EMNLP2016参加報告

第3回自然言語処理シンポジウム

2016/12/21

東北大学 高瀬翔

EMNLP2016にNN関連の論文はどの程度あったのか？

増村さんのEMNLP2015参加報告と同じ解析をしてみる

NAACL2016では

71 / 182

NAACL-HLT 2016参加報告 [鈴木] より

EMNLP2016では

92 / 264

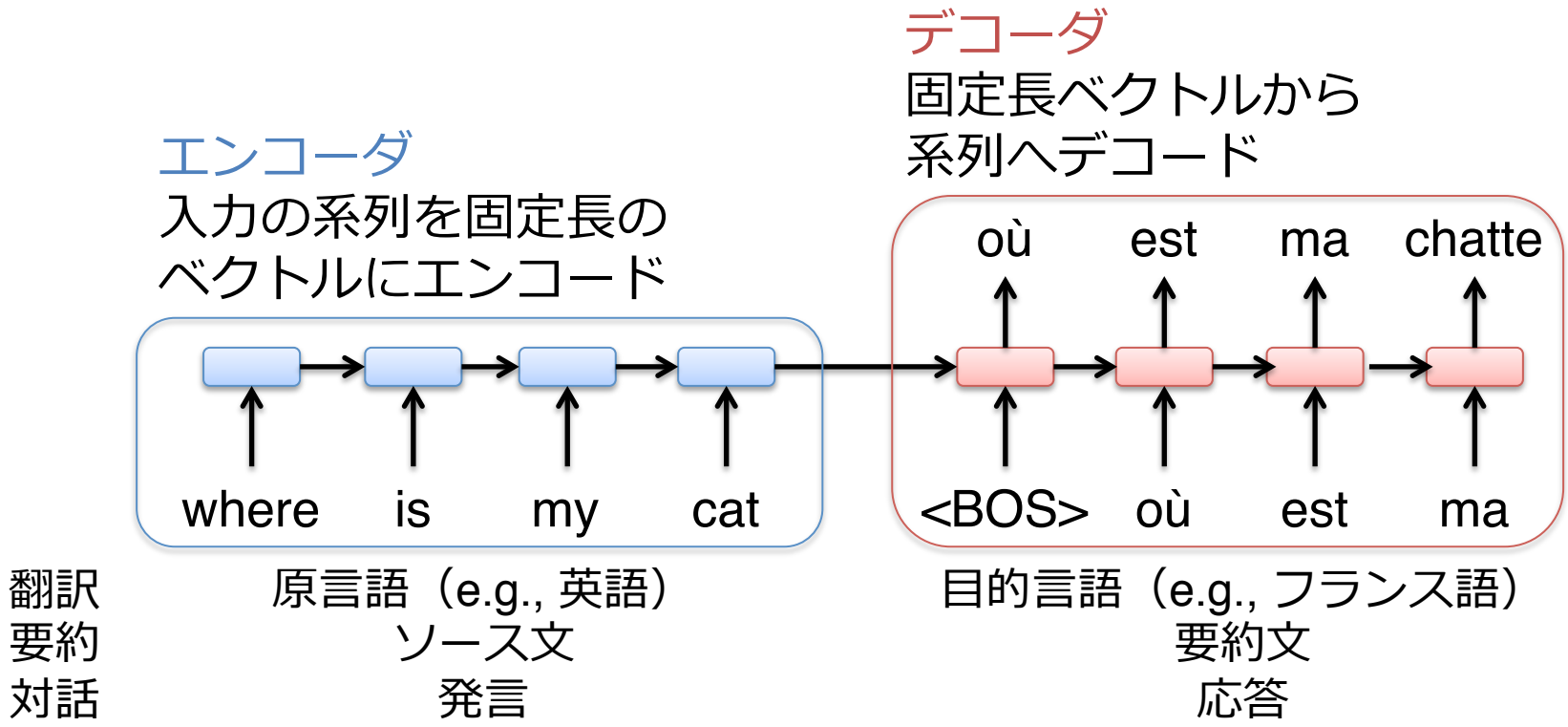
= 34.8%

基礎解析（統語解析など）から
応用（機械翻訳など）まで
様々なタスクでニューラルが活躍

ニューラルネットを用いた自然言語生成の研究を紹介

ニューラルネットを用いた自然言語生成

- Sequence-to-sequence（seq2seq） [Sutskever+ 14] が主流
