



nlpaper.challenge NLP/CV交流勉強会

[2019.1.19]

筑波大学大学院 システム情報工学研究科
知能機能システム専攻 自然言語処理研究室

飯田 頌平





自己紹介 - s.iida

- 所属：筑波大学大学院 自然言語処理研究室 M1
ーニューラル機械翻訳(NMT)を研究



- 論文、発表：
 - ー大規模名詞句フレーズのSMT訳への後処理型置き換えによるNMTの評価・・・NLP2018
 - http://www.anlp.jp/proceedings/annual_meeting/2018/pdf_dir/P6-6.pdf
 - ーフレーズ・トークン込み NMT モデルおよび SMT による大語彙フレーズ翻訳によるハイブリッド翻訳方式・・・電子情報通信学会論文誌 D
 - https://search.ieice.org/bin/summary_advpub.php?id=2018PDP0003&category=D&lang=J
 - ーその他、Japio YEAR BOOK 2018等に寄稿



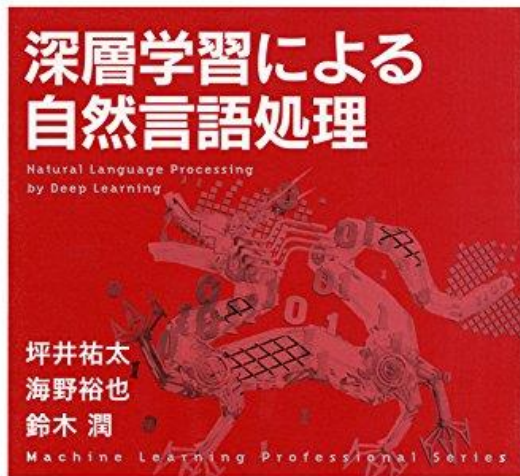
本日のおしながき

1. 自己紹介

2. 『出た!!本』 第1章 自然言語処理のアプローチ

1. 伝統的な自然言語処理
2. 深層学習への期待
3. テキストデータの特徴
4. 他分野への広がり

3. 質疑応答など



出た!!

