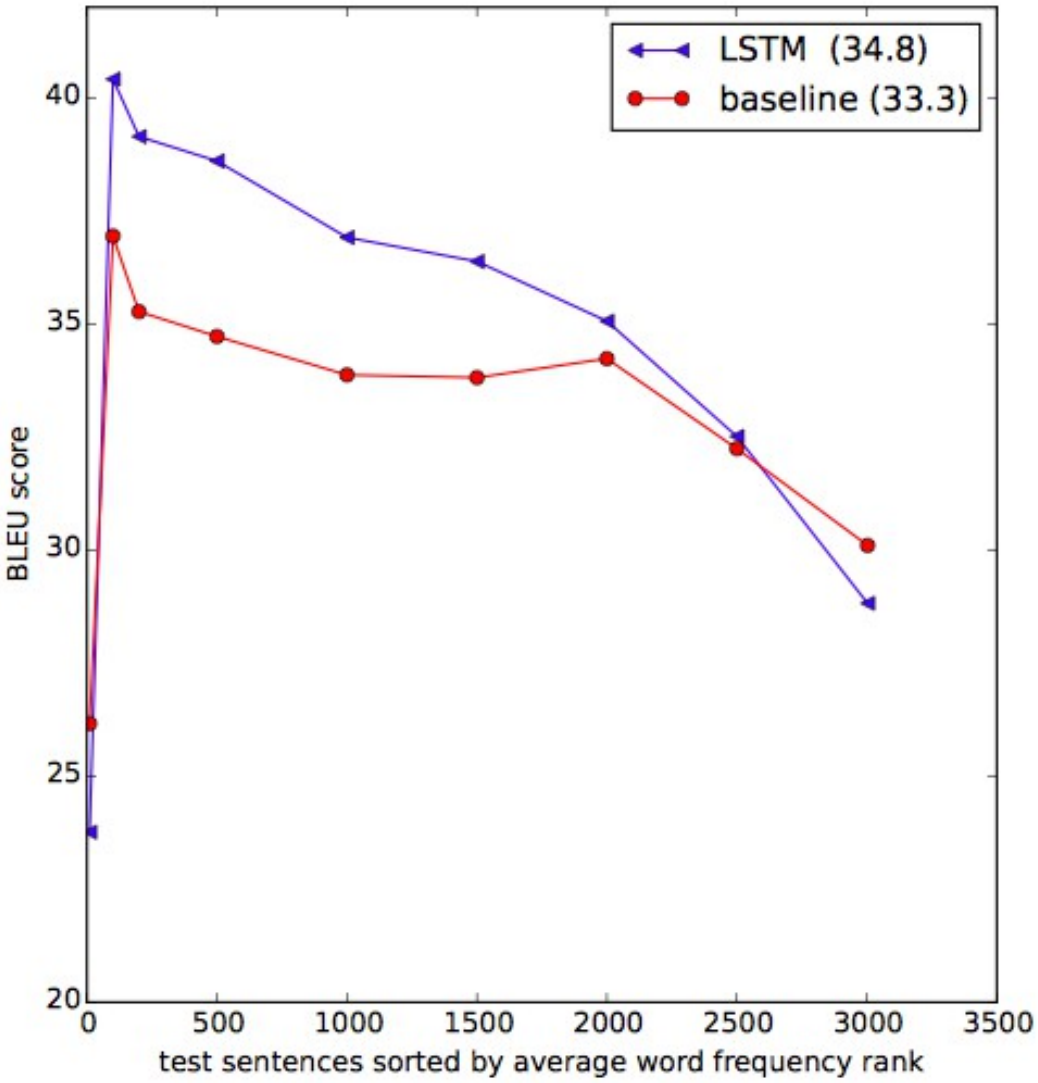
Base:
ここでは，「リスクコミュニケーション」の教育現場研究の実情を前提として，以下の項目について整理した。

