自然言語処理における深層ニューラルネットワーク

東北大学大学院情報科学研究科
岡崎　直観（okazaki@ecei.tohoku.ac.jp）

http://www.chokkan.org/
@chokkanorg

自然言語処理とは

- 言葉を操る賢いコンピュータを作る
  - 応用: 情報検索，機械翻訳，質問応答，自動要約，対話生成，評判分析，SNS分析，…
  - 基礎: 品詞タグ付け（形態素解析），チャンキング，固有表現抽出，構文解析，共参照解析，意味役割付与，…

- 多くのタスクは「入力xから出力$\hat{y}$を予測」

$$\hat{y} = \underset{y \in Y}{\operatorname{argmax}} P(y|x)$$

※確率ではないモデルもあります

単語列からラベル: $\hat{y} = \underset{y \in Y}{\operatorname{argmax}} P(y|\boldsymbol{x})$

$\boldsymbol{x}$: 単語列　　$P(y|\boldsymbol{x})$　　$\hat{y}$

The movie is the best I've ever seen! → 

The movie is coming soon on cinemas. → 

This movie is rubbish!!! → 

- モデル：ナイーブ・ベイズ，パーセプトロン，
  ロジスティック回帰，サポート・ベクトル・マシン

単語列から系列ラベル: $\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in Y^m}{\operatorname{argmax}} P(\boldsymbol{y}|\boldsymbol{x})$

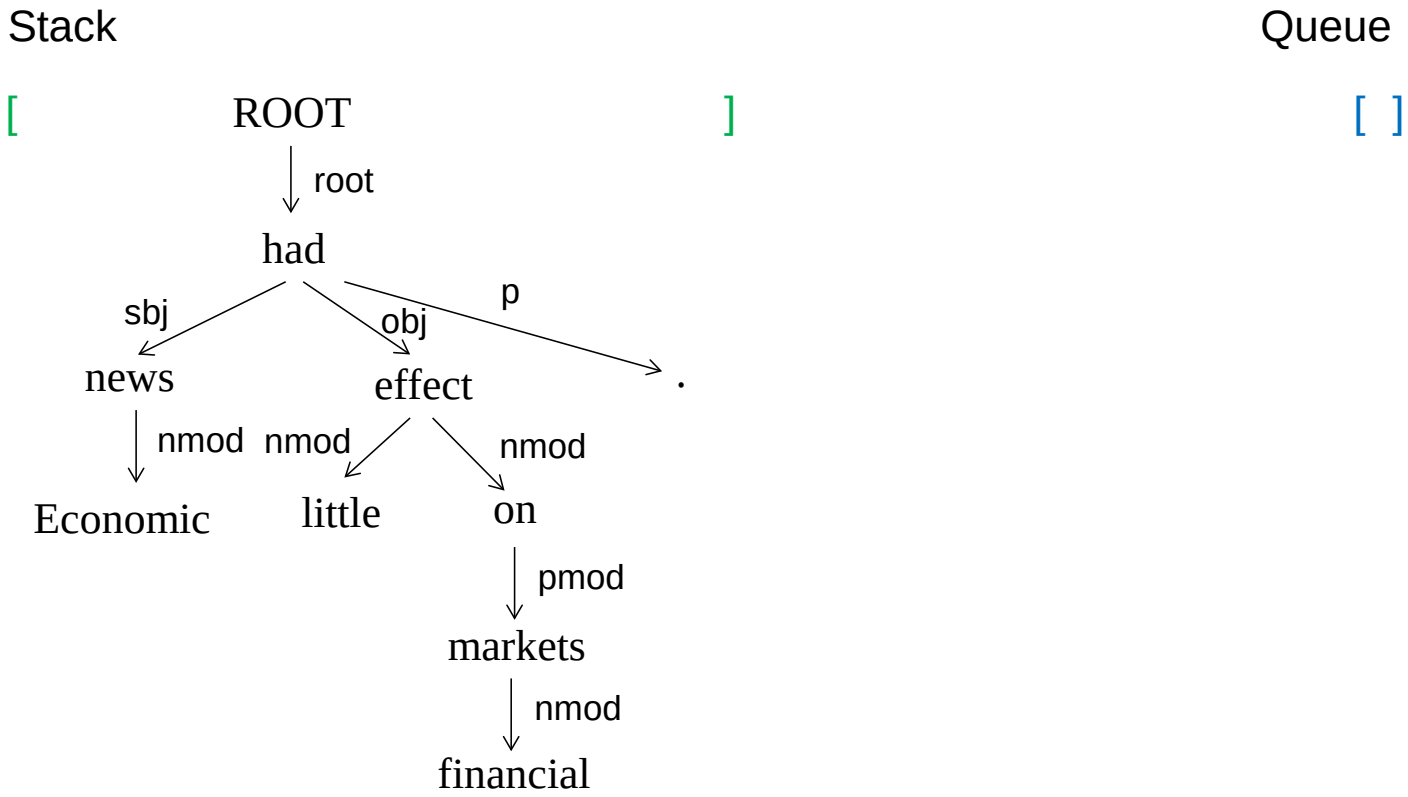
(入力)	*In*	*March*	*2005,*	*the*	*New*	*York*	*Times*	*acquired*	*About,*	*Inc.*
(品詞)	IN	NNP	CD	DT	NNP	NNP	NNP	VBD	NNP	NNP
(句)	O	B-NP	I-NP	B-NP	I-NP	I-NP	I-NP	B-VP	B-NP	B-NP

(翻訳) 2005年 3月 , ニューヨーク・タイムズ は About 社 を 買収 し た .

(対話) I heard Google and Yahoo were among the other bidders.

- モデル：隠れマルコフモデル，条件付き確率場，符号・復号
- 探索法：点予測，動的計画法，ビーム探索，…

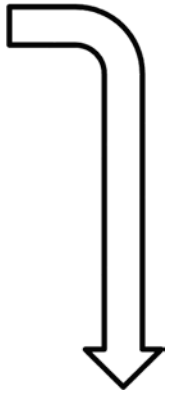
単語列から木構造: $\hat{y} = \underset{y \in \mathrm{Gen}(x)}{\operatorname{argmax}} P(y|x)$



- モデル：確率的文脈自由文法，条件付き確率場
- 探索法：Shift-Reduce法，Eisner法，CKY法，最小全域木

DNN以前の特徴抽出: 固有表現抽出の例

```
B-ORG    EU       NNP   B-NP
O        rejects  VBZ   B-VP
B-MISC   German   JJ    B-NP
O        call     NN    I-NP
O        to       TO    B-VP
O        boycott  VB    I-VP
B-MISC   British  JJ    B-NP
O        lamb     NN    I-NP
O        .        .     O
```



- 単語n-gram
- 品詞n-gram
- チャンクn-gram
- 大文字パターン
- 辞書マッチなど他にも沢山

```
w[-2]=EU w[-1]=rejects w[0]=German w[1]=call w[2]=to w[-2]|w[-
1]=EU|rejects w[-1]|w[0]=rejects|German w[0]|w[1]=German|call
w[1]|w[2]=call|to pos[-2]=NNP pos[-1]=VBZ pos[0]=JJ pos[1]=NN
pos[2]=TO pos[-2]|pos[-1]=NNP|VBZ pos[-1]|pos[0]=VBZ|JJ
pos[0]|pos[1]=JJ|NN pos[1]|pos[2]=NN|TO chk[-2]=B-NP chk[-1]=B-VP
chk[0]=B-NP chk[1]=I-NP chk[2]=B-VP chk[-2]|chk[-1]=B-NP|B-VP
chk[-1]|chk[0]=B-VP|B-NP chk[0]|chk[1]=B-NP|I-NP chk[1]|chk[2]=I-
NP|B-VP iu[-2]=True iu[-1]=False iu[0]=True iu[1]=False
iu[2]=False iu[-2]|iu[-1]=True|False iu[-1]|iu[0]=False|True
iu[0]|iu[1]=True|False  iu[1]|iu[2]=False|False
```