- 「実装上の工夫」など、本書でしか読めない実践的な内容!
- 自然言語処理の応用(機械翻訳、文書要約、対話、質問応答)に焦点を当て、深層学習の利用方法を解説。

MLP 機械学習
プロフェッショナル
シリーズ

講談社



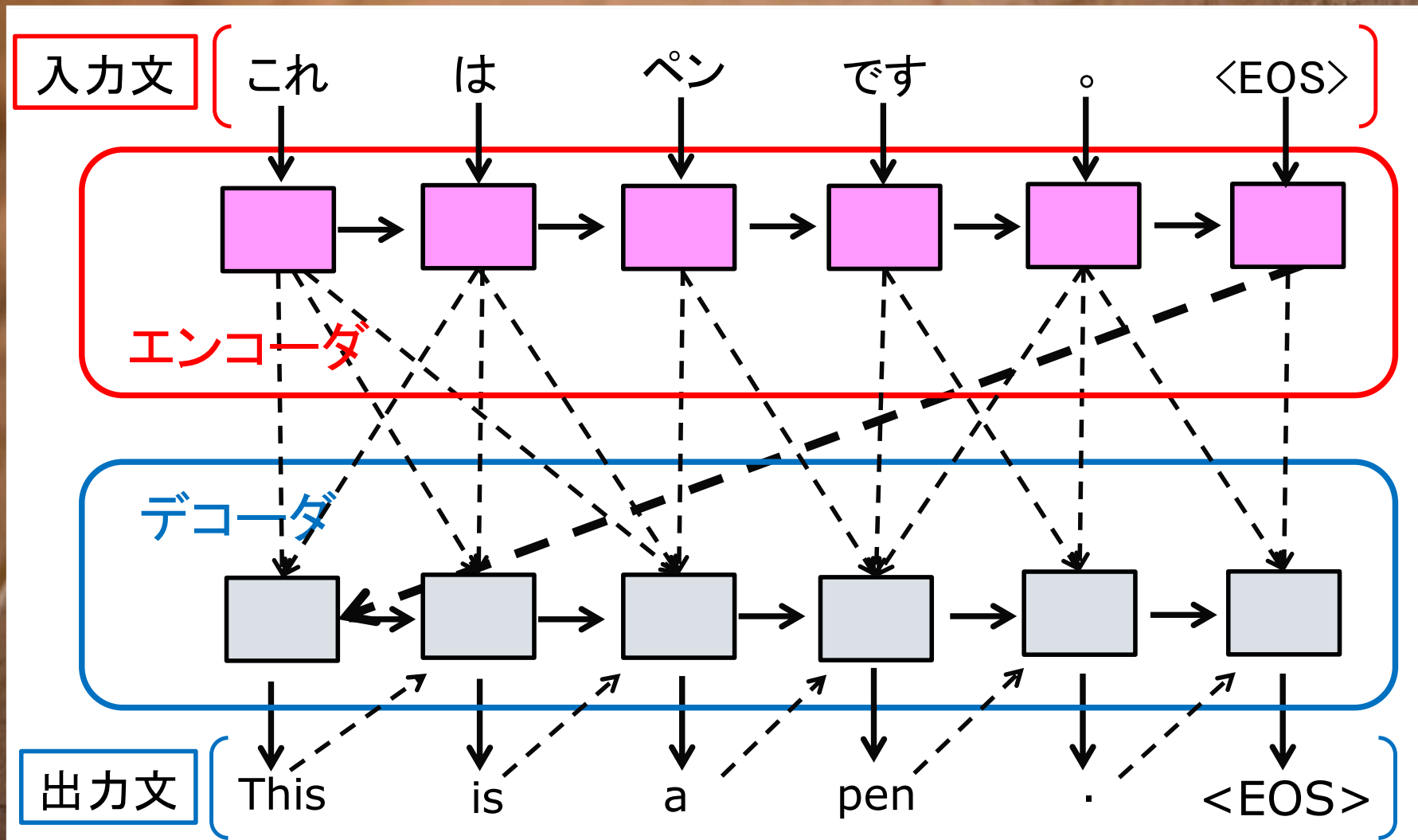


1.1 伝統的な自然言語処理

- 自然言語処理の応用タスク
 - 機械翻訳
 - 文書分類
- 部分問題としての言語解析タスク
 - 機械翻訳における部分問題



再帰型ニューラルネットワーク (RNN)による sequence to sequence 機械翻訳モデル



例：文書分類

A four star performance . . . script

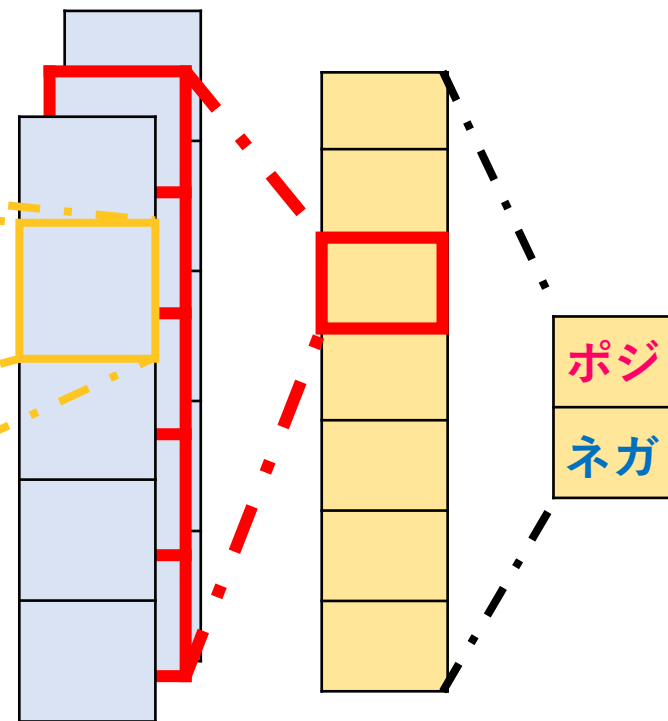
文

分散表現

A four star performance . . . script

0.1	0.2	- 0.1	0.6	0.4
- 0.5	- 0.5	- 0.6	- 0.1	0.3
0.3	0.4	- 0.1	0.2	- 0.3
⋮	⋮	⋮	⋮	⋮
- 0.3	- 0.7	0.1	- 0.1	- 0.2