Rerank:
ここでは，教育現場における「リスクコミュニケーション」の実態を前提として，以下の項目について整理した

今後の課題？

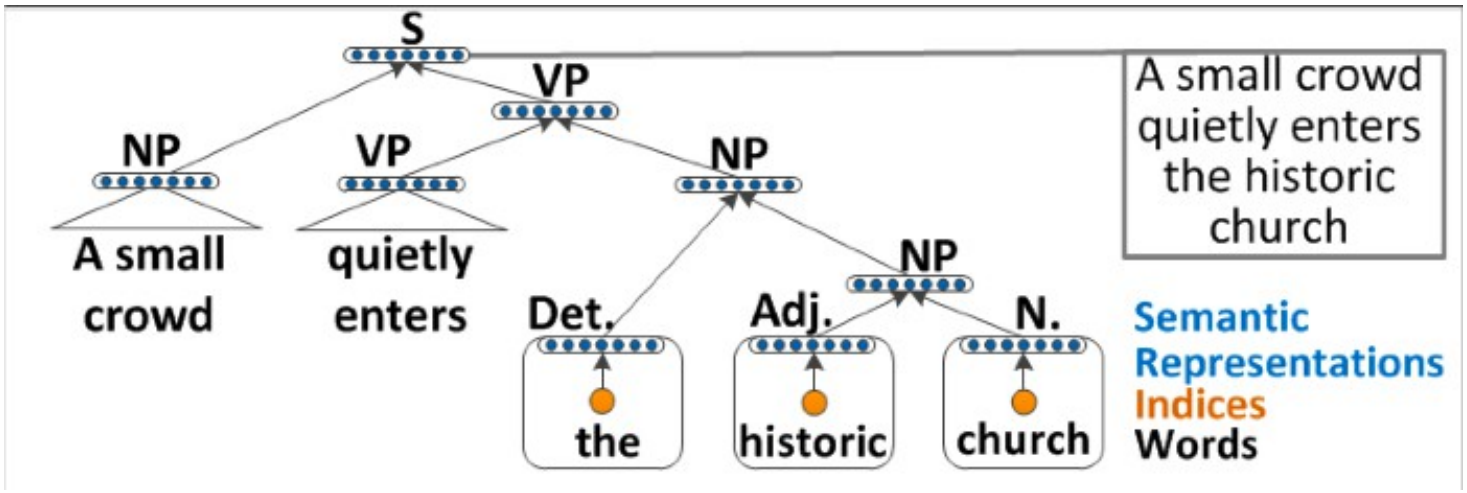
大語彙

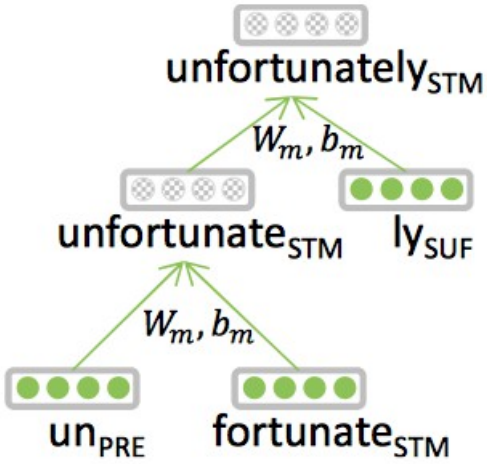
- 学習の都合上、出力語彙が増えると大変
- 低頻度後に弱い→
- 未知語処理で対応 [Luong+ 15]
- 効率的な学習法
  - ノイズ対照推定 (NCE) [Vaswani+ 13]
  - 学習データの分割 [Jean+ 15]



統語・形態論情報の利用

- 現在は言語構造はいっさい未考慮
- 統語情報を使った事前並べ替え＋系列モデル [外山 +15]
- 統語情報を考慮したニューラルネットは利用可？ [Socher+11, Luong+13]