  - 入力-出力対を変えることで様々なタスクに適用可能



EMNLP2016における seq2seqモデルの改良

デコーダの改良

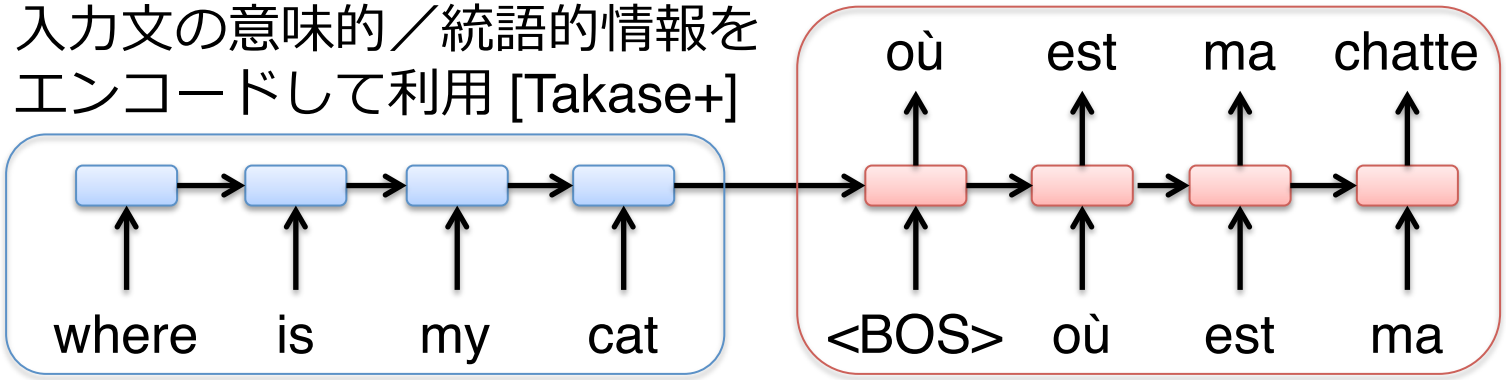
1 : • 系列の学習を真面目に [Wiseman+]

2 : • 外部メモリの利用 [Wang+]

• 出力長の制御 [Kikuchi+]

エンコーダの改良

• 入力文の意味的／統語的情報を エンコードして利用 [Takase+]



全体の改良

3 : • モデルの簡素化 [Kim+]

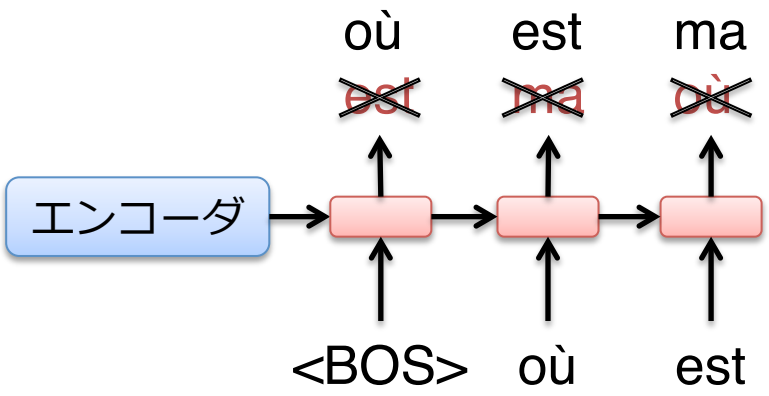
• アライメントを陽に行う [Yu+]

1 ：Sequence-to-Sequence Learning as Beam-Search Optimization [Wiseman+]