CNNで文の感情
(Positive / Negative)
を分類するタスク



出展：筑波大学大学院 自然言語処理研究室



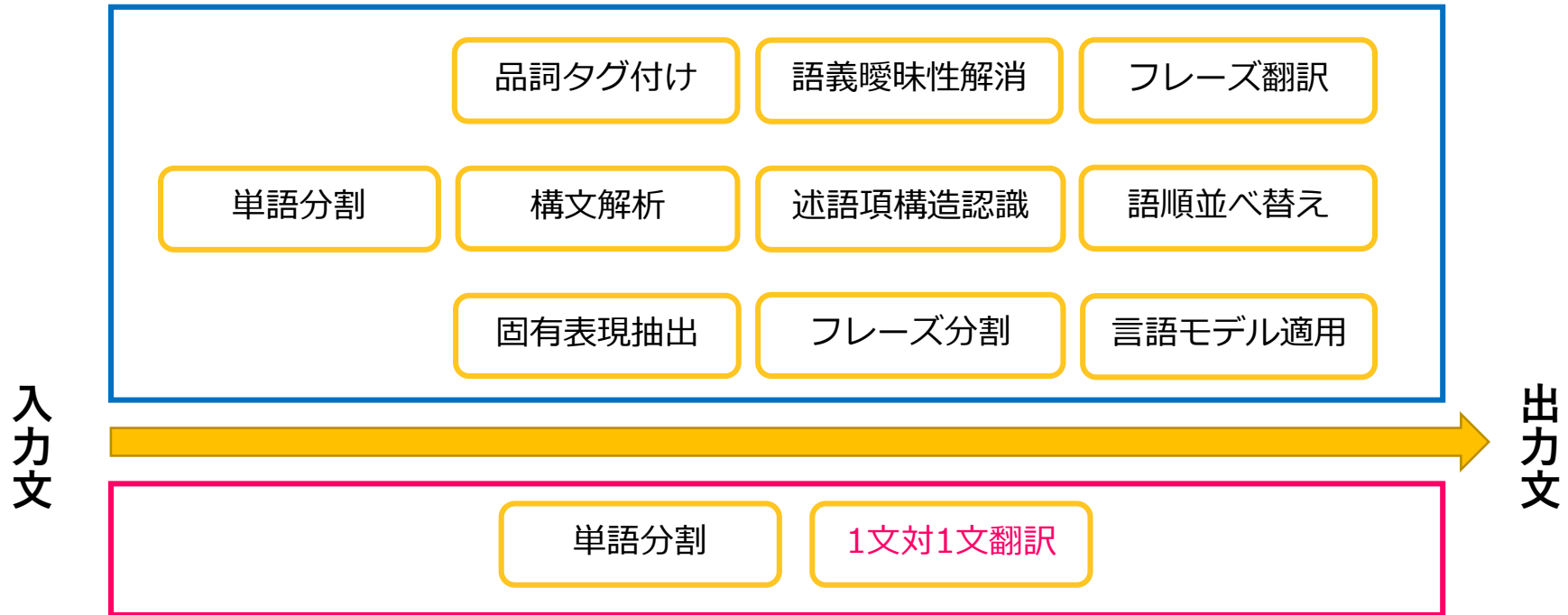
1.1 伝統的な自然言語処理

- 自然言語処理の応用タスク
 - 機械翻訳
 - 文書分類
- 部分問題としての言語解析タスク
 - 機械翻訳における部分問題



機械翻訳における部分問題

伝統的手法・・・統計的機械翻訳(SMT)の部分問題



深層学習を用いた手法・・・ニューラル機械翻訳(NMT)の部分問題



例：SMTの部分問題

入力文：これは必ずしも必要はないがシャフト11を予備とする

(1) フレーズ分割

分割パターン1

これは	必ずしも	必要	は	ない
-----	------	----	---	----

分割パターン2

これは	必ずしも	必要	は	ない
-----	------	----	---	----

分割パターン3

これは	必ずしも	必要	は	ない
-----	------	----	---	----

(2) フレーズ翻訳

これは	必ずしも	必要	は	ない
-----	------	----	---	----

this is not necessarily	a
-------------------------	---

(3) 語順並べ替え確率モデル適用

(4) 英語言語モデル適用

フレーズ翻訳テーブル

これは 必ずしも 必要 ⇒ this is not necessarily	0.33
は ない ⇒ is not	0.23
は ない ⇒ are not	0.04
は ない ⇒ a	0.002
これは ⇒ this is	0.28
必ずしも 必要 は ない ⇒ not always necessary	0.5
必ずしも 必要 は ない ⇒ not necessarily required for	0.5
...	...

日英翻訳モデル

語順並べ替え確率モデル

英語言語モデル

(英文中の単語の並びの確率モデル)

これは	は	必ずしも	必要	は	ない
-----	---	------	----	---	----

this is	not necessarily required for
---------	------------------------------

this is	not always necessary
---------	----------------------

語順＝逆順

語順＝正順

not always necessary this is

this is not always necessary

$P(\text{not}|\text{文頭})$
 $\times P(\text{always}|\text{not}, \text{文頭})$
 $\times P(\text{necessary}|\text{not always})$
 $\times P(\text{this}|\text{always necessary})$
 $\times P(\text{is}|\text{necessary this}) \times \dots$

$P(\text{this}|\text{文頭})$
 $\times P(\text{is}|\text{this}, \text{文頭})$
 $\times P(\text{not}|\text{this is})$
 $\times P(\text{always}|\text{is not})$
 $\times P(\text{necessary}|\text{not always}) \times \dots$

総確率値算出

総確率値算出



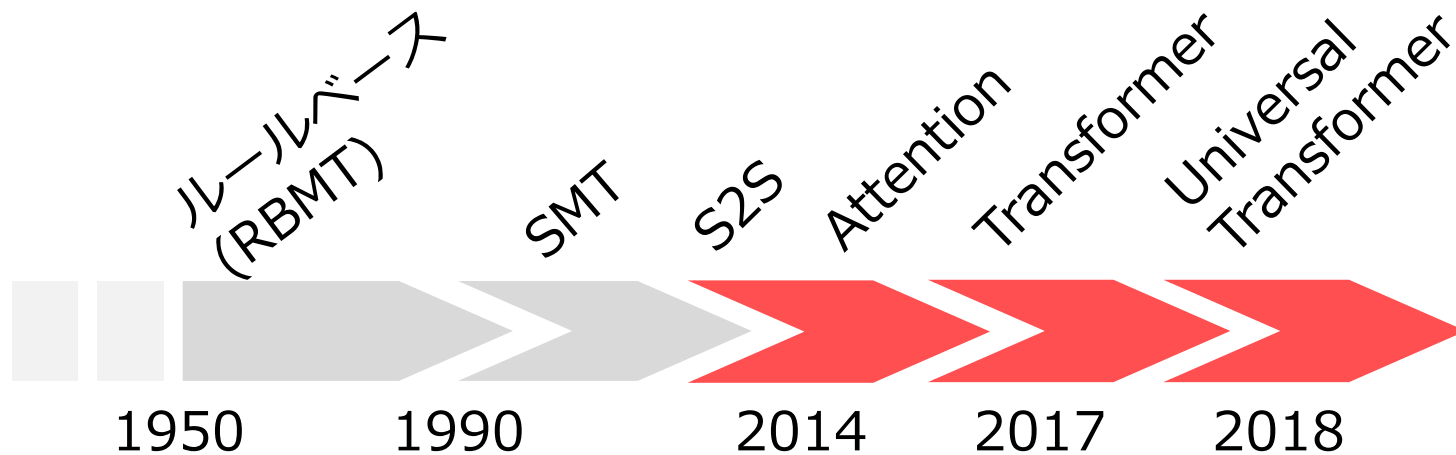
1.2 深層学習への期待

- 深層学習による自然言語処理システム
 - NMTの発展
 - ニューラル系列変換モデル
- ブラックボックス化に対する取り組み
 - 注意機構における重要度の可視化
 - 転移学習による分析



NMTの発展(1)

