

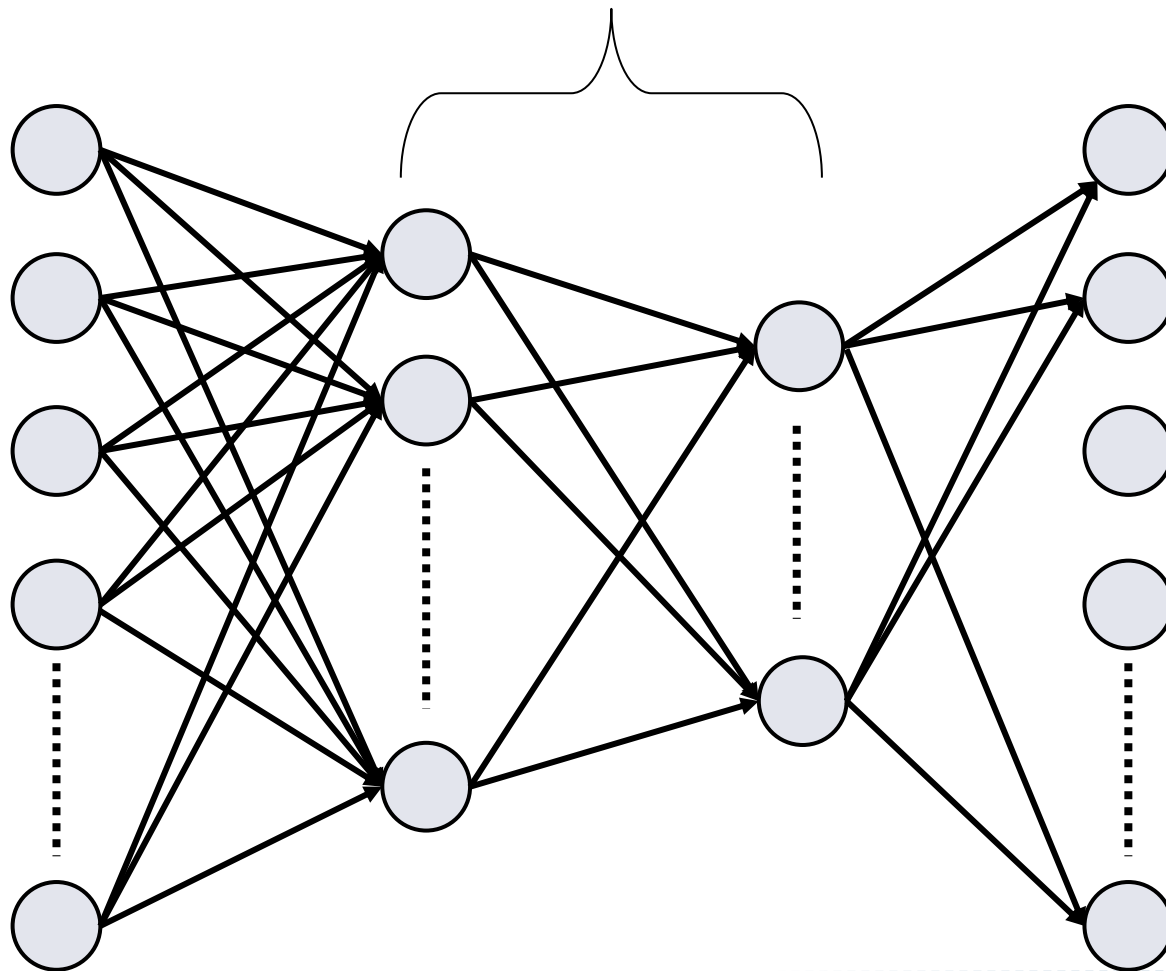
14. 深層学習と自然言語処理

多重パーセプトロン

入力層

隠れ層

出力層



単語のベクトル表現（その1）

一つの要素だけ
「1」

one-hotベクトル表現

使用する単語
...
机
コンピュータ
キーボード
本
ノート
...

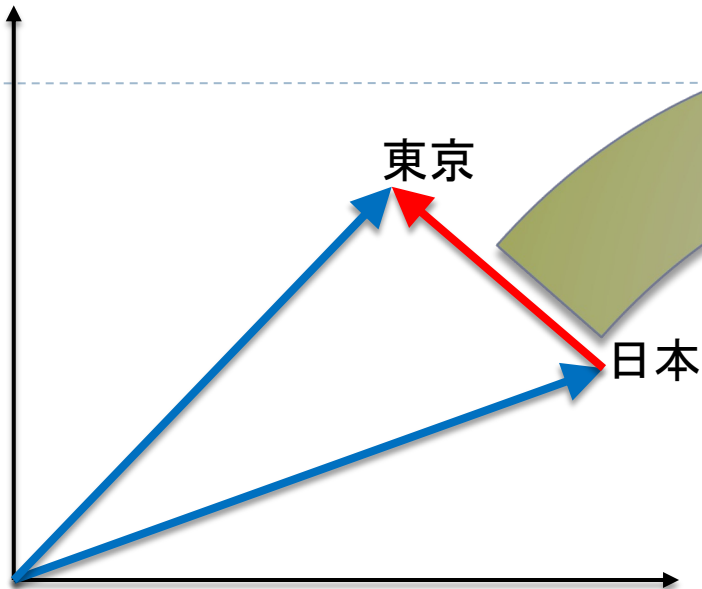
...	机	コンピュータ	キーボード	本	ノート	...
...
..	1	0	0	0	0	..
..	0	1	0	0	0	..
..	0	0	1	0	0	..
..	0	0	0	1	0	..
..	0	0	0	0	1	..
...



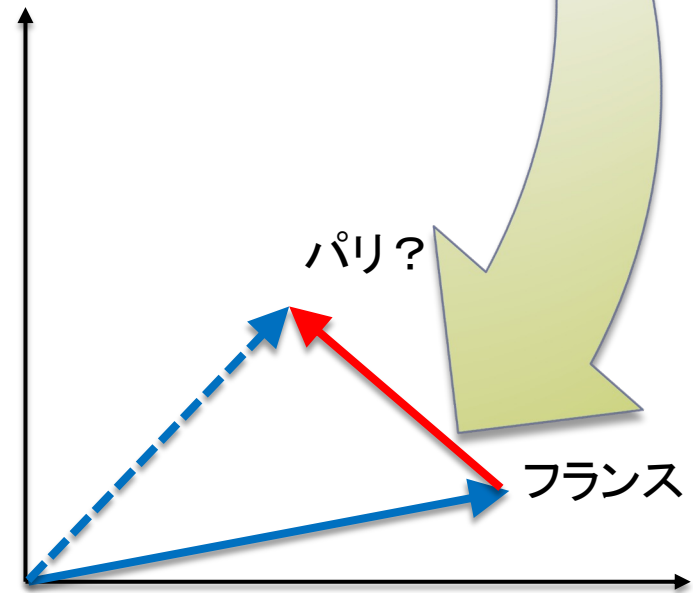
単語の個数 = ベクトルの次元

単語 ベクトル表現
「コンピュータ」 = (0,0,0,...,0,1,0,0,0,...)
「キーボード」 = (0,0,0,...,0,0,1,0,0,...)

word2vec (発想)