- Seq2seqの問題点：学習時とテスト時で状況が違う

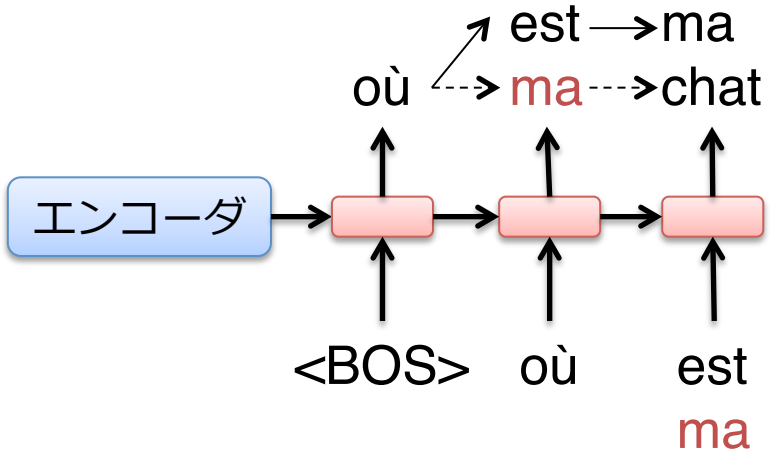
学習時

正しい単語を出力できるよう学習



テスト時

確からしい系列を出力



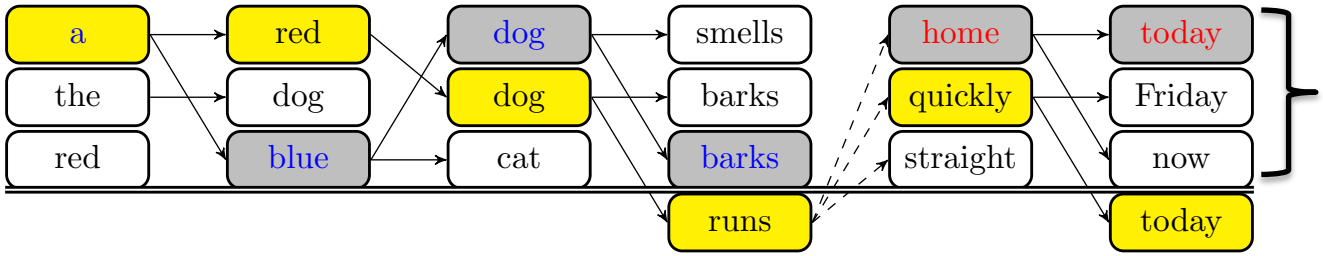
正しい単語から次の単語を予測
（前のステップでの誤りを考慮しない）

前のステップでの出力から次の単語を予測

問題の解決のために

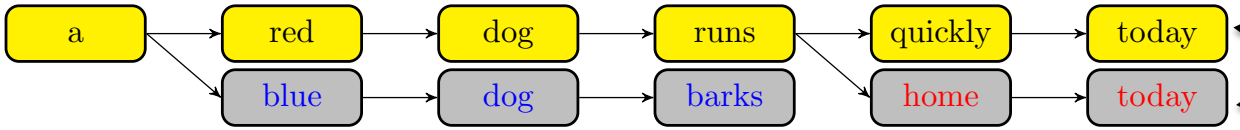
- ビームサーチで正解系列を出力できるよう学習
  - 確からしい系列を出力できるよう学習
  - ビーム内に正解があれば良い（モデルが頑健）

正解の単語列



学習時の
ビームサーチ結果
（ビーム幅：3）

正解の単語列がビームから落ちた箇所について
正解の単語列がビーム内に来るよう学習



ビーム内の単語列と
順位が入れ替わるよう
学習

実験結果

	Machine Translation (BLEU)		
テスト時のビーム幅	$K_{te} = 1$	$K_{te} = 5$	$K_{te} = 10$
seq2seq	22.53	24.03	23.87
BSO, SB-Δ	**23.83**	**26.36**	**25.48**