- 深層学習による自然言語処理システム
 - S2S以降、常にニューラル機械翻訳(NMT)が開拓してきた





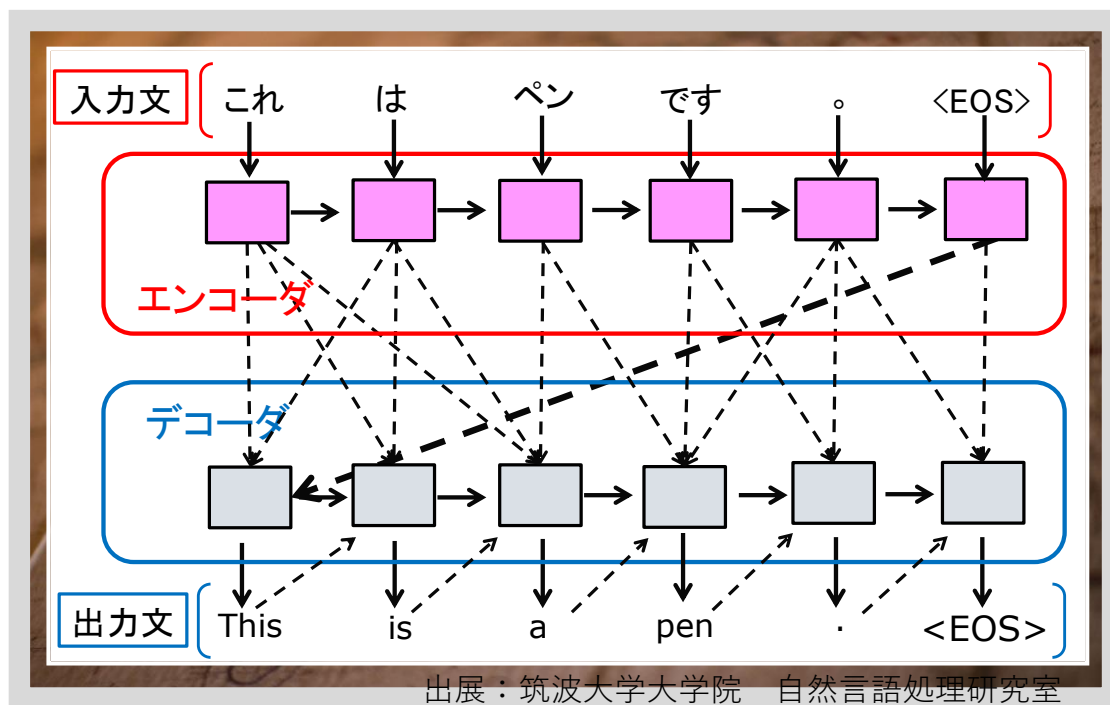
NMTの発展(2)

- 文の生成に長けたEncoder-Decoderモデルの登場
 - sequence-to-sequence(S2S)[Sutskever et al., 2014]
- 注意機構によるS2Sの改良
 - MLP attention[Bahdanau et al., 2015]
 - dot product attention[Luong et al., 2015]
- 双方向RNNをS2Sに採用
 - BRNN[Schuster et al., 1997]
 - BRNN S2S[See et al., 2016]
- RNNからの脱却
 - Transformer[Vaswani et al., 2017]
 - ConvS2S[Gehring et al., 2017]
 - Universal Transformer[Dehghani et al., 2018]



sequence-to-sequenceモデル

- エンコーダとデコーダにそれぞれRNN(LSTM)を使用
 - エンコーダが入力文の全ステップの情報を符号化
 - 符号化された中間表現からデコーダの各ステップで単語生成



$$h_i^{(s)} = \Psi^{(s)}(\bar{x}_i, h_{i-1}^{(s)})$$

Ψ : LSTM, s : 層, i : ステップ,
 $\bar{x}$: 入力文, h : 内部状態

$$h_j^{(t)} = \Psi^{(t)}(\bar{y}_j, h_{j-1}^{(t)})$$

Ψ : LSTM, t : 層, j : ステップ,
 $\bar{y}$: 入力文, h : 内部状態





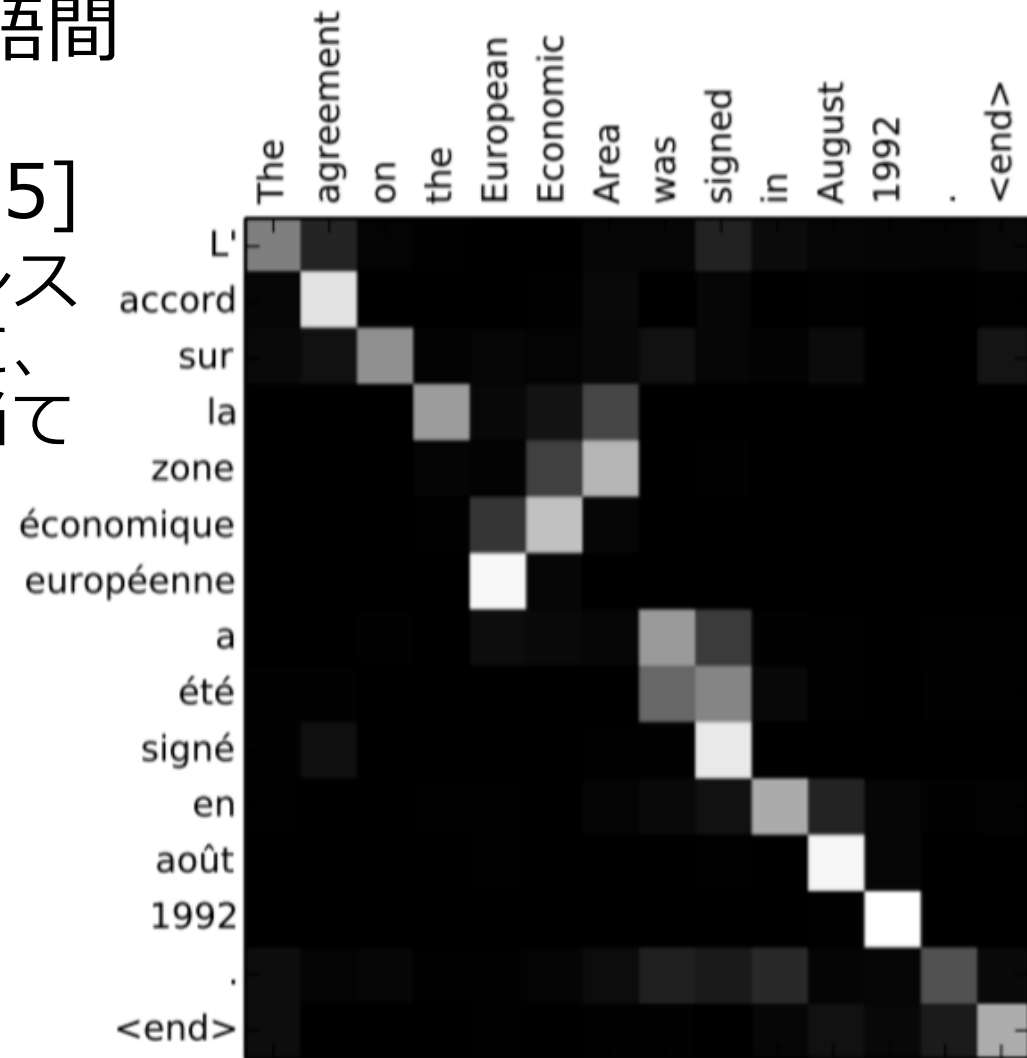
1.2 深層学習への期待

- 深層学習による自然言語処理システム
 - NMTの発展
 - ニューラル系列変換モデル
- ブラックボックス化に対する取り組み
 - 注意機構における重要度の可視化
 - 転移学習による分析