DNN時代の到来

- 画像認識でブレークスルー (2012年)
  - エラー率が10%以上減少 (ILSVRC 2012)

- 当時，言語処理での衝撃は限定的だった
  - 文字や単語などの明確な特徴量があったため？
  - 画像認識のようなブレークスルーはまだ無い

- 現在，多くのタスクでDNNが使われる
  - 地道な研究成果の積み重ね
  - 前スライドの固有表現抽出もDNNで最高精度

言語処理における(D)NNの主な成果

- Neural Language Model (Bengio 03)
- SENNA (CNNでNLP) (Collobert+ 08)
- Recursive Neural Network (Socher+ 11)
- Skip-gram & CBOW model (Mikolov+ 13)
- Encoder-decoder (Sutskever+ 14; Cho+ 14)
- Memory networks (Weston+ 14)
- Attention mechanism (Bahdanau+ 15)

分散表現　分散表現の合成　符号・複号化

単語の分散表現
(distributed representations)

分散表現 (Hinton+ 1986)

- 局所表現（local representation）
  - 各概念に１つの計算要素 (記号, ニューロン, 次元) を割り当て

バス

萌えバス

- 分散表現（distributed representation）
  - 各概念は複数の計算要素で表現される
  - 各計算要素は複数の概念の表現に関与する

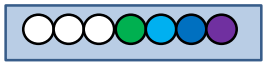
ニューロンの興奮パターン≒ベクトル表現

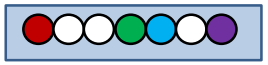
バス

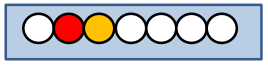
トラック

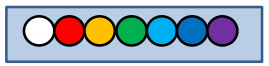
萌え

萌えバス





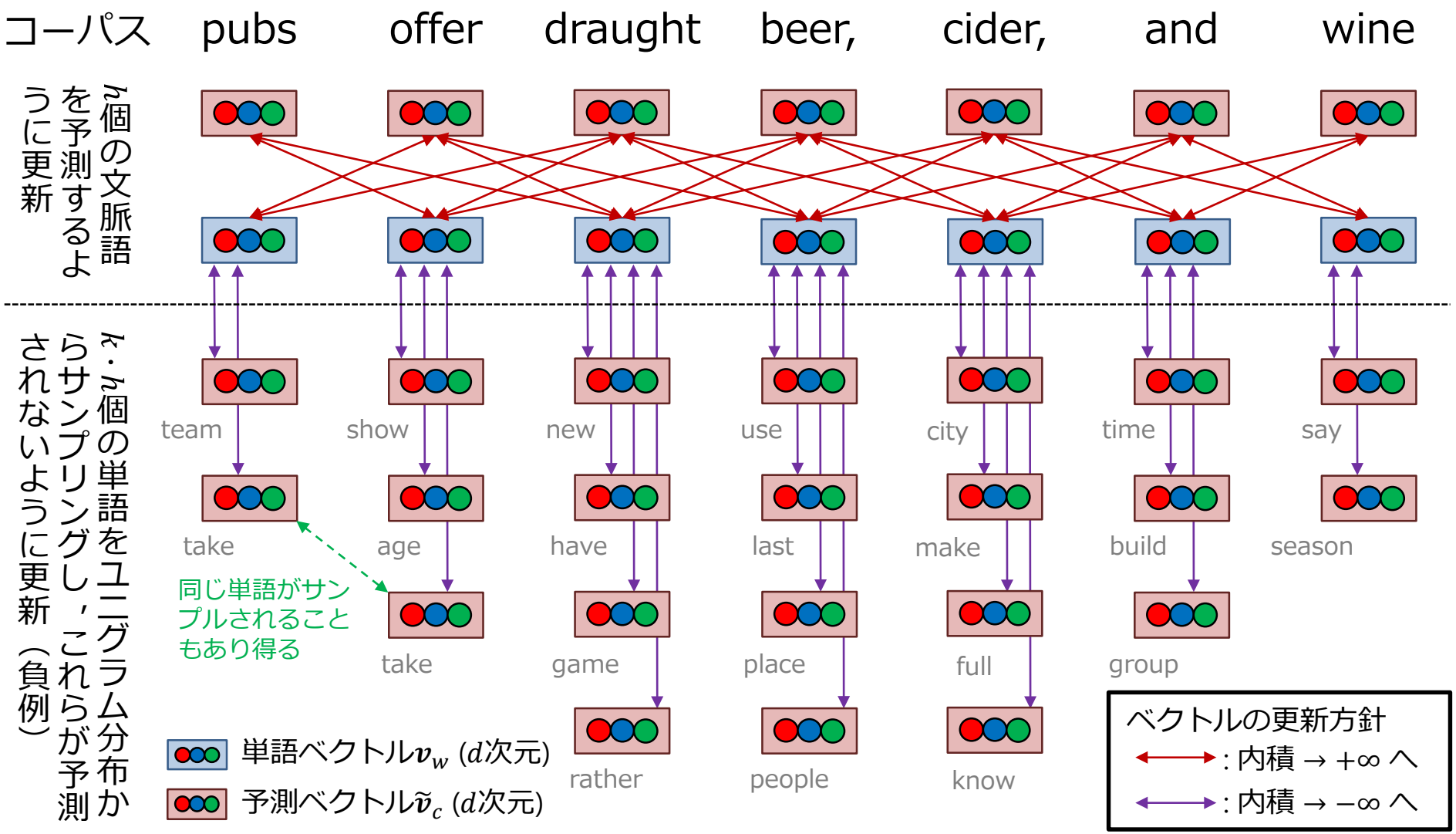




http://ja.wikipedia.org/wiki/富士急山梨バス　http://saori223.web.fc2.com/

Skip-gram with Negative Sampling (SGNS)