制御可能性

- 細かく訳出結果を制御することは不可
- 今のところ、データの追加以外の改良法はない？

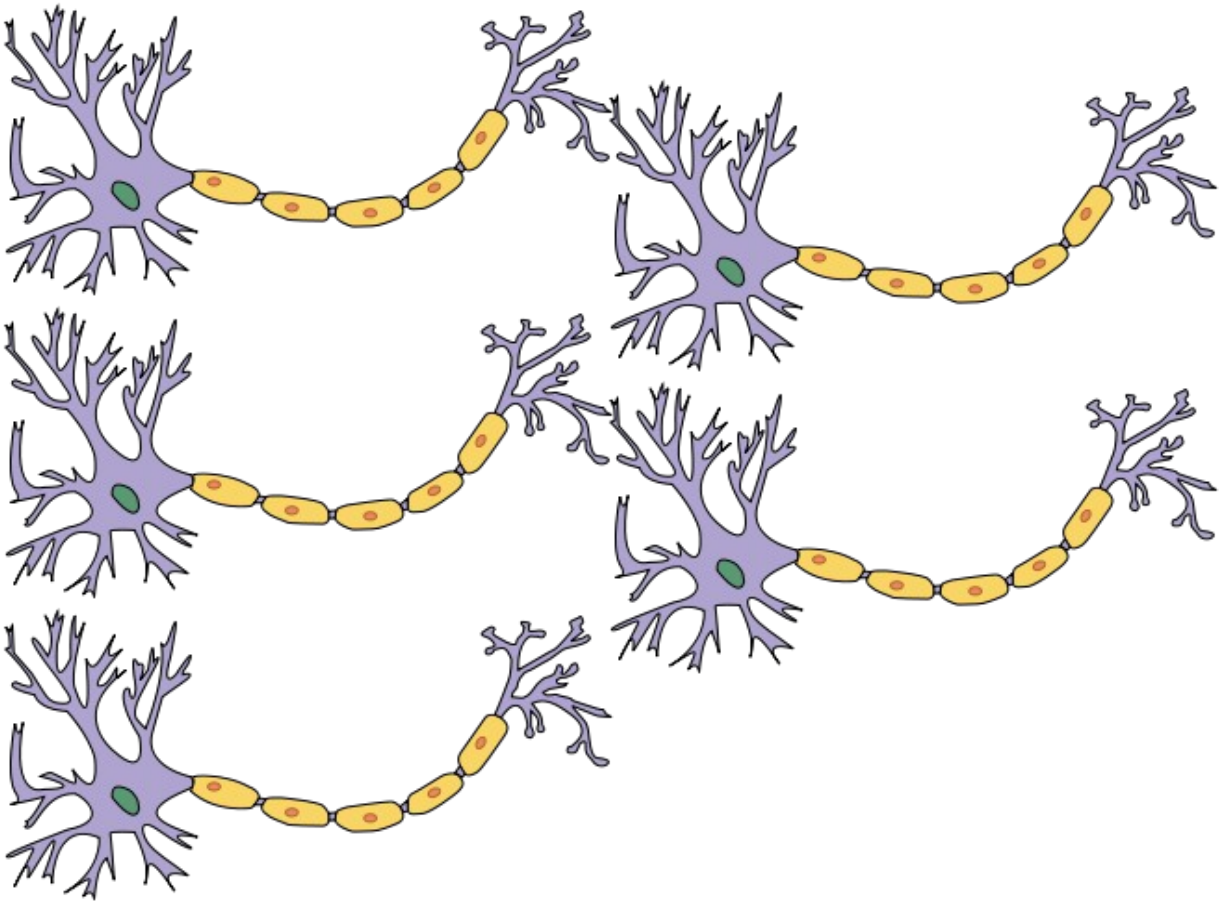
参考資料・文献

- Kevin Duh: Deep Learning for Machine Translation http://cl.naist.jp/~kevinduh/notes/cwmt14tutorial.pdf

- D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Proc. ICLR, 2015.
- Y. Bengio, H. Schwenk, J.-S. Sen´ecal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning, 2006.
- S. F. Chen and R. Rosenfeld. A survey of smoothing techniques for me models. Speech and Audio Processing, IEEE Transactions on, 8(1):37–50, Jan 2000.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.
- S. Jean, K. Cho, R. Memisevic, and Y. Bengio. On using very large target vocabulary for neural machine translation. In Proc. ACL, 2015.
- N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In Proc. EMNLP, pages 1700–1709, Seattle, Washington, USA, 2013. Association for Computational Linguistics.
- M.-T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. In Proc. ACL, 2015.
- T. Luong, R. Socher, and C. Manning. Better word representations with recursive neural networks for morphology. pages 104–113, 2013.
- T. Mikolov, M. Karafi´at, L. Burget, J. Cernocky`, and S. Khudanpur. Recurrent neural network based language model. In Proc. InterSpeech, pages 1045–1048, 2010.
- M. Nakamura, K. Maruyama, T. Kawabata, and K. Shikano. Neural network approach to word category prediction for English texts. In Proc. COLING, 1990.
- R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. pages 129–136, 2011.
- I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Proc. NIPS, pages 3104–3112, 2014.
- A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang. Decoding with large-scale neural language models improves translation. In Proc. EMNLP, pages 1387–1392, 2013.