注意機構における重要度の可視化

- NMTの原言語・目的言語間での重要度を可視化
[Bahdanau et al., 2015]
– 英仏翻訳においてフランス語の単語を生成する際に、重要度がどの英単語に当てられているかを可視化

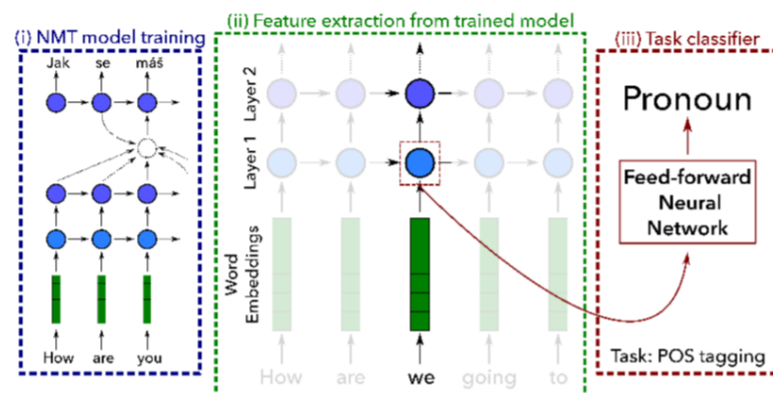




転移学習による分析(1)

- RNNの層ごとの傾向の調査[Belinkov et al., 2017]

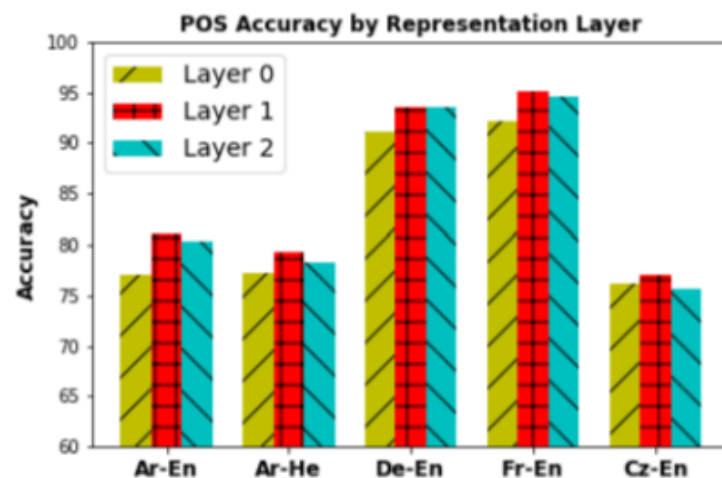
- まず機械翻訳タスクでS2Sモデルを訓練
- その後、訓練したRNNの層を品詞タグ付けタスクに転用



- 結論

- RNNの1層目(Layer1)は2層目よりも正答率が高い
- 品詞のような文法的な要素に必要な特徴は低次の層で得られている

NMTでは、翻訳という応用問題の中に、品詞タグ付けという部分問題が含まれる

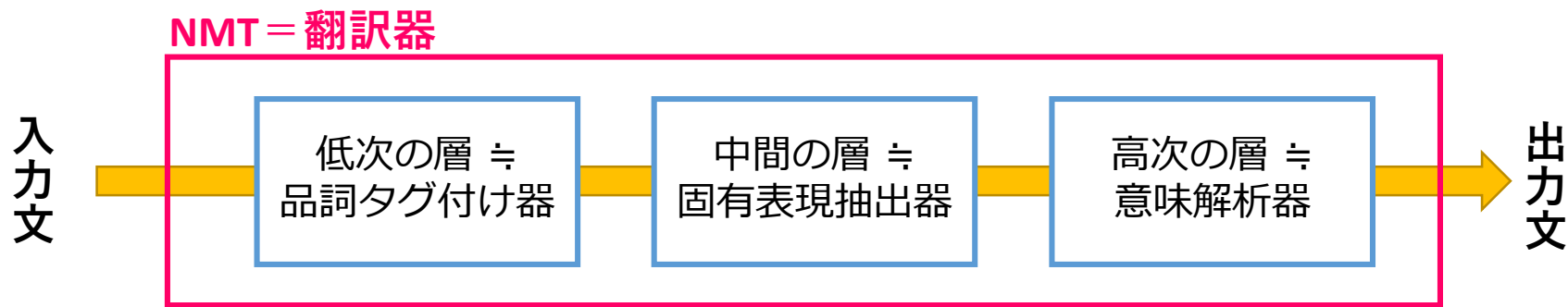




転移学習による分析(2)

- Transformerの層ごとの傾向の調査[Raganato et al., 2018]
 - 機械翻訳のタスクでTransformerモデルを訓練
 - その後、訓練したself-attentionの層を4タスクに転用

タスク	タスクの説明	得意な層の傾向
品詞タグ付け	単語の品詞を分類問題として推定	低次の層
チャンキング	文節の切れ目を推定	低次の層
固有表現抽出	名詞の固有表現（人名, 地名, …）を推定	中間の層
語義タグ付け	単語の意味を分類問題として推定	高次の層





1.3 テキストデータの特徴

- NLPでよく使われるデータセット
 - 機械翻訳：対訳コーパス
 - 質問応答：SQuAD
- 前処理：単語分割
 - 日本語の単語分割手法
 - 英語の単語分割手法
 - 未知語問題