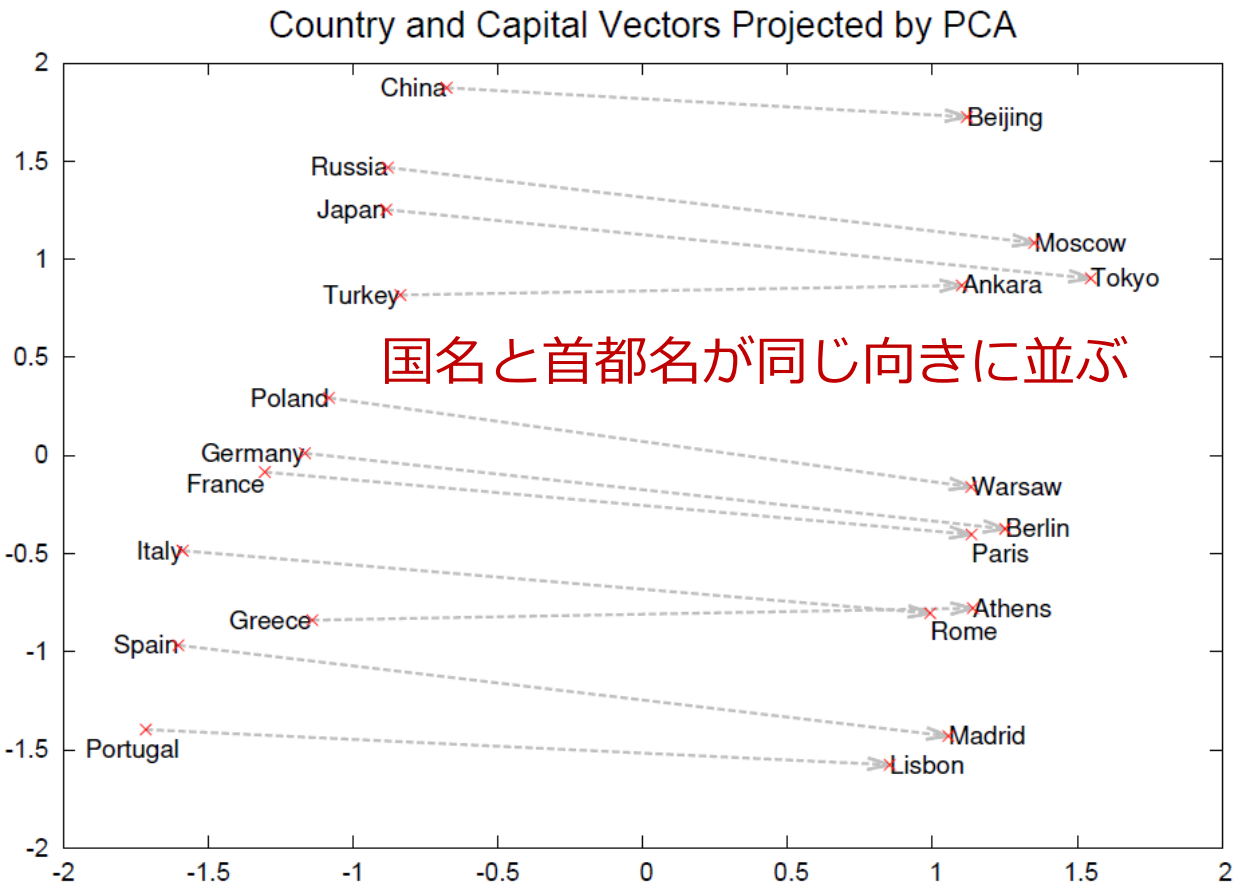
(Mikolov+ 2013)　（文脈幅$h = 2$, 負例サンプル数$k = 1$の場合の例）



(Mikolov+ 2013)

SGNSで学習した分散表現は加法構成性を持つ

- 有名な例: $\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$



(Mikolov+ 2013)

分散表現の合成 (composition)

句や文をDNNでモデル化したい

- 単語の分散表現の有用性が明らかになった
- 次は句や文の分散表現を獲得したい
- 構成性の原理に基づき，単語の分散表現から句や文の分散表現を合成する
  - 句や文の意味は，その構成要素の意味とその合成手続きから計算できる

- 最も単純かつ強力な手法は平均

$$\boldsymbol{p} = \frac{1}{2}(\boldsymbol{u} + \boldsymbol{v})$$

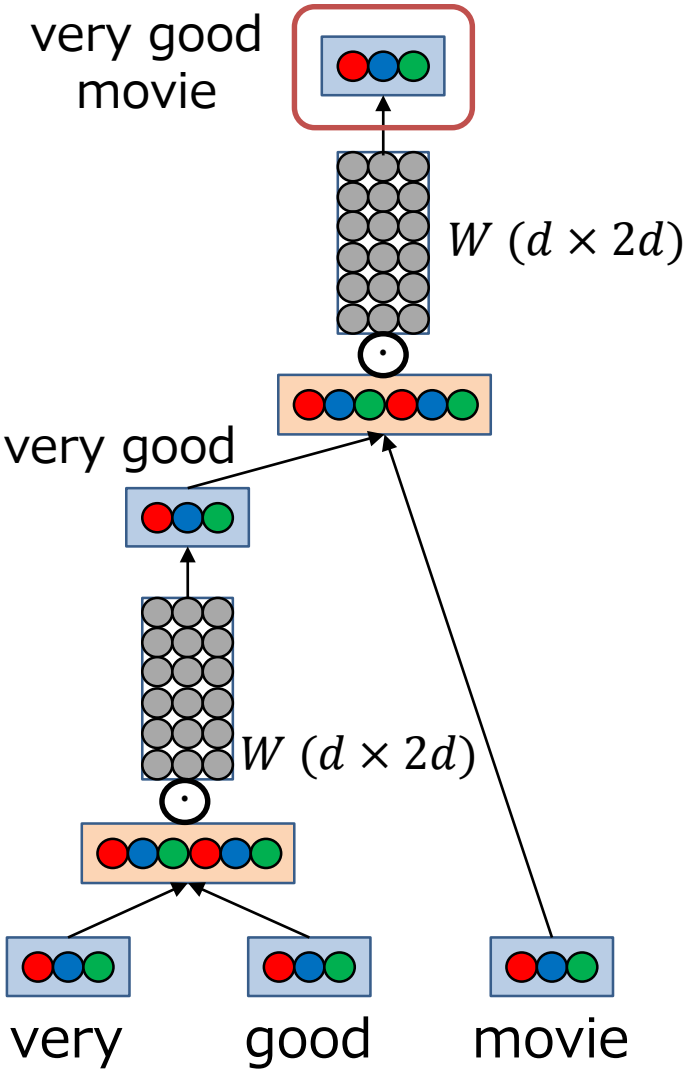
Recursive Neural Network (RNN)
(Socher+ 2011)

- 句ベクトルを次式で合成

$$\boldsymbol{p} = f(\boldsymbol{u}, \boldsymbol{v}) = g\left(W \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}\right)$$

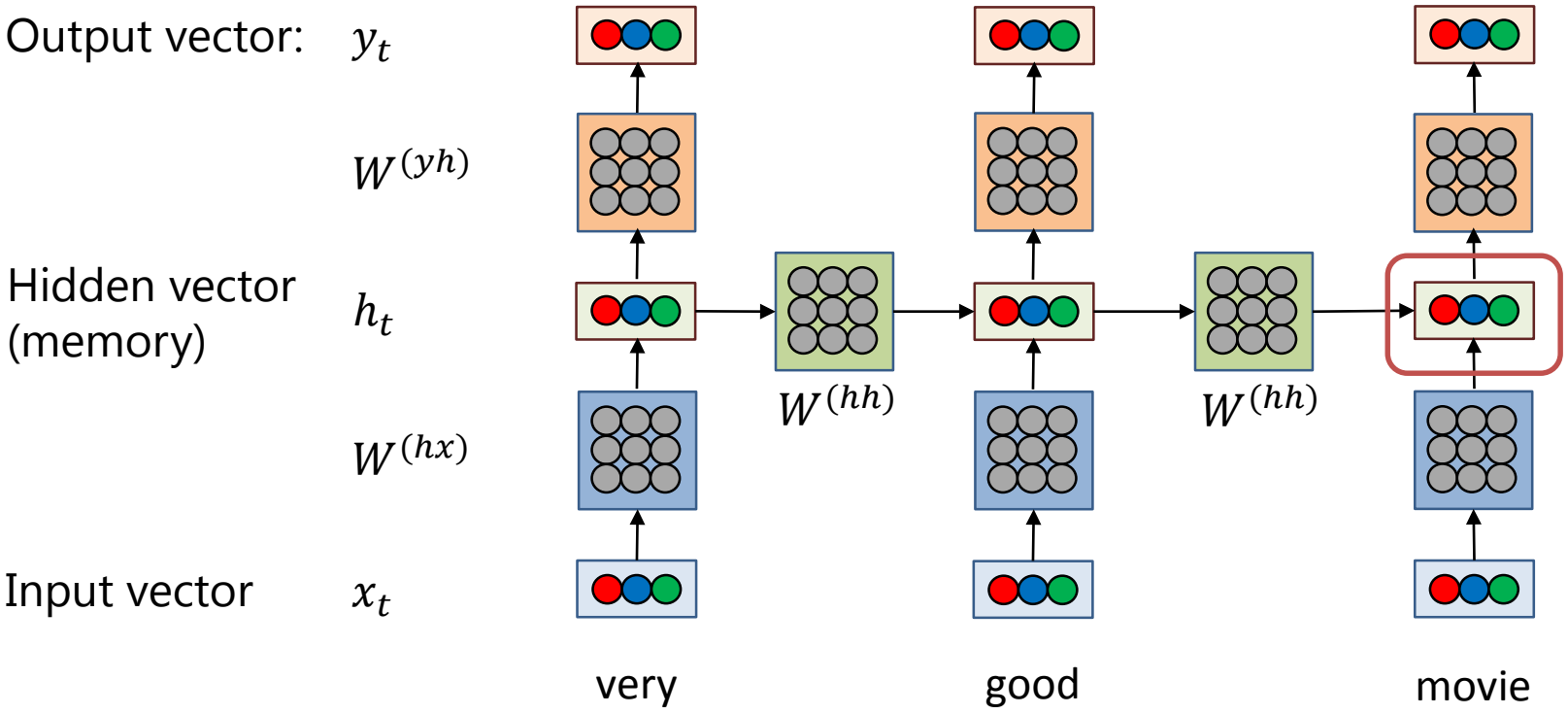
  - $W \in \mathbb{R}^{d \times 2d}$: $\mathbb{R}^{2d} \rightarrow \mathbb{R}^{d}$の変換行列
  - g: 活性化関数 (σやtanh)