ビーム幅 = 6で学習した結果

- 独-英の翻訳タスクでseq2seqモデルより良い性能
- 単語並べ替え，依存構造解析でもseq2seqモデルより良い性能だった

EMNLP2016における seq2seqモデルの改良

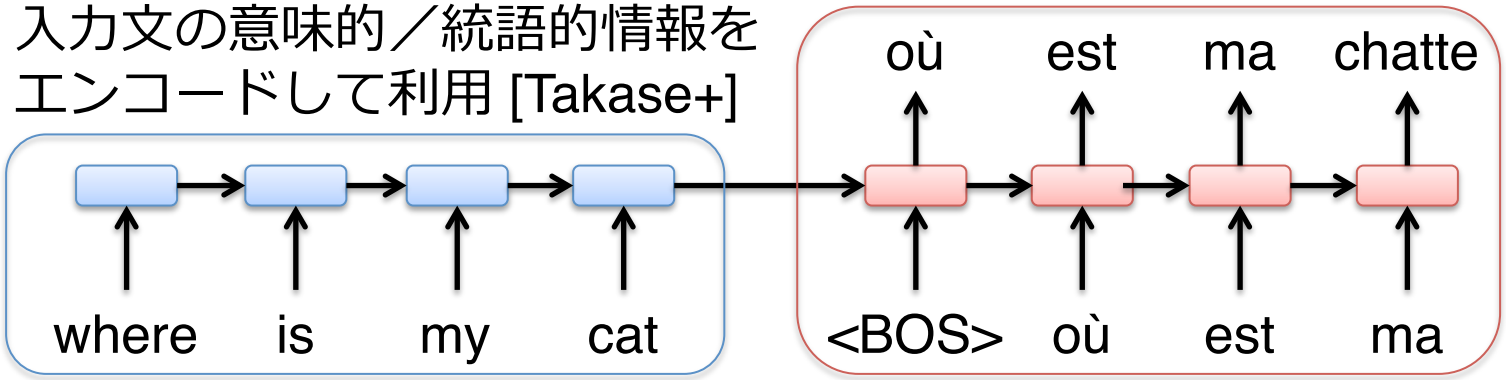
デコーダの改良

1 :
- 系列の学習を真面目に [Wiseman+]

2 :
- 外部メモリの利用 [Wang+]
- 出力長の制御 [Kikuchi+]

エンコーダの改良

- 入力文の意味的／統語的情報をエンコードして利用 [Takase+]



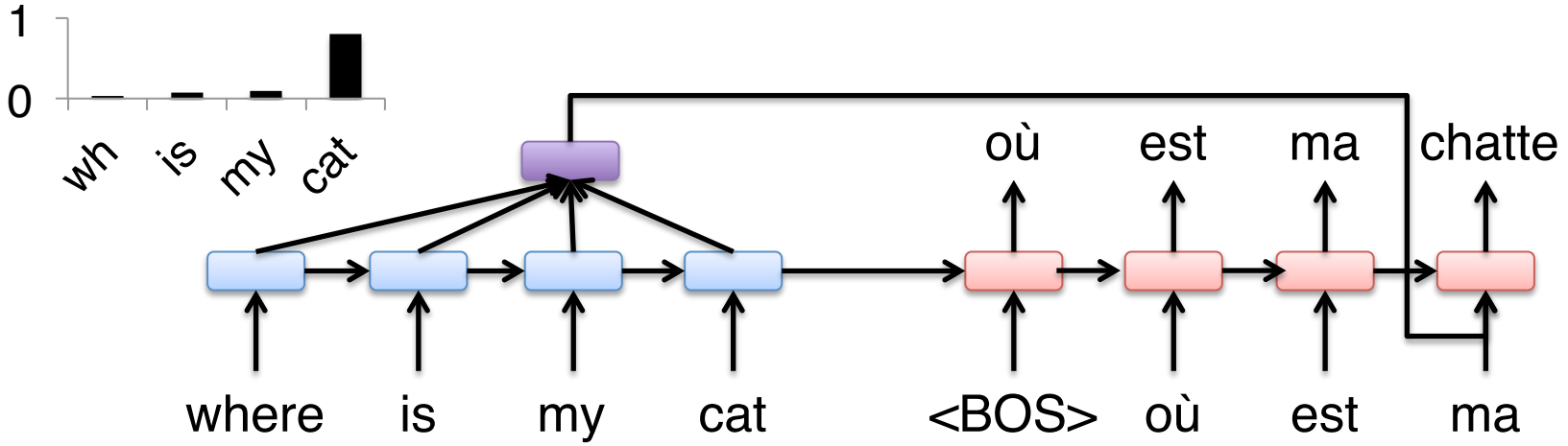
全体の改良

3 :
- モデルの簡素化 [Kim+]
- アライメントを陽に行う [Yu+]