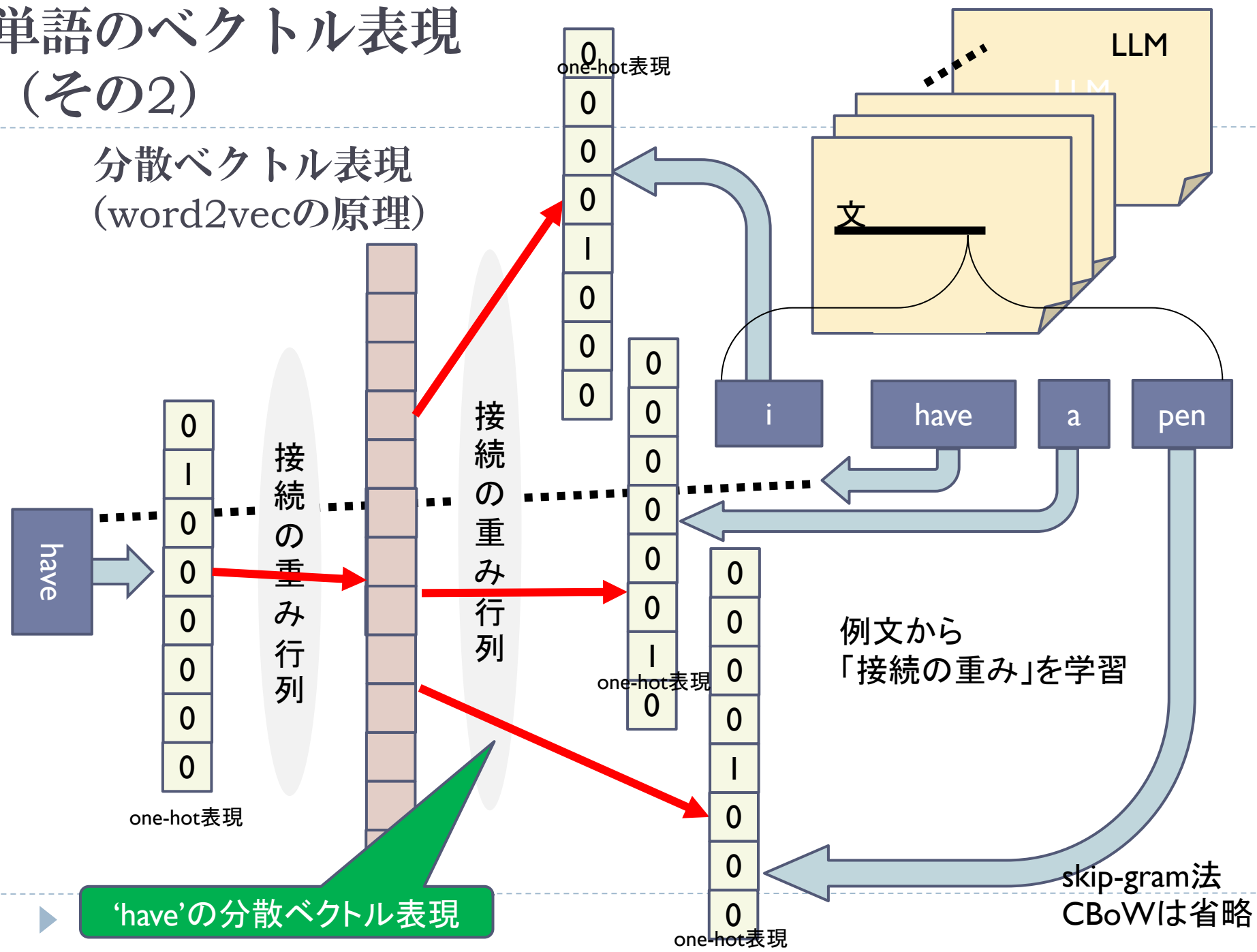


$$? \quad V_{パリ} = V_{フランス} + (V_{東京} - V_{日本})$$

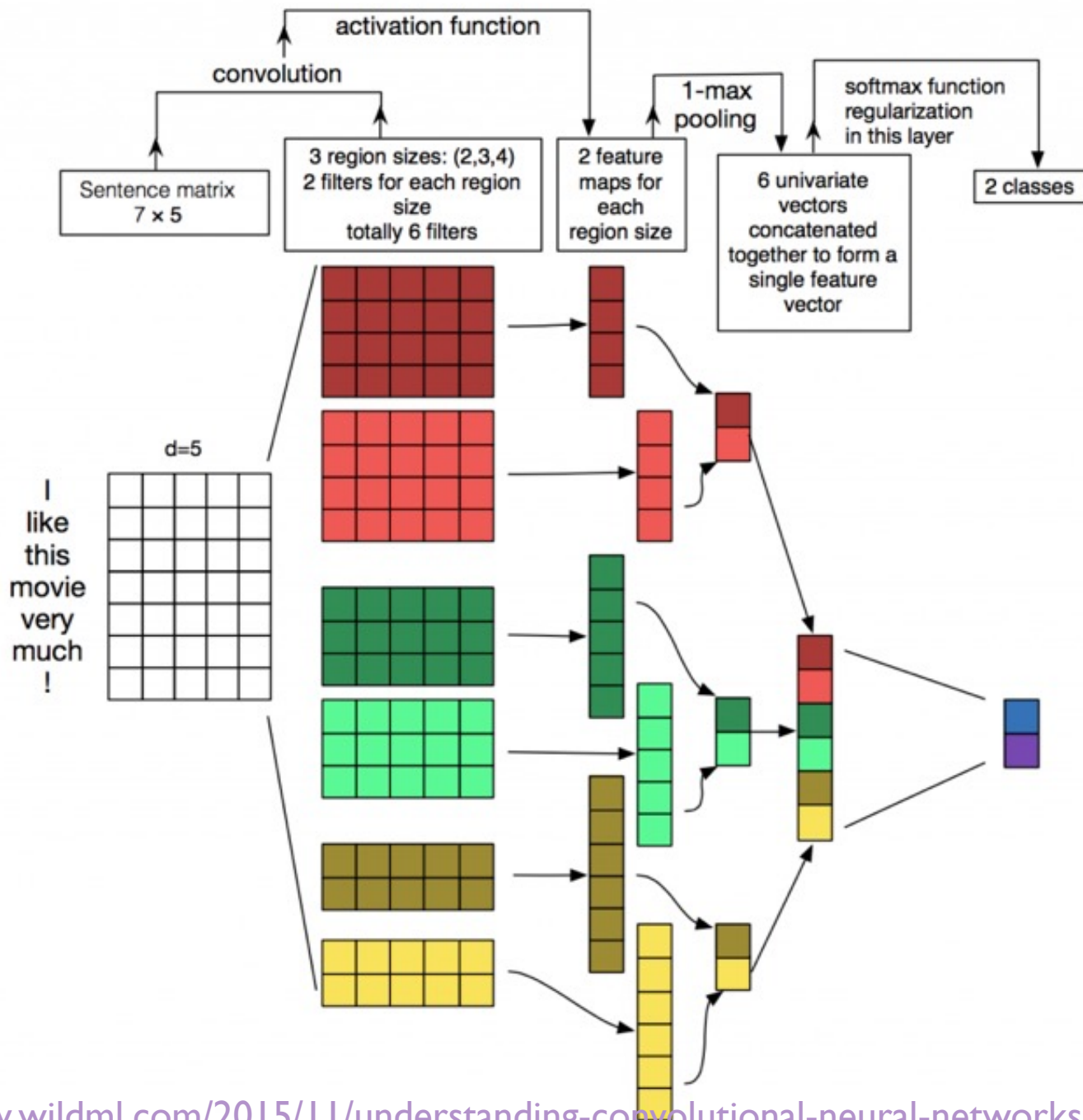


単語のベクトル表現 (その2)

分散ベクトル表現
(word2vecの原理)