- 文の句構造に従って再帰的に句ベクトルを計算
- 訓練データを用い，誤差逆伝搬法でWを学習
- 単語ベクトルも同時に学習



Recurrent Neural Network (RNN)

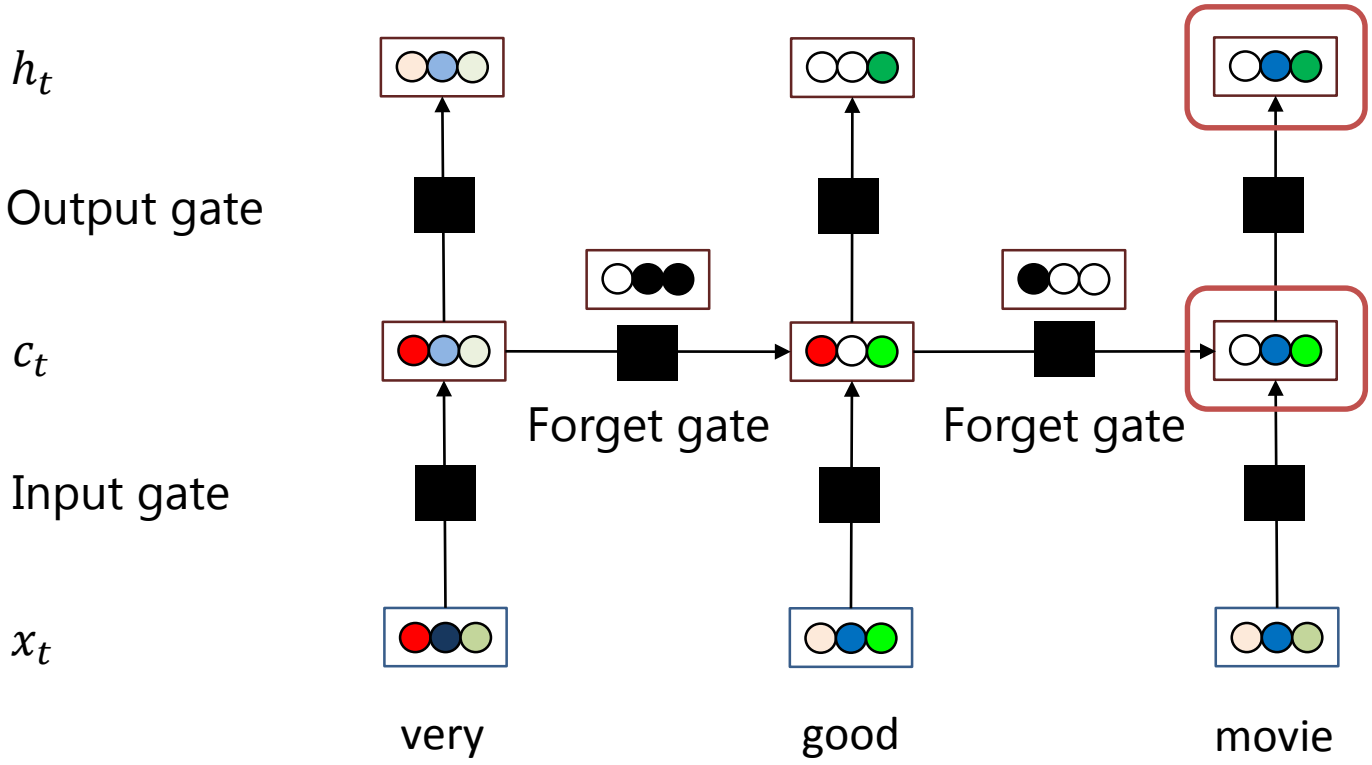
(Sutskever+ 2011)



潜在変数: $h_t = \sigma\left(W^{(hx)}x_t + W^{(hh)}h_{t-1} + b_h\right)$

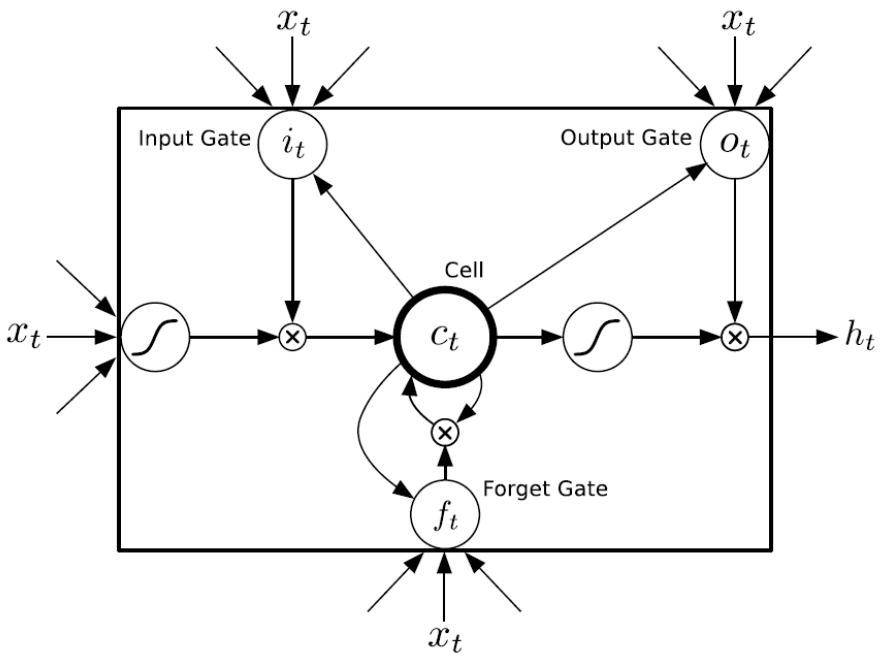
出力: $y_t = \sigma\left(W^{(yh)}h_t + b_y\right)$