追加資料

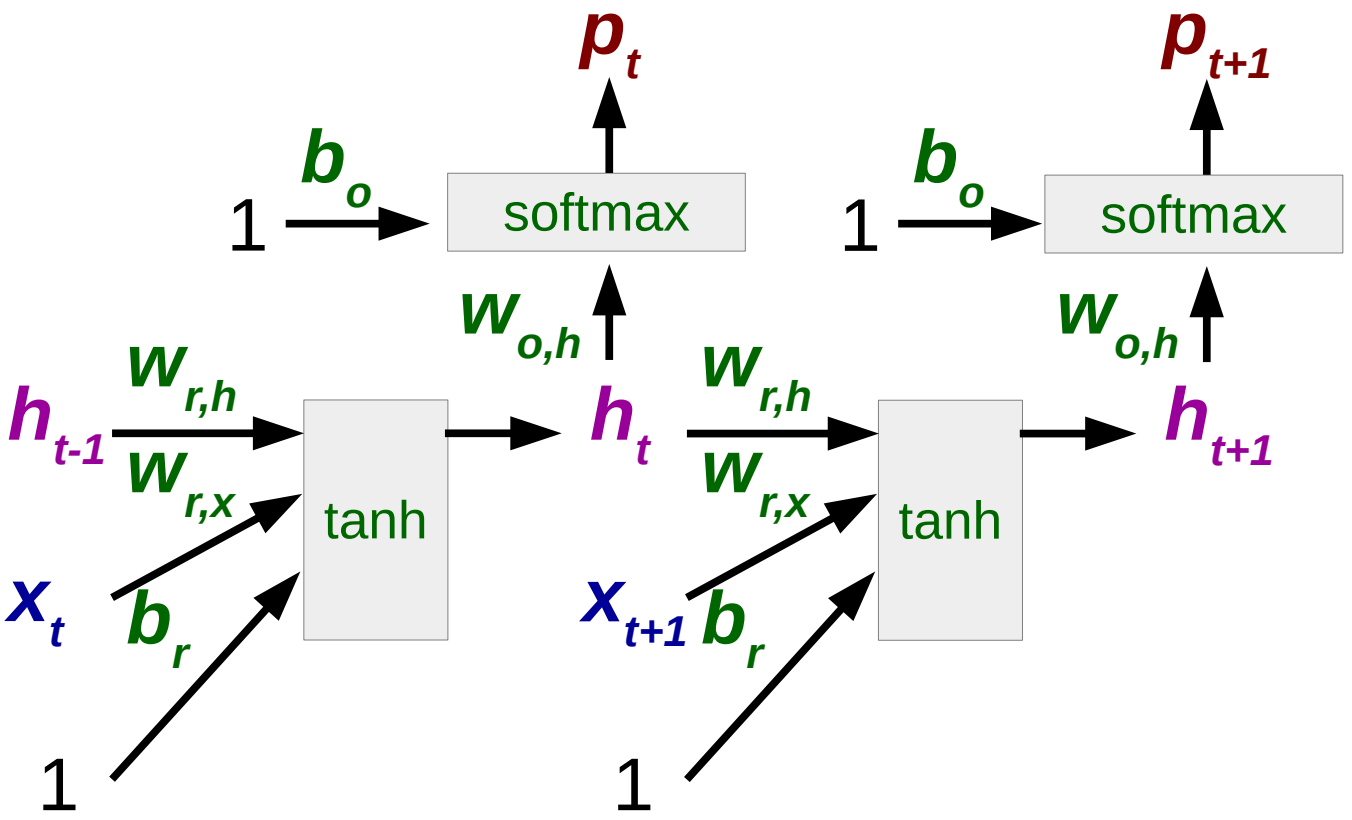
NN に関する考え方 (1980 年代ごろ？)



- 生理学的解釈を重視

画像： Wikipedia

NN に対する考え方 (2010 年代)



- ただの（微分可能な）関数のつながり