対訳コーパスの例

- OpenSubtitles2018[Lison et al., 2018]
 - 映画字幕の対訳コーパス
 - 原言語と目的言語の文が一文対一文で対応している

(src)="1"> " THE CABINET OF DR. CALIGARI "
(trg)="1"> カリガリ 博士 の 小屋

(src)="2"> Act 1
(trg)="2"> 幕 1

(src)="3"> " Spirits surround us on every side – they have driven me from hearth and home , from wife and child . "
(trg)="3"> 私 は 亡霊 の おかげ で 家族 や 家庭 を 捨て ざる を え な かった

(src)="4"> " She is my bride ... "
(trg)="4"> 僕 の 婚約 者 だ

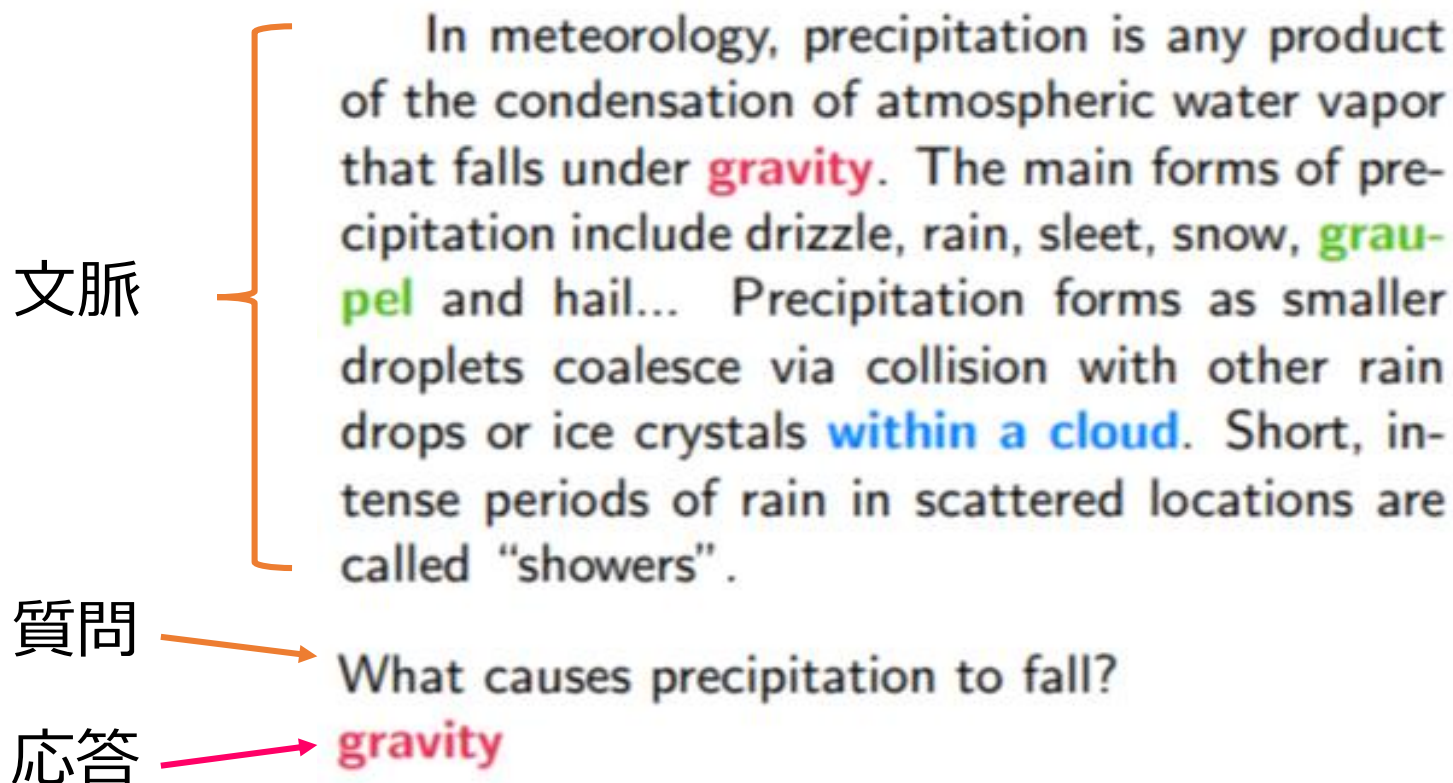
(src)="5"> " What she and I have experienced is yet more remarkable than the story you have told me . "
(trg)="5"> 僕 と 彼女 の 体験 は あなたの 物語 より も 恐ろしい





質問応答データセットの例

- SQuAD[Rajpurkar et al., 2016]
 - 質問文、応答文、そして応答のヒントになる文脈のセット





1.3 テキストデータの特徴

- NLPでよく使われるデータセット
 - 機械翻訳：対訳コーパス
 - 質問応答：SQuAD
- 前処理：単語分割
 - 日本語の単語分割手法
 - 英語の単語分割手法
 - 未知語問題



日本語の単語分割手法

- MeCab[Kudo, 2004]
 - 日本語の形態素解析器（実は言語非依存）
 - 単語辞書によって単語分割を行う
 - MeCab + ipadic + neologdによる分割結果の例：

本山は、足利義満により建立された京都の相国寺。

MeCab



本山 / は / 、 / 足利義満 / により / 建立 / さ / れ / た / 京都 / の / 相国寺 / 。

辞書に基づき、「足利義満」のような固有名詞が一単語として認識される





英語の単語分割手法

- Moses Tokenizer[Koehn et al., 2007]
 - 英語における単語分割器
 - カンマやピリオドを処理するためのもの
 - Moses Tokenizerによる分割例：

Head temple is Shokoku-ji in Kyoto, established by Yoshimitsu ASHIKAGA.

Moses Tokenizer



Head temple is Shokoku-ji in Kyoto , established by Yoshimitsu ASHIKAGA .