Long Short-Term Memory (LSTM)
(Graves 2013) (単純化したもの)



- 各ゲートはマスクの役割を担う（ベクトルの要素ごとの積）
- 各ゲートのマスクパターンを入力x_t，記憶h_{t-1}，出力h_{t-1}などで制御する
- 長い系列での誤差逆伝搬時の勾配消失をゲートで防止する（→長期依存の保存）

LSTMもNNの一種



⊗（数式中は⊙）は要素ごとの積

Graves (2013)

Input gate: $i_t = \sigma\left(W^{(xi)}x_t + W^{(hi)}h_{t-1} + W^{(ci)}c_{t-1} + b_i\right)$

Forget gate: $f_t = \sigma\left(W^{(xf)}x_t + W^{(hf)}h_{t-1} + W^{(cf)}c_{t-1} + b_f\right)$

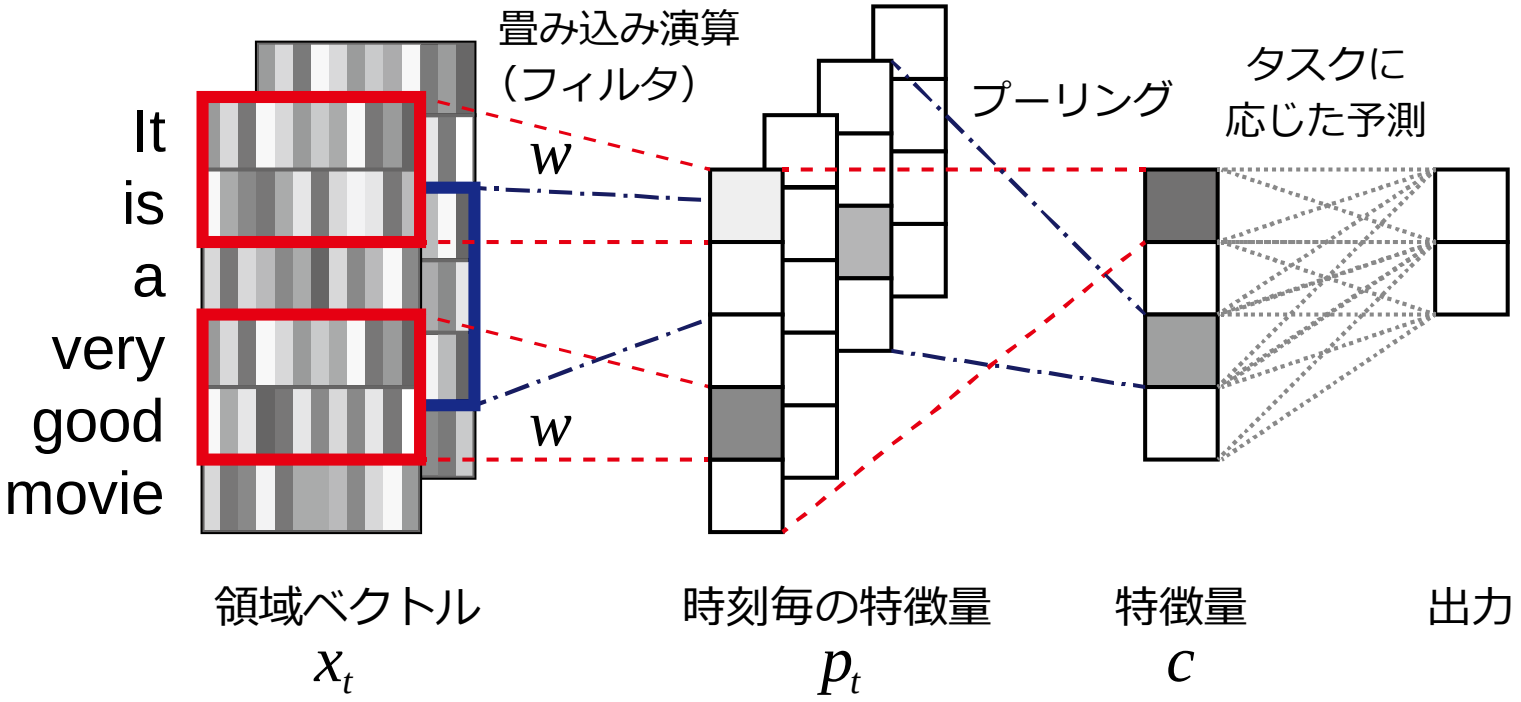
Cell: $c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(W^{(xc)}x_t + W^{(hc)}h_{t-1} + b_c\right)$

Output gate: $o_t = \sigma\left(W^{(xo)}x_t + W^{(ho)}h_{t-1} + W^{(co)}c_t + b_o\right)$

Hidden variable: $h_t = o_t \odot \tanh(c_t)$

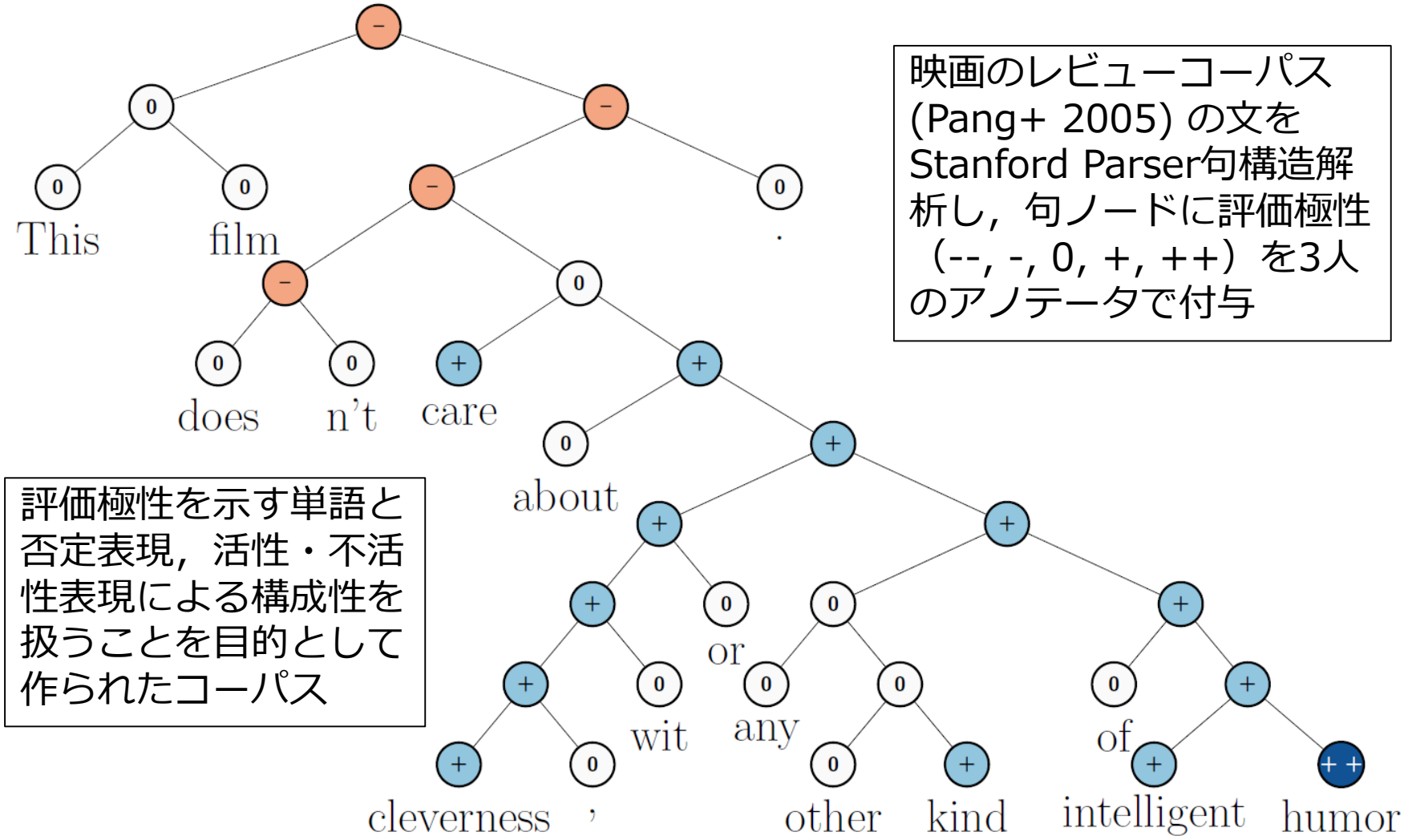
Convolutional Neural Network (CNN)