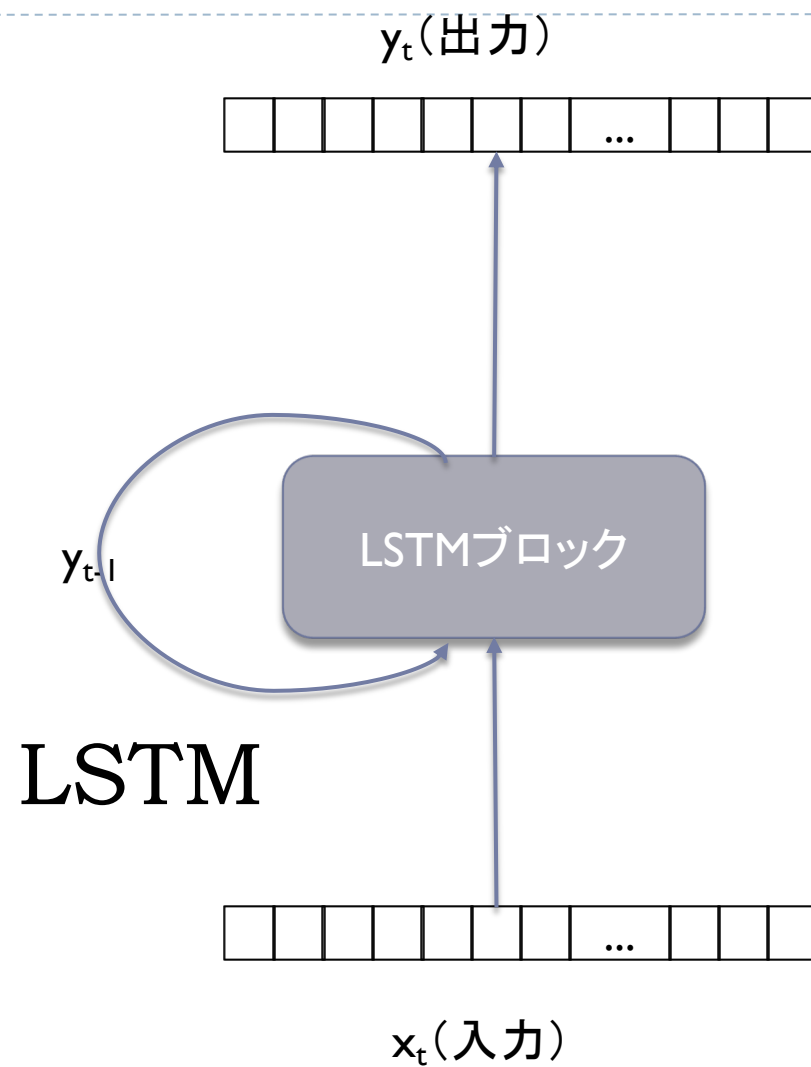
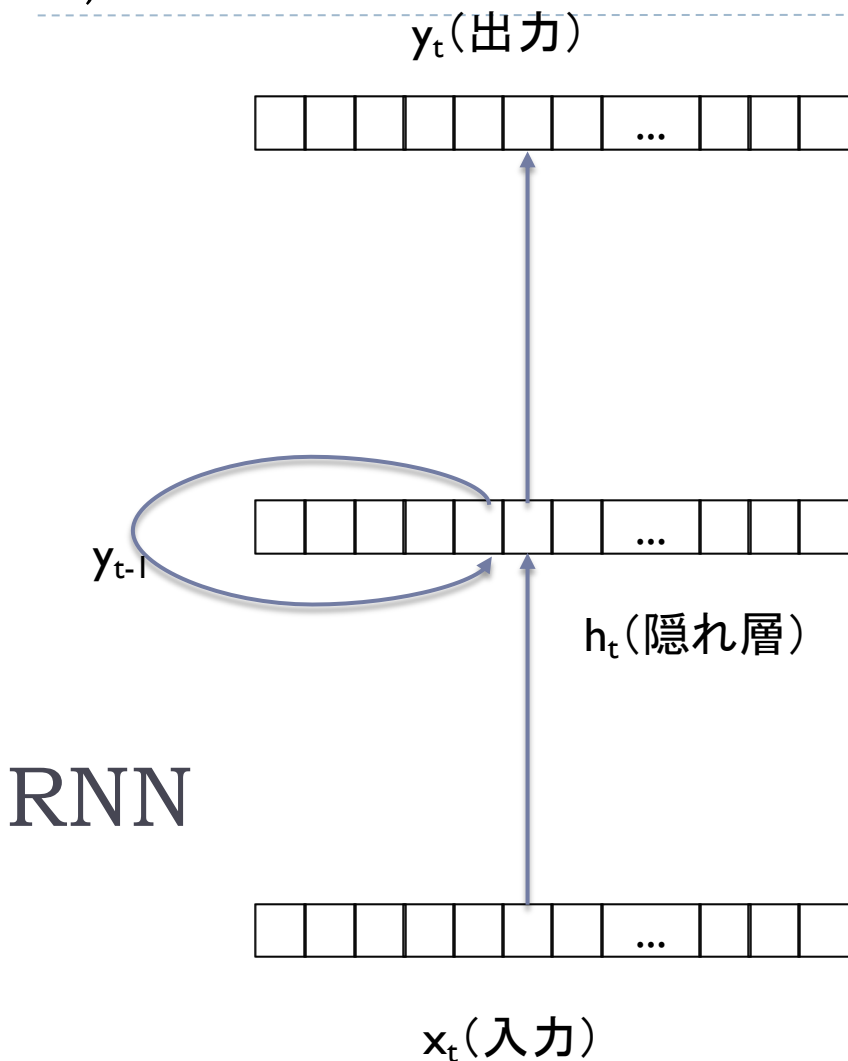


CNN

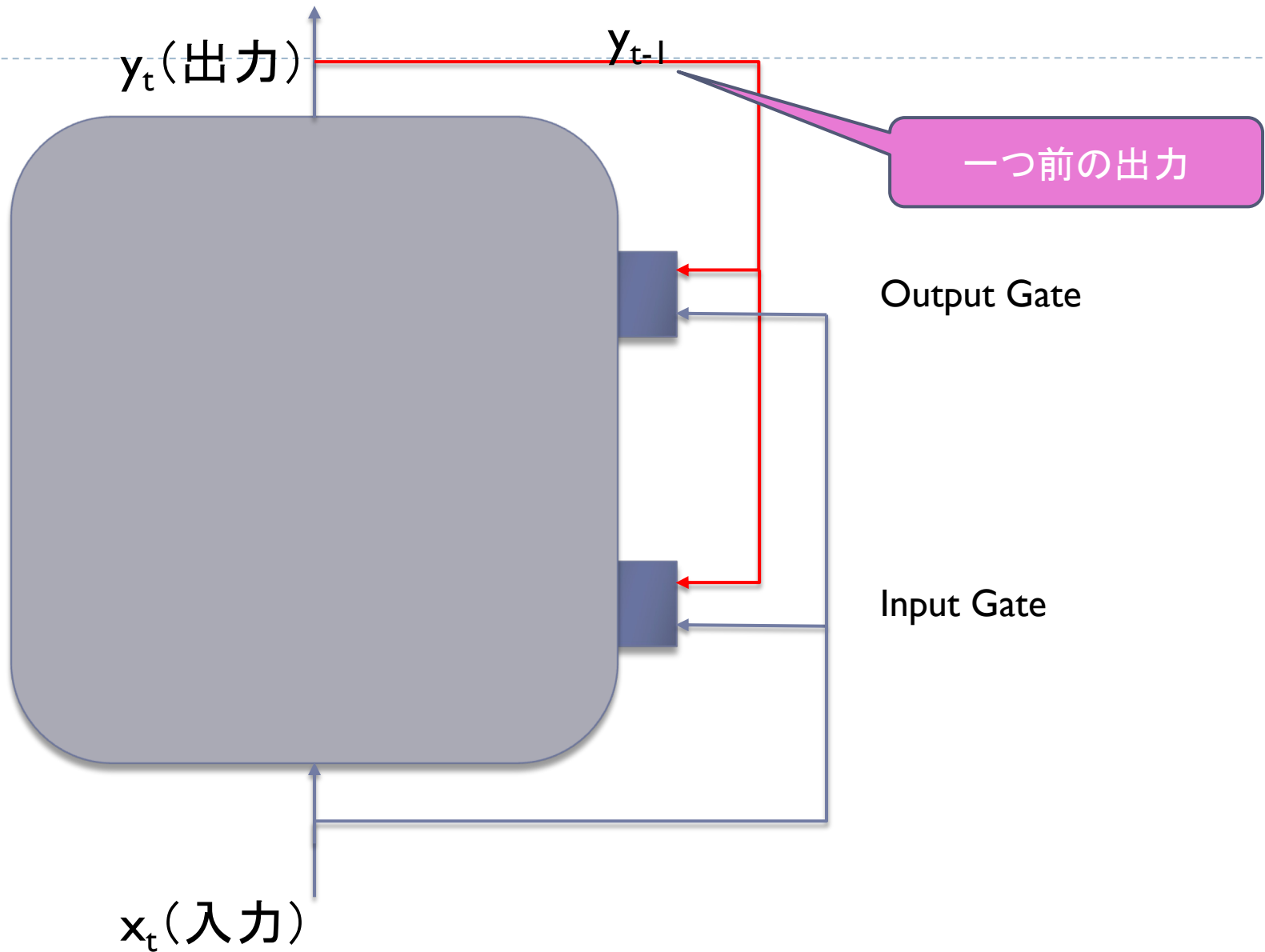


深層学習による自然言語処理

RNN(recurrent NN)とLSTM(long short time memory)、CNN(convolutional NN)