["Kyoto,"]で1単語となっていたものを、["Kyoto", ","]の2単語に分割
["ASHIKAGA."]で1単語となっていたものを、["ASHIKAGA", "."]の2単語に分割



未知語問題

- MeCabなど単語単位で分割する手法では、単語の種類の数だけ語彙が際限なく増え、分散表現で表せない
- よって、語彙の数にしきい値を設け、出現頻度の低い単語は未知語として無視する必要がある
- 無視された単語はシステムで扱えない（未知語問題）





未知語問題への対策

- Subword[Wu et al., 2016]

- 未知語を発生させないよう、単語未満の文字単位で分割
- 統計的に頻度の高い文字の並びをまとめる
- Subword(SentencePiece[Kudo et al., 2018])による分割例：

本山は、足利義満により建立された京都の相国寺。

SentencePiece



本 / 山 / は / 、 / 足利義 / 満 / により / 建立 / された / 京都 / の / 相 / 国 / 寺 / 。

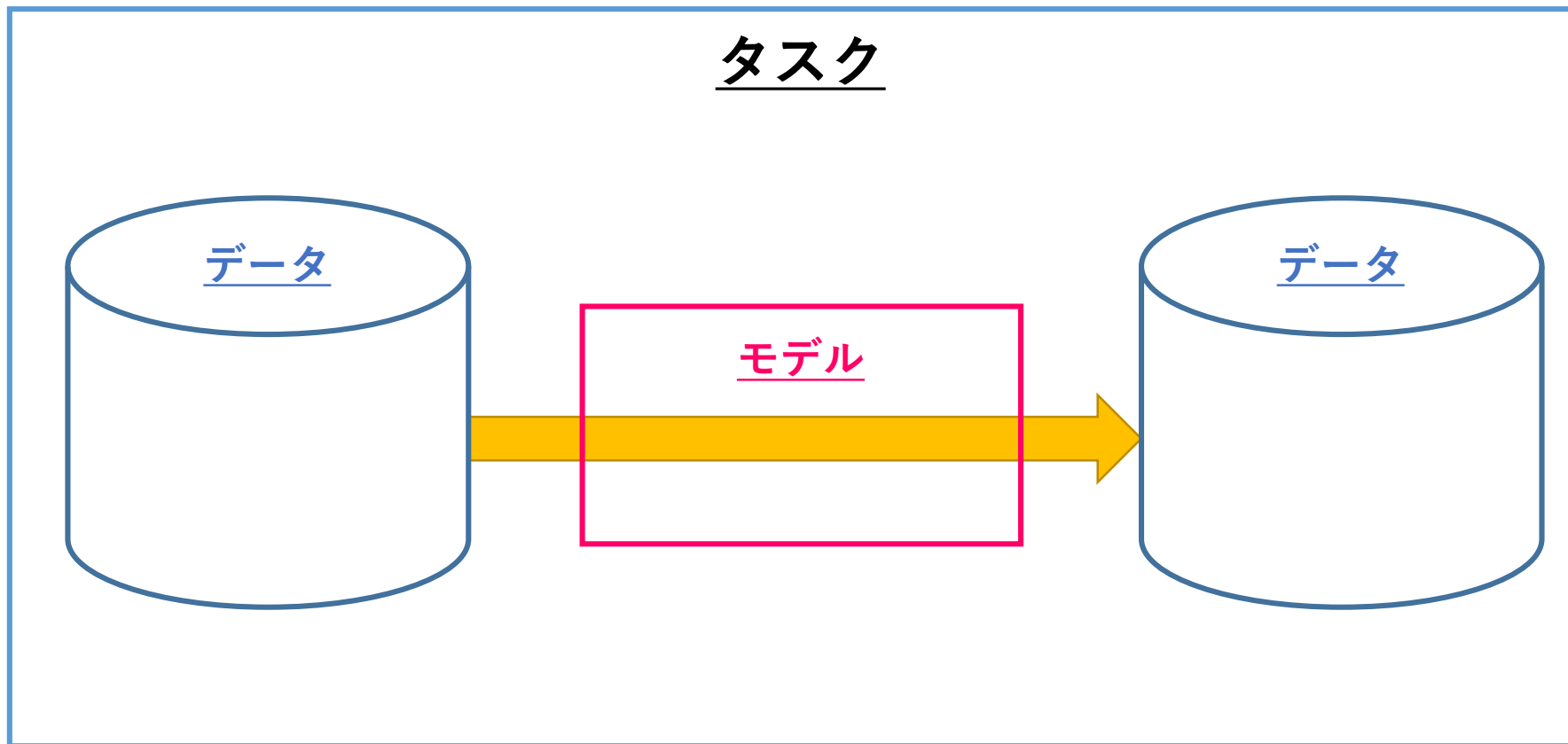


「足利義政」など、「足利義」・・・という文字の並びが多いため、
「足利義」でひとつの単語（トークン）で認識される
「視」という文字と組み合わせれば「足利義視」も語彙に入る