(Kim 14)



- 領域ベクトル: $x_{t:t+\delta} = x_t \oplus x_{t+1} \oplus \cdots \oplus x_{t+\delta-1}$
- 時間毎の特徴量: $p_t = g(w \cdot x_{t:t+\delta} + b)$
- 特徴量 (maxプーリング): $c = \max_{1<t<T-\delta+1} p_t$

Stanford Sentiment Treebank (Socher+ 2013)



映画のレビューコーパス (Pang+ 2005) の文を Stanford Parser句構造解析し，句ノードに評価極性（--, -, 0, +, ++）を3人のアノテータで付与

評価極性を示す単語と否定表現，活性・不活性表現による構成性を扱うことを目的として作られたコーパス

Stanford Sentiment Treebank での性能

Method	Fine-grained	Binary
RAE (Socher et al., 2013)	43.2	82.4
MV-RNN (Socher et al., 2013)	44.4	82.9
RNTN (Socher et al., 2013)	45.7	85.4
DCNN (Blunsom et al., 2014)	48.5	86.8
Paragraph-Vec (Le and Mikolov, 2014)	48.7	87.8
CNN-non-static (Kim, 2014)	48.0	87.2
CNN-multichannel (Kim, 2014)	47.4	**88.1**
DRNN (Irsoy and Cardie, 2014)	49.8	86.6
LSTM	46.4 (1.1)	84.9 (0.6)
Bidirectional LSTM	49.1 (1.0)	87.5 (0.5)
2-layer LSTM	46.0 (1.3)	86.3 (0.6)
2-layer Bidirectional LSTM	48.5 (1.0)	87.2 (1.0)
Dependency Tree-LSTM	48.4 (0.4)	85.7 (0.4)
Constituency Tree-LSTM		
– randomly initialized vectors	43.9 (0.6)	82.0 (0.5)
– Glove vectors, fixed	49.7 (0.4)	87.5 (0.8)
– Glove vectors, tuned	**51.0** (0.5)	88.0 (0.3)

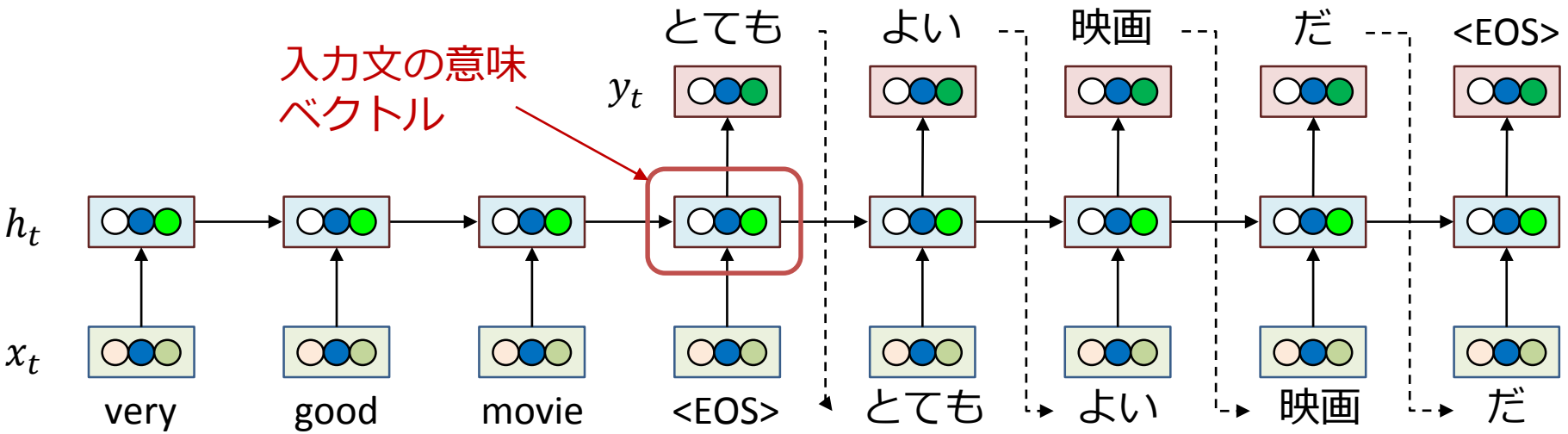
Tai+ (2015)

符号化・復号化
(Encoder-decoder models)

分散表現から文を生成できるか？

- 単語から句や文の分散表現を合成できる
- 句や文の分散表現から単語列を取り出せるか？
- 符号化・復号化 (encoder-decoder) モデル
  - 与えられた文の内容をベクトルで表現し，文を出力する
  - 単語列から単語列を予測できる
- DNNの応用がさらに広がる
  - 機械翻訳 (Sutskever+ 14, Cho+ 14, Luong+ 15)
  - 対話文生成 (Vinyals+ 15)
  - 自動要約 (Rush+ 15)
  - 画像説明文の生成 (Vinyals+ 14)

Sequence-to-sequence (Sutskever+ 14)



- 英語 (x_t) から日本語 (y_t) への機械翻訳の例
  - $h_t = \sigma(W^{(hx)}x_t + W^{(hh)}h_{t-1})$
  - $y_t = W^{(yh)}h_t$ (y_tは出力単語のスコアのベクトル表現)
- 出力単語を入力側に戻すことで，翻訳履歴を考慮
- 実際にはRNNではなく2-layer LSTMを用いる

対話の生成 (Vinyals+ 15)