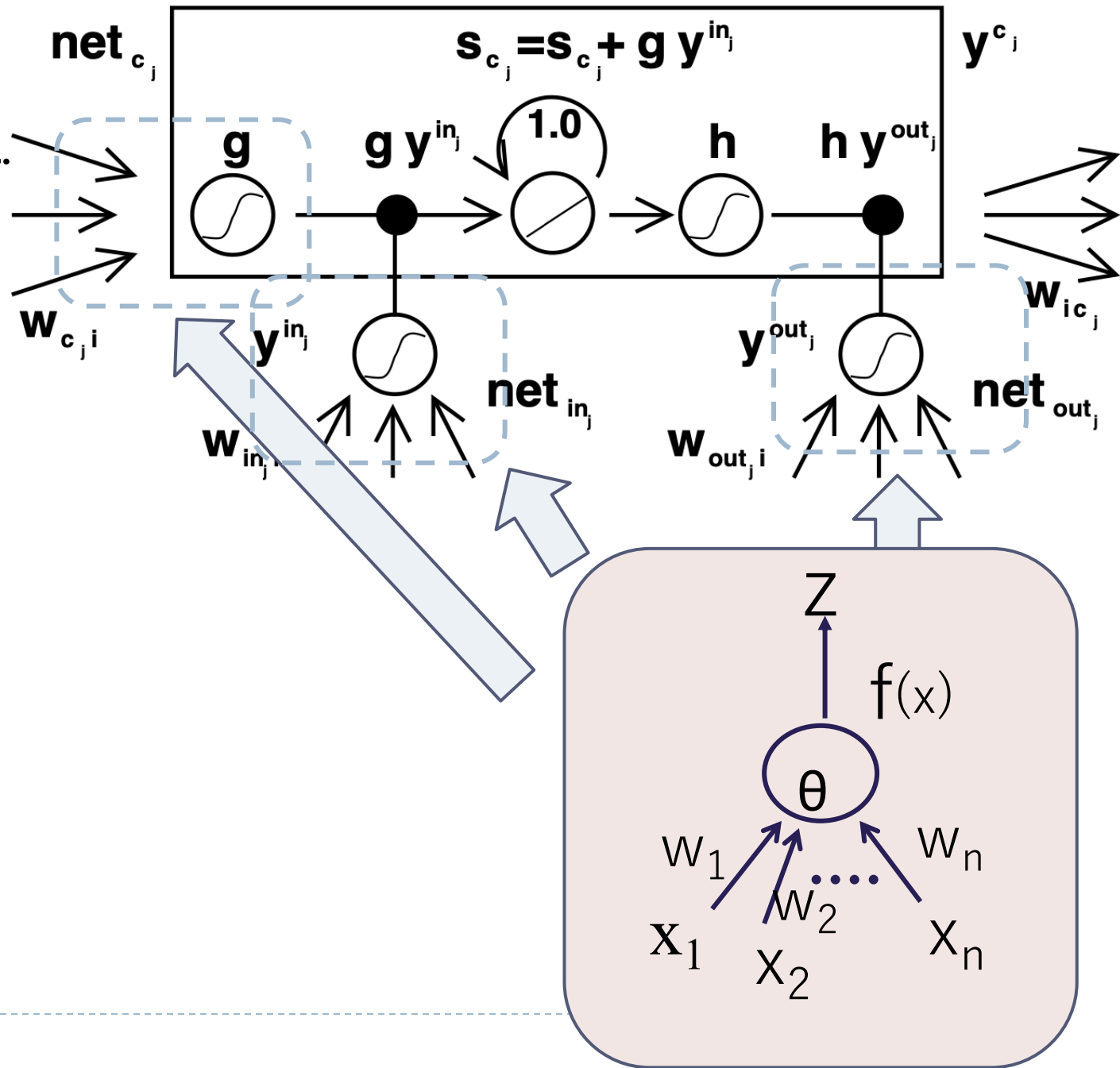


LSTMブロック



LSTM

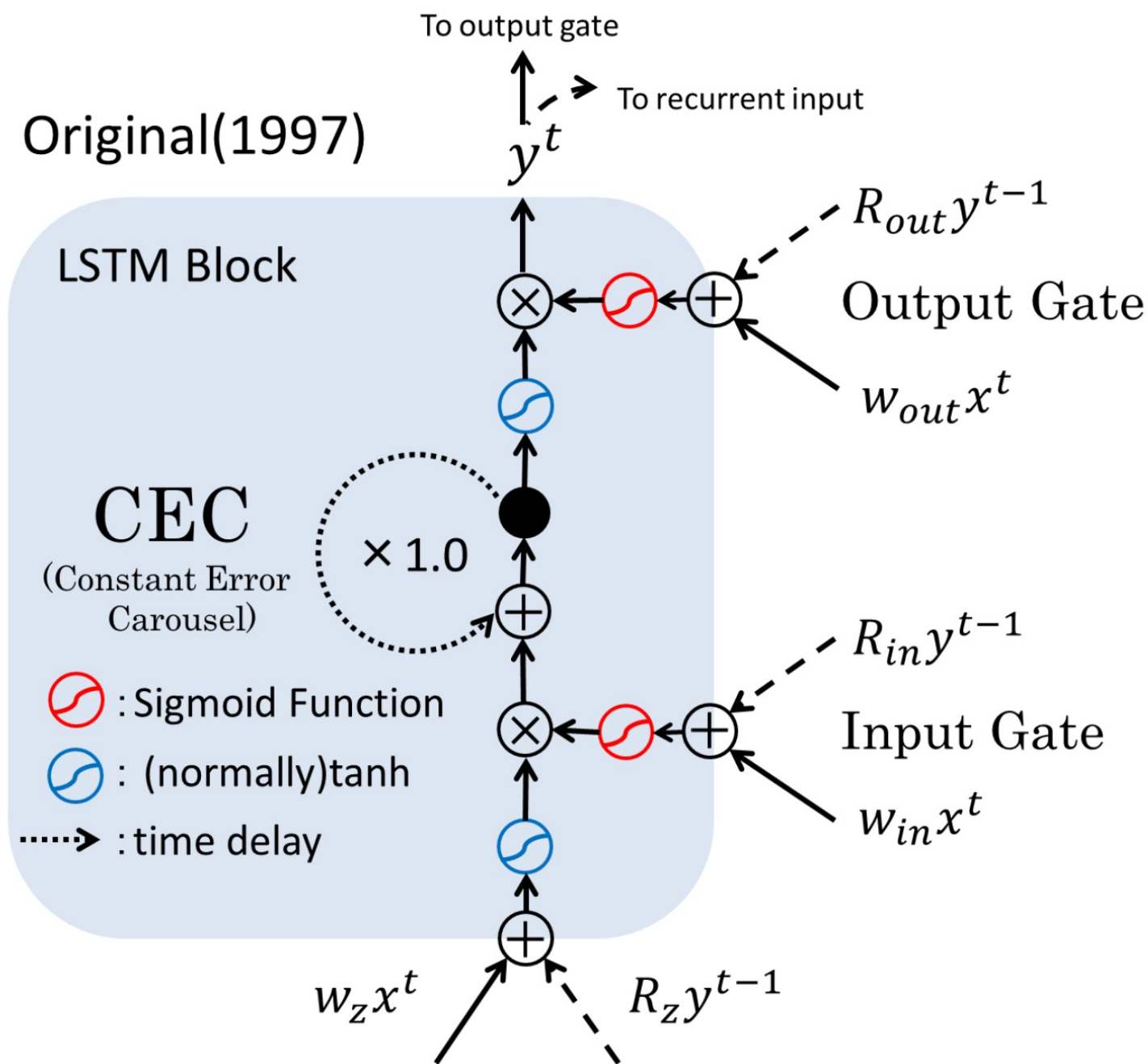
S.Hochreiter, et.
al(1997)