NLPのまとめ

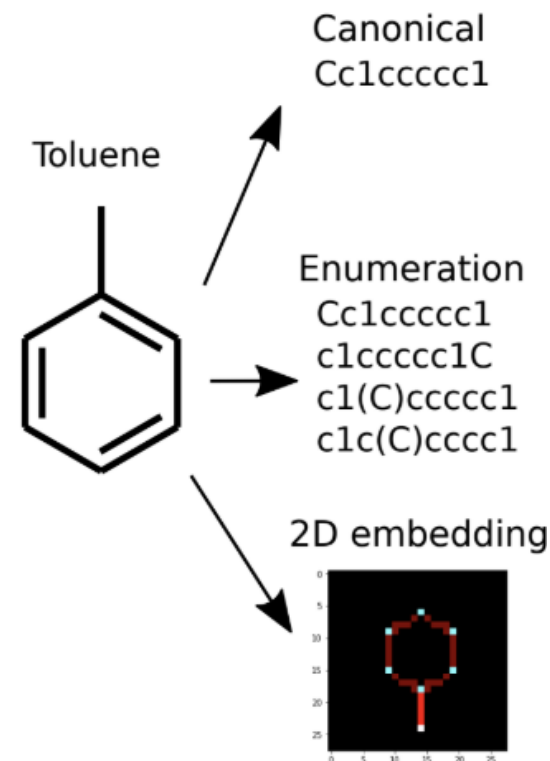
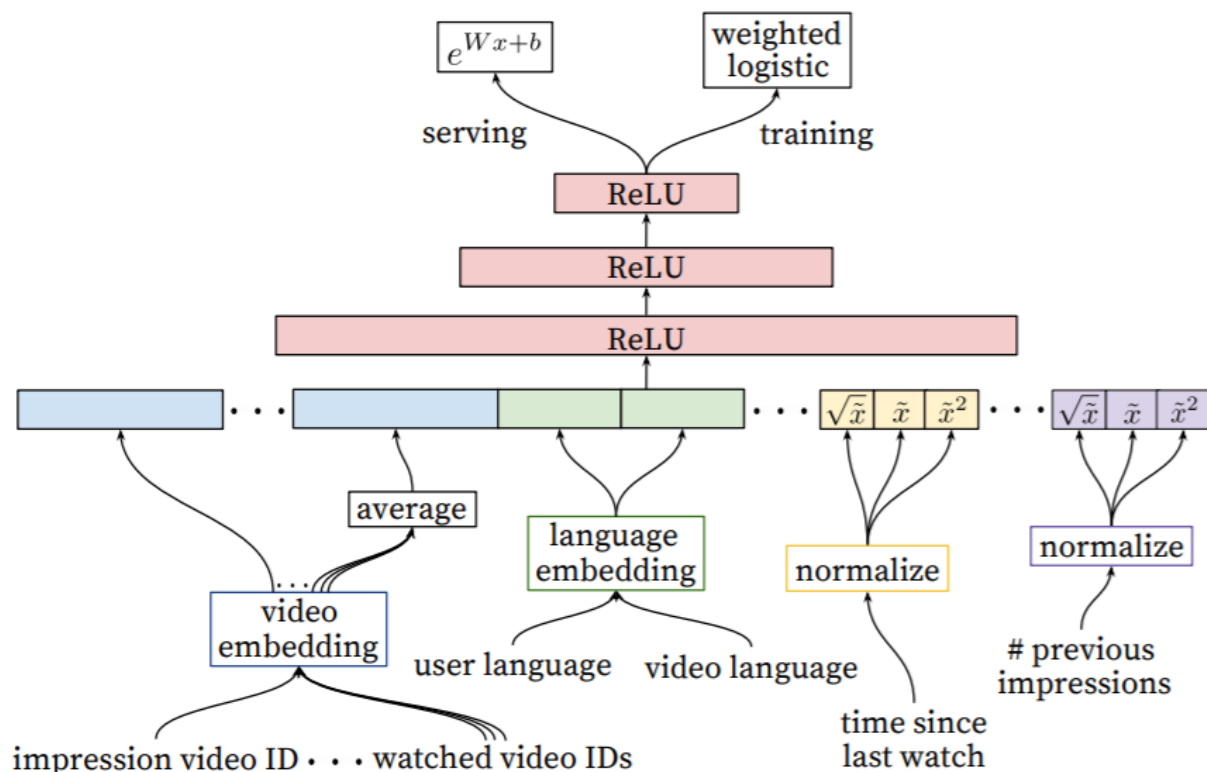
- タスク, モデル, データの3要素でNLPの研究が始まる



1.4 他分野への広がり

• 非言語のデータへの適用

- Youtubeの推薦システム[Covington et al., 2016]
- 分子構造を文字列として扱う[Bjerrum et al., 2018]





参考文献(1)

- Neural Machine Translation by Jointly Learning to Align and Translate[Bahdanau et al., 2015]
- What do Neural Machine Translation Models Learn about Morphology?[Belinkov et al., 2017]
- Improving Chemical Autoencoder Latent Space and Molecular De-novo Generation Diversity with Heteroencoders[Bjerrum et al., 2018]
- Deep Neural Networks for YouTube Recommendations [Covington et al., 2016]
- Universal Transformers[Dehghani et al., 2018]
- Convolutional Sequence to Sequence Learning[Gehring et al., 2017]
- Moses: Open Source Toolkit for Statistical Machine Translation[Koehn et al., 2007]
- Applying Conditional Random Fields to Japanese Morphological Analysis[Kudo, 2004]





参考文献(2)

- OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora[Lison et al., 2018]
- Effective Approaches to Attention-based Neural Machine Translation[Luong et al., 2015]
- An Analysis of Encoder Representations in Transformer-Based Machine Translation[Raganato et al., 2018]
- SQuAD: 100,000+ Questions for Machine Comprehension of Text[Rajpurkar et al., 2016]
- Bidirectional Recurrent Neural Networks[Schuster et al., 1997]
- Compression of Neural Machine Translation Models via Pruning[See et al., 2016]
- Sequence to Sequence Learning with Neural Networks[Sutskever et al., 2014]
- Attention is All You Need[Vaswani et al., 2017]
- Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[Wu et al., 2016]





Special Thanks

- 筑波大学大学院 自然言語処理研究室（宇津呂研究室）
 - －研究室資料：ニューラル機械翻訳(p5, p13)
 - <http://nlp.iit.tsukuba.ac.jp/research/listxx-NN-NMT.html>
 - －研究室資料：文書分類(p6)
 - <http://nlp.iit.tsukuba.ac.jp/research/listxx-NN-sentiment.html>
 - －研究室資料：統計的機械翻訳(p9)
 - "機械翻訳の研究の現状 --- 経験的アプローチを中心として ---"
宇津呂 武仁
日本語学, 第28巻, 第12号, pp.44-61, October 2009.
 - <https://ci.nii.ac.jp/naid/40016839007>
- エンジニアサークル『デジらくだ』
 - －スライド素材
 - <https://digirakuda.org/>