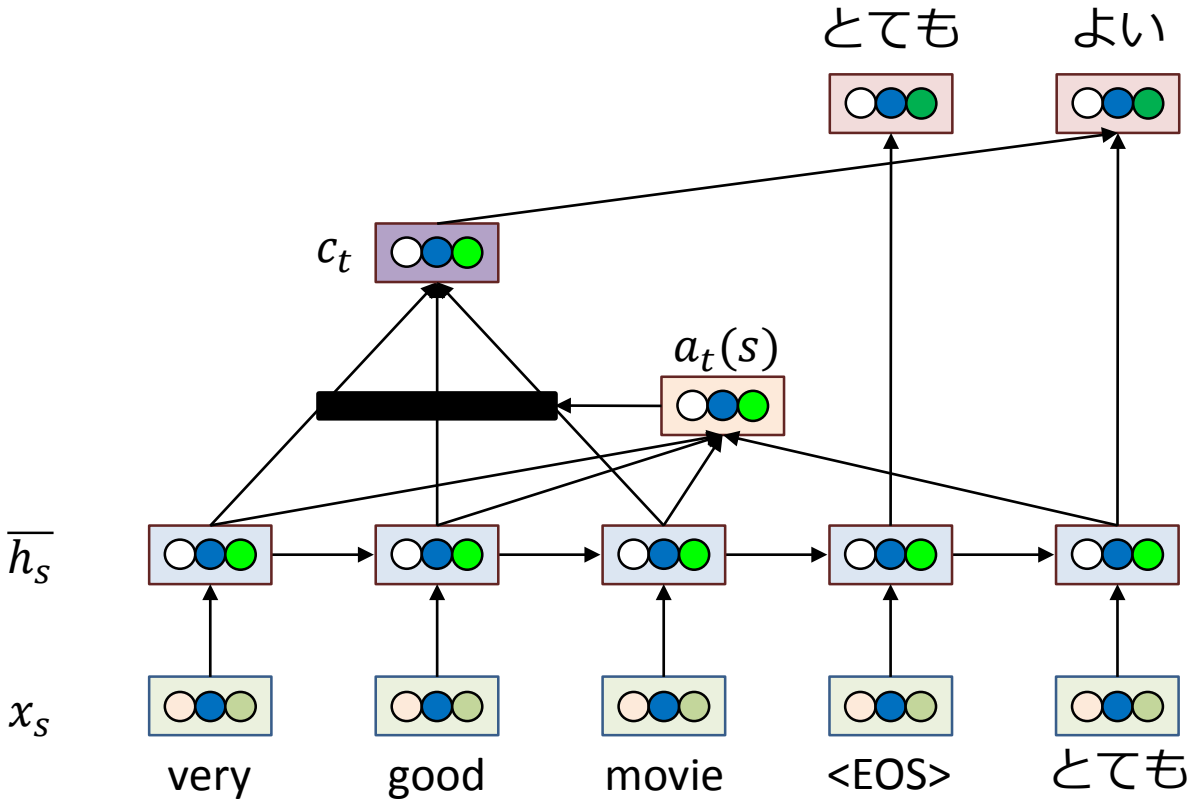
- 学習データ: OpenSubtitles
  - 映画の字幕から抽出した対話文（約6200万文）
- Sequence-to-sequenceで生成した例

Human: *hello !*
Machine: *hello !*
Human: *how are you ?*
Machine: *i 'm good .*
Human: *what 's your name ?*
Machine: *i 'm julia .*
Human: *when were you born ?*
Machine: *july 20th .*
Human: *what year were you born ?*
Machine: *1977 .*
Human: *where are you from ?*
Machine: *i 'm out in the boonies.*
Human: *are you a man or a woman ?*
Machine: *i 'm a woman .*
Human: *why are we here ?*
Machine: *i 'm not sure .*
Human: *okay, bye !*
Machine: *bye .*
Human: *see you later .*
Machine: *bye .*

Attention mechanism

(Bahdanau+ 15, Luong+, 15)

- 固定長のベクトルで文の意味をエンコードするには限界がある
- 位置tの単語を生成するときに，入力のどの単語に着目すべきかの重み$a_t(s)$を求め，入力単語のベクトル$\overline{h_s}$の重み付き平均ベクトルc_tも用いて出力単語を生成する



$$\tilde{h_t} = \tanh(W_c[c_t; h_t])$$

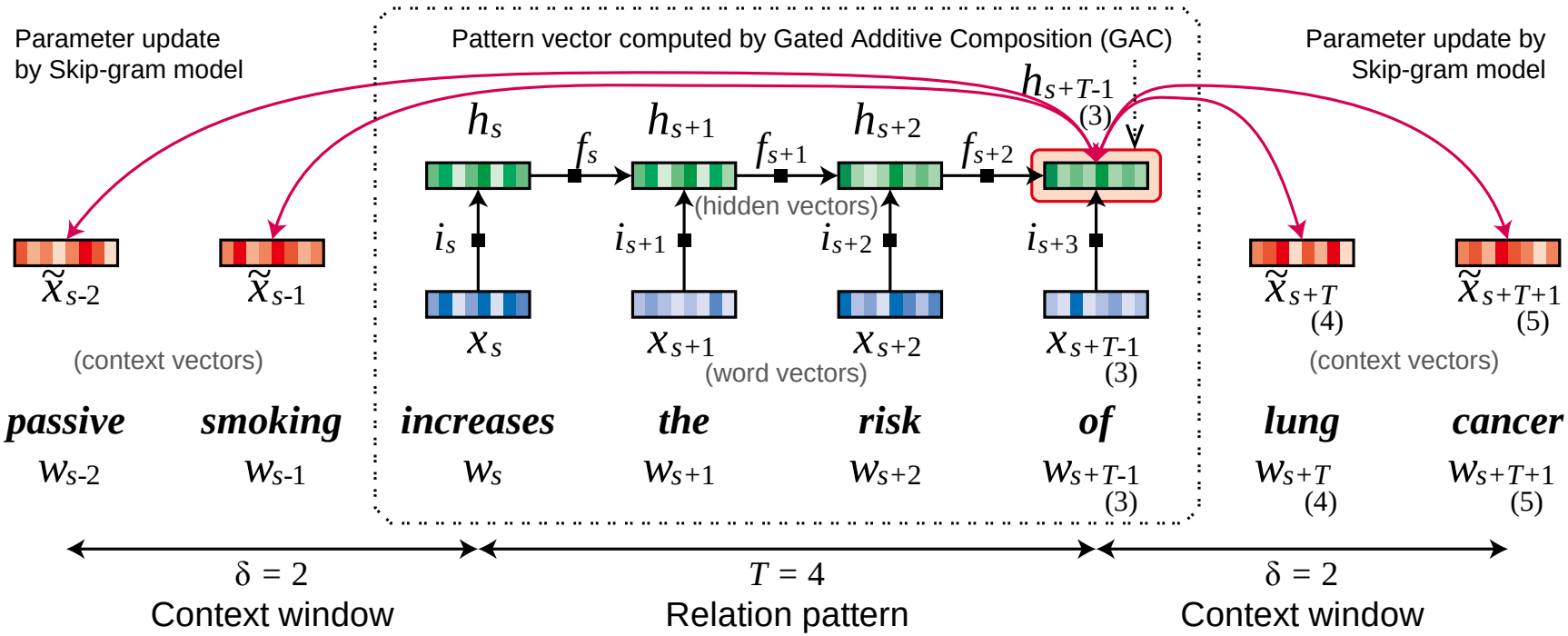
$$c_t = \sum_s a_t(s)\overline{h_s}$$

$$a_t(s) = \frac{\exp \text{score}(h_t, \overline{h_s})}{\sum_{s'} \exp \text{score}(h_t, \overline{h_{s'}})}$$