LSTM

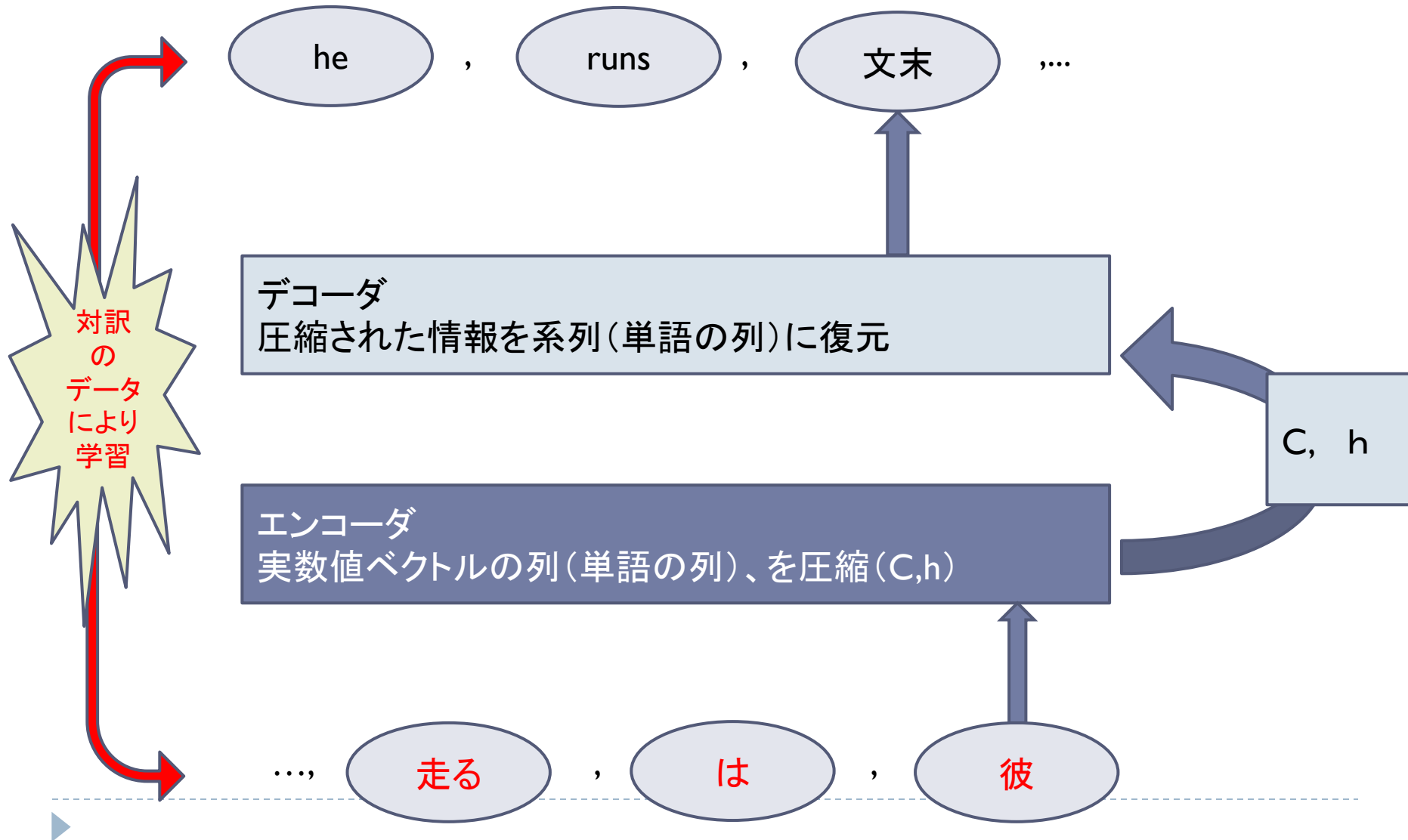
https://qiita.com/t_Signull/items/21b82be280b46f467d1b

Original(1997)



ニューラル機械翻訳

原理: encoder/decoder (系列の学習)



言語資源: Wikiペディアからダウンロード(2.7GB)

jawiki-latest-pages-articles.xml-012.txt
jawiki-latest-pages-articles.xml-013.txt
jawiki-latest-pages-articles.xml-014.txt
jawiki-latest-pages-articles.xml-015.txt
jawiki-latest-pages-articles.xml-016.txt
jawiki-latest-pages-articles.xml-017.txt
jawiki-latest-pages-articles.xml-018.txt
jawiki-latest-pages-articles.xml-019.txt
jawiki-latest-pages-articles.xml-020.txt
jawiki-latest-pages-articles.xml-021.txt
jawiki-latest-pages-articles.xml-022.txt
jawiki-latest-pages-articles.xml-023.txt
jawiki-latest-pages-articles.xml-024.txt
jawiki-latest-pages-articles.xml-025.txt
jawiki-latest-pages-articles.xml-026.txt
jawiki-latest-pages-articles.xml-027.txt
jawiki-latest-pages-articles.xml-028.txt
jawiki-latest-pages-articles.xml-029.txt
jawiki-latest-pages-articles.xml-030.txt
jawiki-latest-pages-articles.xml-031.txt
jawiki-latest-pages-articles.xml-032.txt
jawiki-latest-pages-articles.xml-033.txt
jawiki-latest-pages-articles.xml-034.txt
jawiki-latest-pages-articles.xml-035.txt
jawiki-latest-pages-articles.xml-036.txt
jawiki-latest-pages-articles.xml-037.txt
jawiki-latest-pages-articles.xml-038.txt
jawiki-latest-pages-articles.xml-039.txt
jawiki-latest-pages-articles.xml-040.txt
jawiki-latest-pages-articles.xml-041.txt
jawiki-latest-pages-articles.xml-042.txt
jawiki-latest-pages-articles.xml-043.txt

[[バイキング]]

#REDIRECT [[ヴァイキング]]

[[ロマン主義]]

ロマン主義（ロマンしゅぎ、Romanticism、Romantisme、Romantik、Romanticismo、Romanticismo、Romantismo）は、主として18世紀末から19世紀前半にヨーロッパで、その後ヨーロッパの影響を受けた諸地域で起こった精神運動の一つである。それまでの理性偏重、合理主義などに対し感受性や主観に重きをおいた一連の運動であり、古典主義と対をなす。恋愛賛美、民族意識の高揚、中世への憧憬といった特徴をもち、近代国民国家形成を促進した。その動きは文芸・美術・音楽・演劇など様々な芸術分野に及んだ。のちに、その反動として写実主義・自然主義などをもたらした。

==概要==

ロマン主義は教条主義、古典主義の対概念としてとらえられるもので、アメリカの哲学者・アーサー・ラブジョイ（:en:Arthur Oncken Lovejoy）は「ロマン主義の時代」を1780年から1830年としている。また、ロマン主義は産業革命の大きな原動力となった。その萌芽は既にベルナルダン・ド・サン＝ピエールやディドロに見られ、セナンクール、スタール夫人、パンジャマン・コンスタン、フランソワ＝ルネ・ド・シャトーブリアンといった初期ロマン派作家によってそれまで教条主義によって抑圧されてきた個人の根本的独自性を根本とした表現が特徴とされる。これらはナポレオン1世の第一帝政に対する文化的抵抗運動の中で文芸サロンやサークルの

jawiki-latest-pages-articles.xml-022.txt

標準テキスト書類 - 10.6 MB

タグ タグを追加...