$$\text{score}\left(h_t, \overline{h_s}\right) = h_t \cdot \overline{h_s}$$

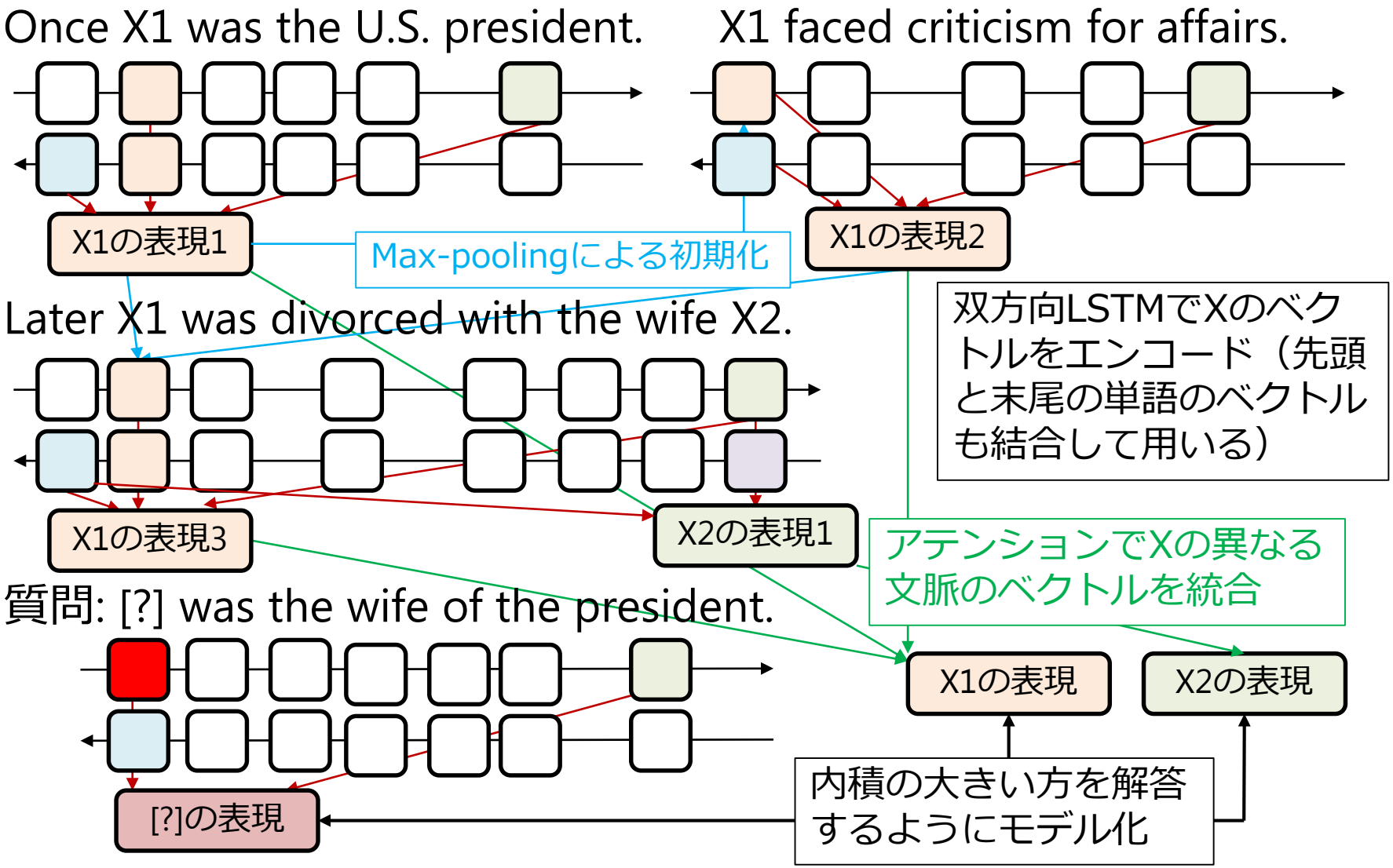
研究室での取り組み

関係パタンの意味合成 (Takase+ 2016)



- 因果関係などの「関係」を表す言語表現の分散表現を，構成単語の分散表現から合成する手法を提案
- 関係の分散表現の評価データを作成・公開

動的分散表現による読解 (Kobayashi+ 2016)



(D)NNを用いたその他の研究

- 構文解析のための分散表現の学習 (Komatsu+, 15)
- Dependency-based Compositional Semanticsと加法構成性に基づく意味合成 (Tian+, 16)
- CNNによるツイートの賛否分類 (Igarashi+, 16)
- Wikipedia記事の意味カテゴリ推定 (Suzuki+, submitted)
- 画像内の物体間の関係抽出 (Muraoka+, submitted)

まとめ

- 言語処理における(D)NN研究の最近の流れ
  - 2013年-: 単語の分散表現
  - 2011年-: 構成性に基づく句の分散表現
  - 2014年-: 符号化・復号化モデル
  - 2015年-: アテンション・メカニズム
  - 2016年-: ？
- 今後DNNで取り組みたいと思っているテーマ
  - 文脈のモデリングとその応用
  - 分散表現による知識の表現と推論

さらに詳しく知りたい方へ

岩波DS Vol. 2 (2016年2月)

人工知能 2016年3月号特集記事

参考文献 (1/2)

- Bahdanau, D., Cho, K., and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate, in ICLR (2015)
- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C.: A Neural Probabilistic Language Model, Journal of Machine Learning Research, Vol. 3, pp. 1137–1155 (2003)
- Cho, K., Merrienboer, van B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in Proc. of EMNLP, pp. 1724–1734 (2014)
- Collobert, R. and Weston, J.: A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning, in Proc of ICML, pp. 160–167 (2008)
- Graves, A.: Generating Sequences With Recurrent Neural Networks, CoRR, Vol. abs/1308.0850, (2013)
- Hinton, G., McClelland, J., and Rumelhart, D.: Distributed representations, in Rumelhart, D. E., McClelland, J. L., and Group, P. R. eds., Parallel distributed processing: Explorations in the microstructure of cognition, Vol. I, chapter 3, pp. 77–109, MIT Press, Cambridge, MA (1986)
- Igarashi, Y., Komatsu, H., Kobayashi, S., Okazaki, N., Inui, K.: Feature-based Model versus Convolutional Neural Network for Stance Detection, in Proc. of SemEval (2016)Kim, Y.: Convolutional Neural Networks for Sentence Classification, in Proc. of EMNLP, pp. 1746–1751 (2014)
- Kobayashi, K., Tian, R., Okazaki, N., Inui, K.: Dynamic Entity Representation with Max-pooling Improves Machine Reading, in Proc. of NAACL (2016)
- Komatsu, H., Tian, R., Okazaki, N., Inui, K.: Reducing Lexical features in Parsing by Word Embeddings. in Proc. of PACLIC, pp.106-113 (2015)
- Luong, M.-T., Pham, H., Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, in Proc. of EMNLP, pp. 1412-1421 (2015)

参考文献 (2/2)

- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed Representations ofWords and Phrases and their Compositionality, in Proc. of NIPS, pp. 3111–3119 (2013)
- Rush, A. M., Chopra, S., Weston, J.: A Neural Attention Model for Sentence Summarization
- , in Proc. of EMNLP, pp. 379-389 (2015)
- Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D.: Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions, in Proc. of EMNLP, pp. 151–161 (2011)
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C.: Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, in Proc. of EMNLP, pp. 1631–1642 (2013)
- Sutskever, I., Martens, J., and Hinton, G.: Generating Text with Recurrent Neural Networks, in Proc. of ICML, pp. 1017–1024 (2011)
- Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, in Proc. of NIPS, pp. 3104–3112 (2014)
- Tai, K. S., Socher, R., and Manning, C. D.: Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, in Proc. of ACL-IJCNLP, pp. 1556–1566 (2015)
- Takase, S., Okazaki, N., Inui, K.: Composing Distributed Representations of Relational Patterns, in Proc. of ACL (2016)
- Tian, R., Okazaki, N., Inui, K.: Learning Semantically and Additively Compositional Distributional Representations, in Proc. of ACL (2016)
- Vinyals, O., Le, Q. V., A Neural Conversational Model, in Proc. of ICML Deep Learning Workshop, (2015)
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D.: Show and tell: A neural image caption generator, in Proc. of CVPR (2015)
- Weston, J., Chopra, S., Bordes, A.: Memory Networks, in Proc. of ICLR, (2015)