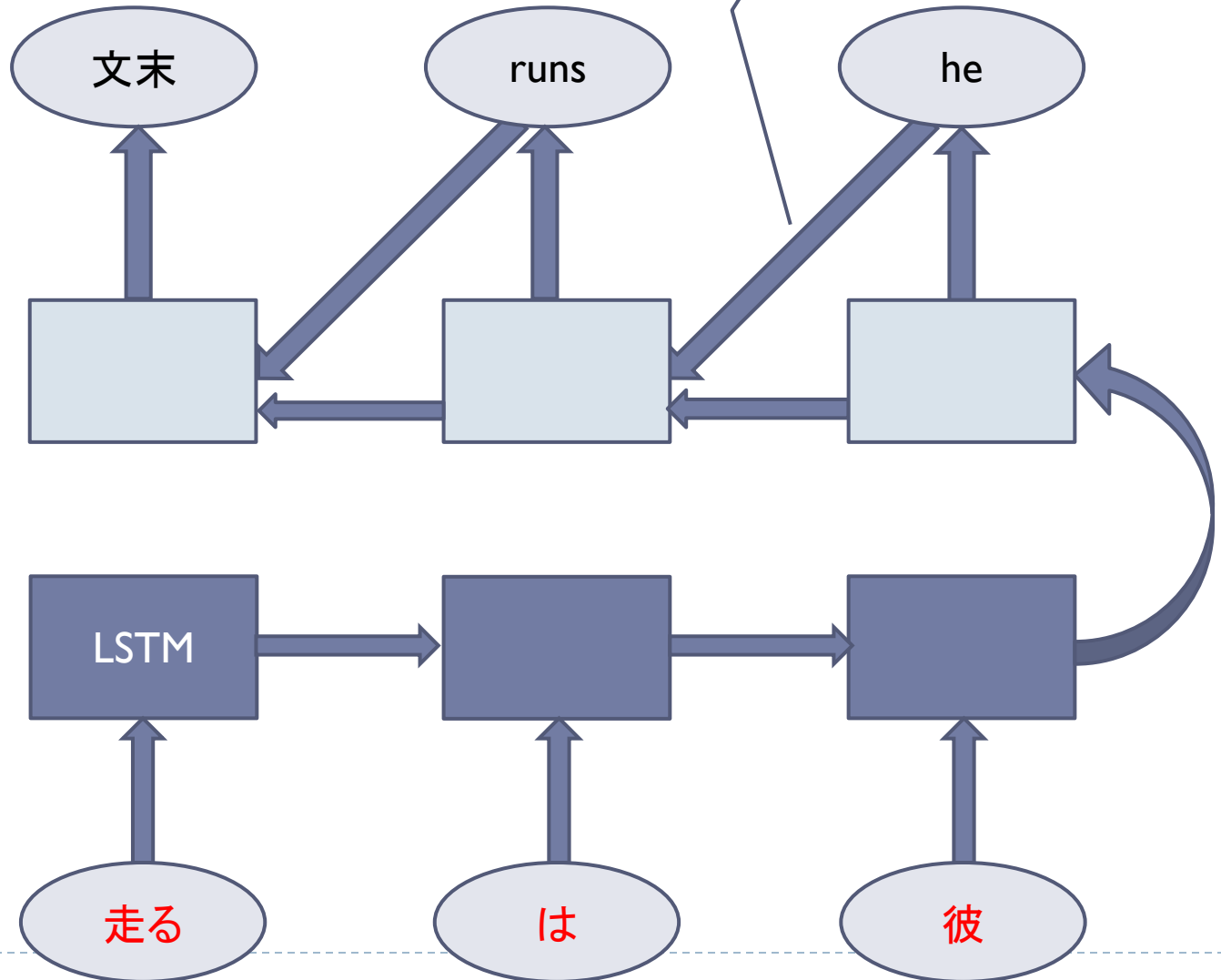
作成日 2017年4月11日 火曜日 20:40

変更日 2017年4月11日 火曜日 20:40

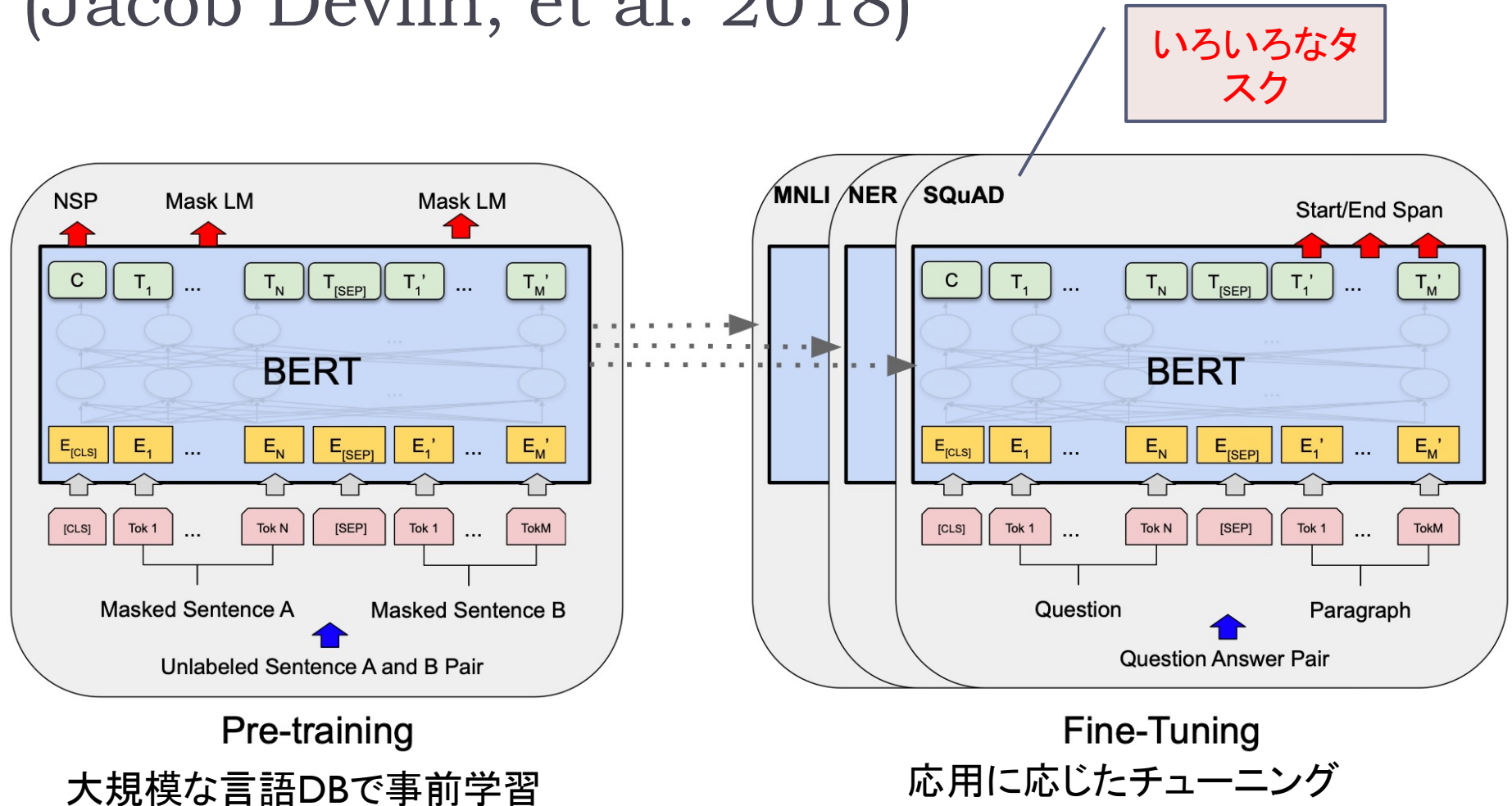
ニューラル機械翻訳

内部情報だけで
無く、表層の単
語も使う

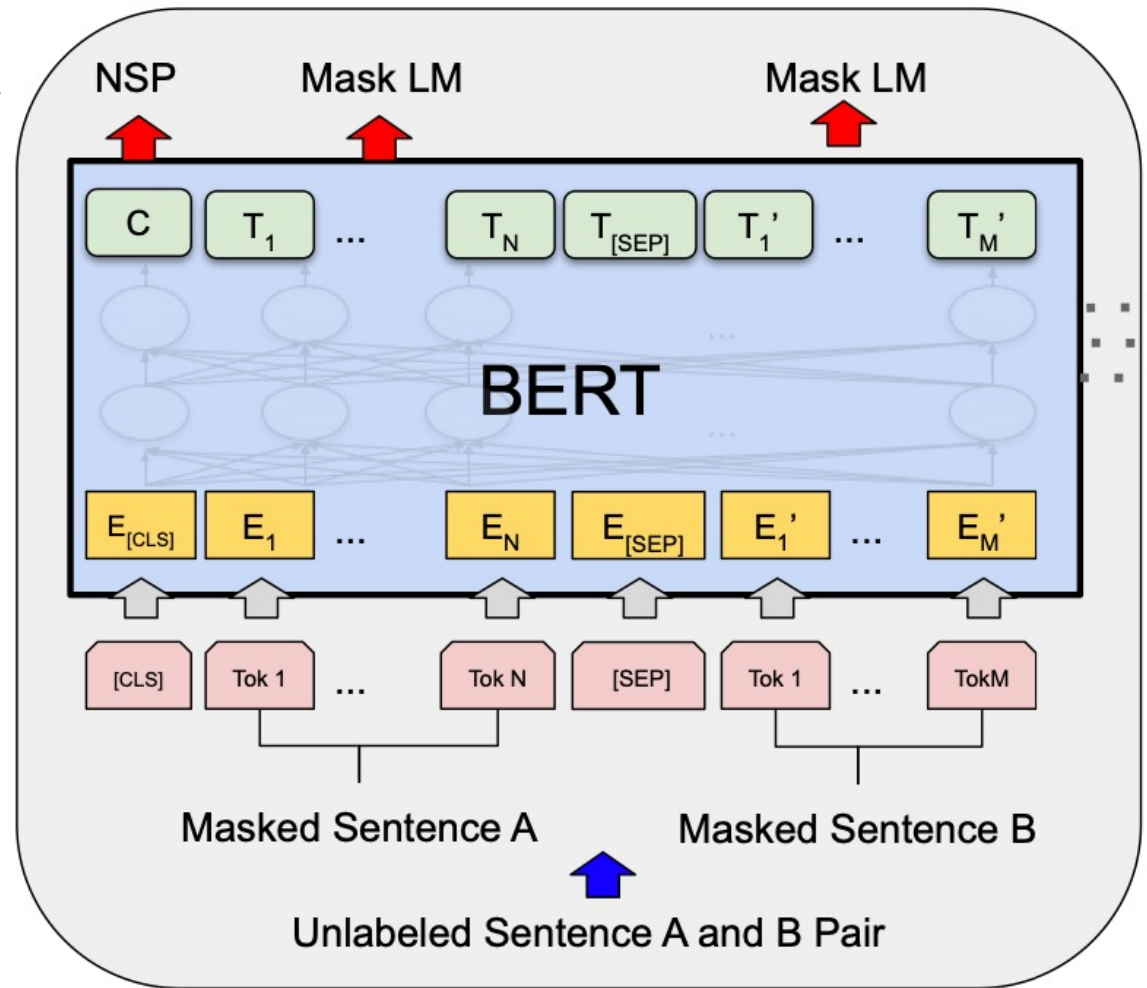
デコーダ
(LSTMI個で実装)



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Jacob Devlin, et al. 2018)



Pre-Training



Pre-training

- ・事前に学習
- ・共通
- ・公開されているものを利用可能

Pre-Training

タスク1)

NSP(next sentence prediction):

2文の連続性(isNext/notNext)についての訓練

タスク2)

MLM(masked language modeling):

いくつかの単語を隠し、それを当てる

ex. This is my pen. That is your pen.

入力: [CLS] this is my pen [SEP] that is [MASK] pen [SEP]

出力: (isNext) this is my pen [SEP] that is your pen [SEP]

タスク1の正解

タスク2の正解

訓練

- TPU4つで4日間
- BooksCorpus(単語数8億)
- English wikipedia(単語数25億)

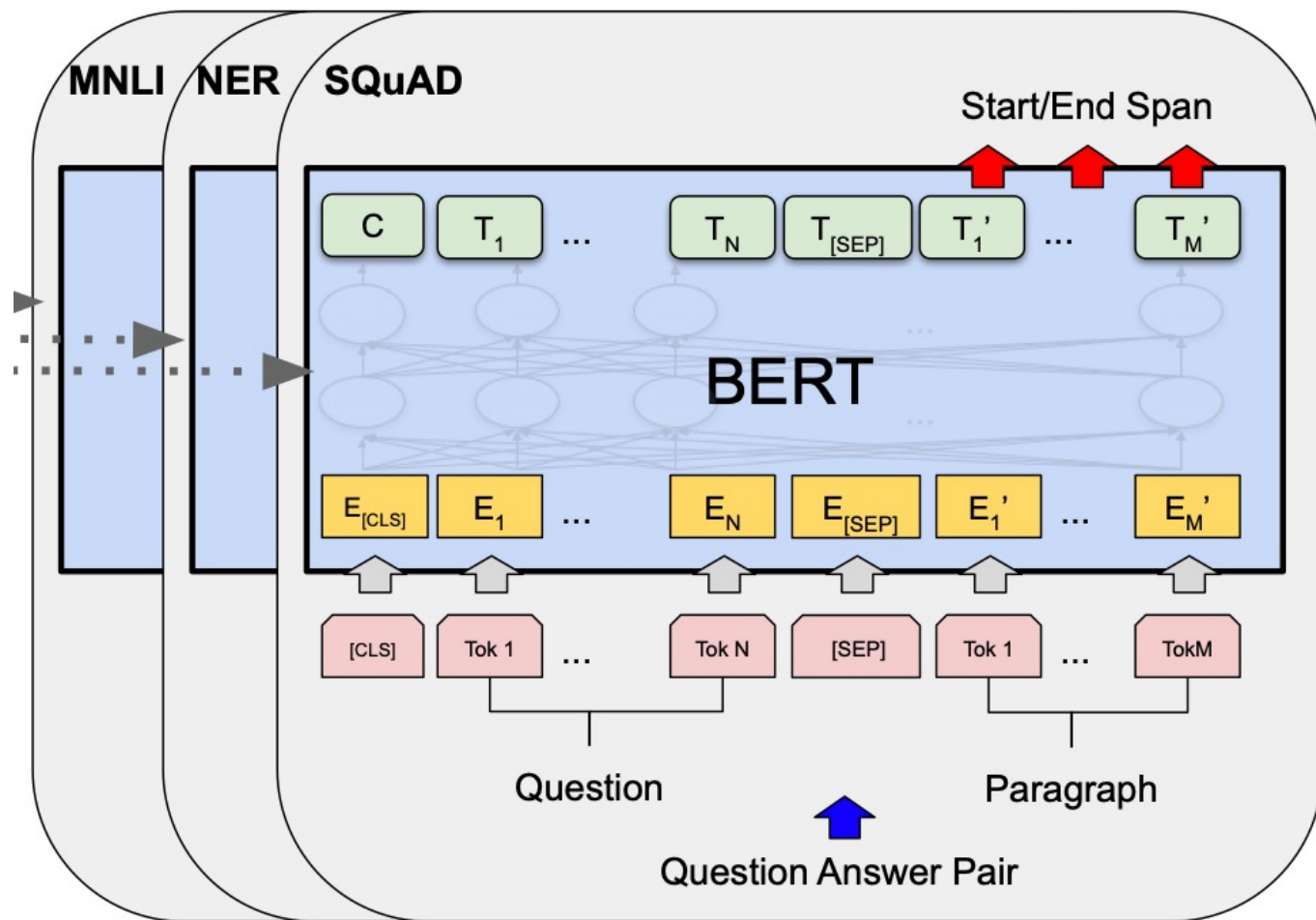
Fine Tuning

・更に応用毎に学習

SQuAD: ウィキペディアから質問の答えとなるテキストを見つける。

MNLI: 代名詞が置換された文が元の文に含まれているか判定

NER: 固有表現抽出



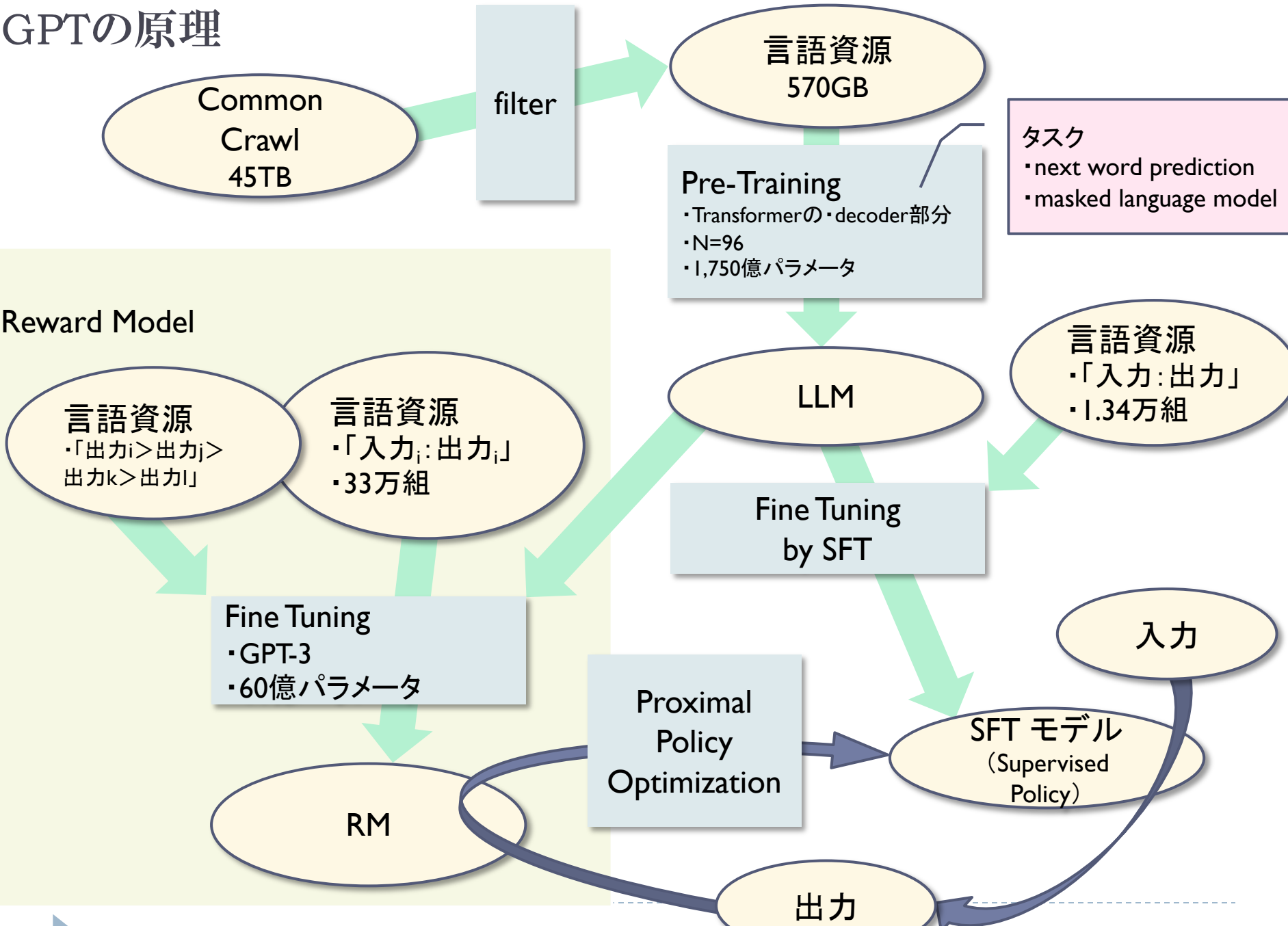
Fine-Tuning

GPT概要

人間どうしの対話を模倣。音楽、小説などの創作ができる。

- ▶ GPT-3 → (Alignment) → InstructGPT → (対話特化) → ChatGPT
 - ▶ Alignment: 人間の好み、道徳性に対応すること
 - ▶ 対話特化: モデル: GPT-3.5, データ: 会話データ
- ▶ GPT-3 : $N \times$ TransformerのDecoder部分
- ▶ $N=96$
- ▶ GPT-3パラメータ数 1, 750億, コーパス570GB以上
- ▶ クローリング45TB → 570GB
- ▶ InstructGPTの原理: Ouyang, L. et al. 2022

GPTの原理



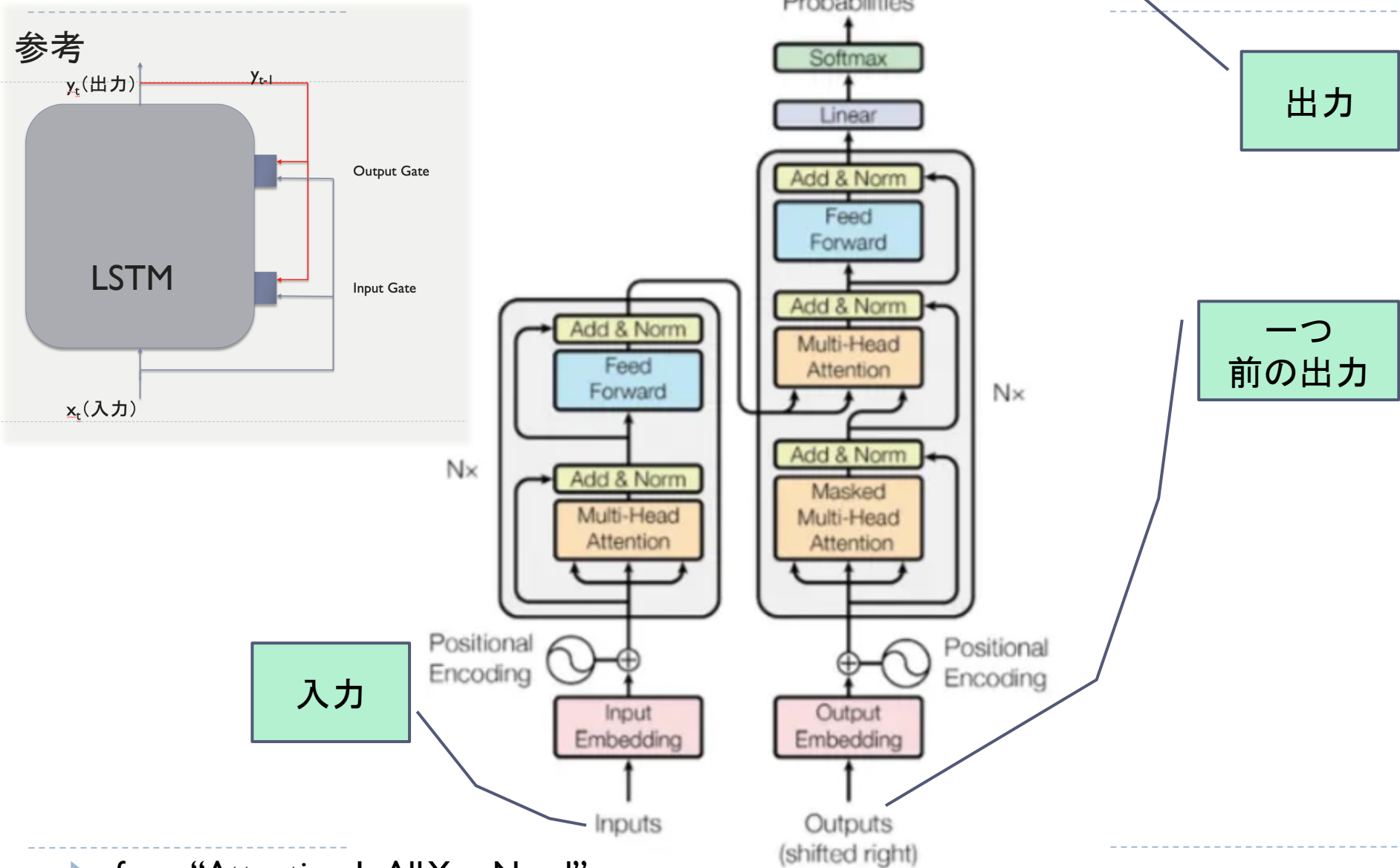
タスク

- next word prediction
- masked language model

Reward Model

* 数値はGPT-3, InstructionGPTから

transformer



▶ from “Attention Is All You Need”

GPT実例

(from <https://chat.openai.com> 2023/5/30)

TA

中国で使えるChatGPTと同等のシステムは何？



中国で利用可能なChatGPTと同等の人工知能システムとしては、以下のいくつかがあります。



1. 百度DuerOS（度秘）

百度DuerOS（度秘）は、中国の大手検索エンジンである百度が開発した人工知能プラットフォームです。DuerOSは音声やテキストに対応しており、自然な対話を行うことができます。音声アシスタントやスマートスピーカーなどのデバイスで利用されています。

2. 腾讯QQ机器人（QQ Chatbot）

腾讯（テンセント）が提供するQQ Chatbotは、中国で広く使用されているチャットプラットフォームQQの中に組み込まれた人工知能チャットボットです。ユーザーとの対話に応じて情報を提供したり、質問に答えたりすることができます。

3. 小i机器人（小i助手）

小i机器人（小i助手）は中国の企業、科大讯飞が開発したAIアシスタントです。テキストチャットや音声対話を通じて、ユーザーの質問に答えたり、タスクを実行したりすることができます。小i机器人は様々なインターネットサービスやアプリケーションに統合されています。

これらのシステムは中国語での対話に特化しており、一般的な会話や質問に対応しています。ただし、完全に人間のような対話を行うことはまだ困難な場合もあります。それぞれのシステムの能力や利用方法については、公式ウェブサイトや関連情報を参照してください。

GPT实例

(from <https://chat.openai.com> 2023/5/30)



これを中国語で教えてください



中国现有可与ChatGPT相当的人工智能系统包括：



1. 百度度秘 (DuerOS)

百度度秘 (DuerOS) 是中国知名搜索引擎百度开发的人工智能平台。DuerOS支持语音和文本交互，能够进行自然对话。它被应用于语音助手和智能音箱等设备。

2. 腾讯QQ机器人 (QQ Chatbot)

腾讯QQ机器人是中国广泛使用的聊天平台QQ内的一个人工智能聊天机器人。它可以根据用户的对话提供信息或回答问题。

3. 小i机器人 (小i助手)

小i机器人是中国科大讯飞开发的人工智能助手。它通过文本聊天或语音对话，回答用户的问题或执行任务。小i机器人被集成到各种互联网服务和应用中。

这些系统专注于中文对话，并且能够应对常见的会话和问题。然而，要实现与人类完全一样的对话仍然是具有挑战性的。有关每个系统的能力和使用方法，请参阅官方网站或相关信息。



GPT問題点

- ▶ 応答文が極めて自然
 - ▶ 応答が自然なので、信じてしまう。
 - ▶ 内容が古い場合がある。LLMを作った時点の情報に基づく応答。Web情報には事実でないものも存在する。
- ▶ プログラムを出力するが、数値、図、画像の直接の表示はしない。それぞれを生成するプログラムを出力する。
 - ✓ ソースコードを吐き出すが、そのままではほぼ動かない。
 - ✓ 修正にはそれなりの知識、技量を要する。
 - ✓ 手順がそっくり抜けてたりする。
- ▶ 目的「システム構築のために、ChatGPTは役立つか」
 - ▶ 検索システムの代替として使える(?)ただし、誤った解答があることを常に意識する必要がある。
 - ▶ 対話処理インタフェース(API)として期待

まとめ

- ▶ 単語のベクトル表現
- ▶ word2vec
- ▶ 文の扱い
 - ▶ CNN
 - ▶ RNN
 - ▶ LSTM
- ▶ ニューラル機械翻訳
- ▶ BERT
 - ▶ Pre-Training
 - ▶ Fine Tuning
- ▶ GPT
 - ▶ 概要
 - ▶ 原理
 - ▶ GPT実例
 - ▶ 問題点

課題

- ▶ GPT同等品式（百度DuerOS、騰訊QQ机器人、小i机器人）について調べ、（出来れば）動作を確認せよ。