2 ：Memory-enhanced Decoder for Neural Machine Translation [Wang+]

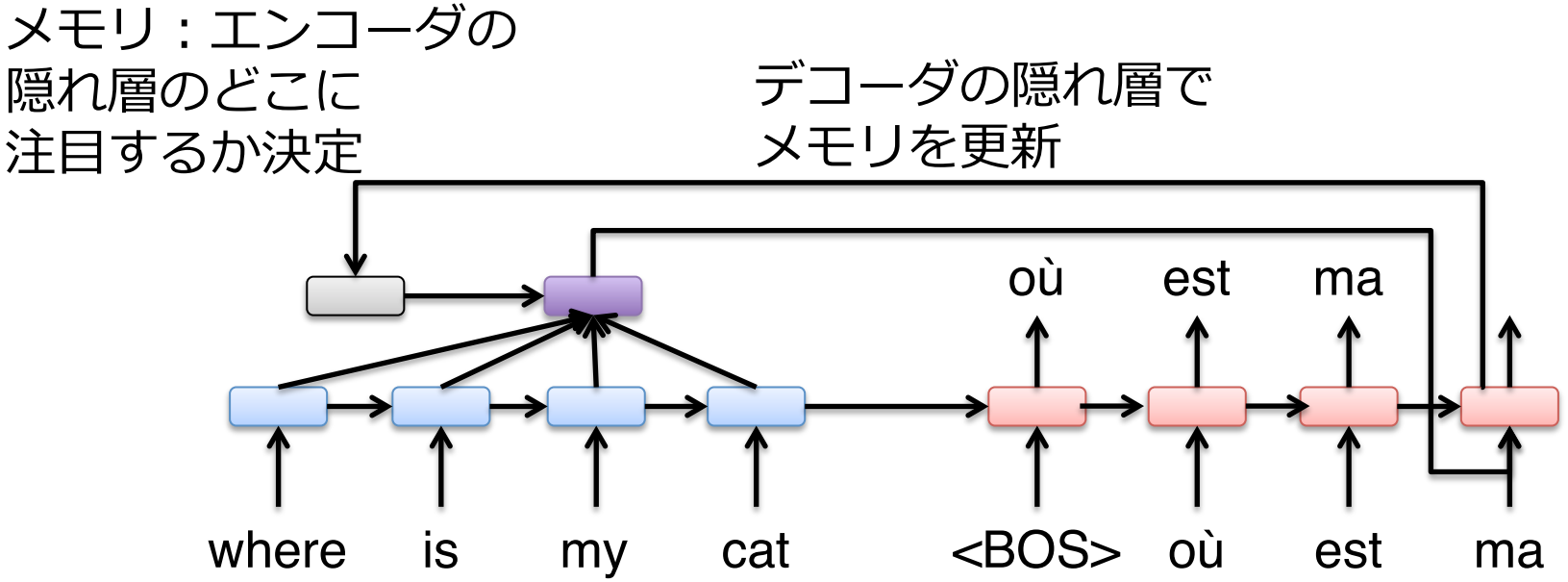
- メモリを利用してデコードする情報を選択
  - メモリ：アテンションの拡張

アテンション：エンコーダの隠れ層の
重み付き和をデコードの各ステップで利用



2 ：Memory-enhanced Decoder for Neural Machine Translation [Wang+]

- メモリを利用してデコードする情報を選択
  - メモリ：アテンションの拡張



実験結果

SYSTEM	MT03	MT04	MT05	MT06	AVE.
Groundhog	31.92	34.09	31.56	31.12	32.17
RNNsearch⋆	33.11	37.11	33.04	32.99	34.06
RNNsearch⋆ + coverage	34.49	38.34	34.91	34.25	35.49
MEMDEC	**36.16**	**39.81**	**35.91**	**35.98**	**36.95**
Moses	31.61	33.48	30.75	30.85	31.67

RNNsearch: アテンション付きseq2seq

- 中-英の翻訳タスクでBLEUが向上

EMNLP2016における seq2seqモデルの改良

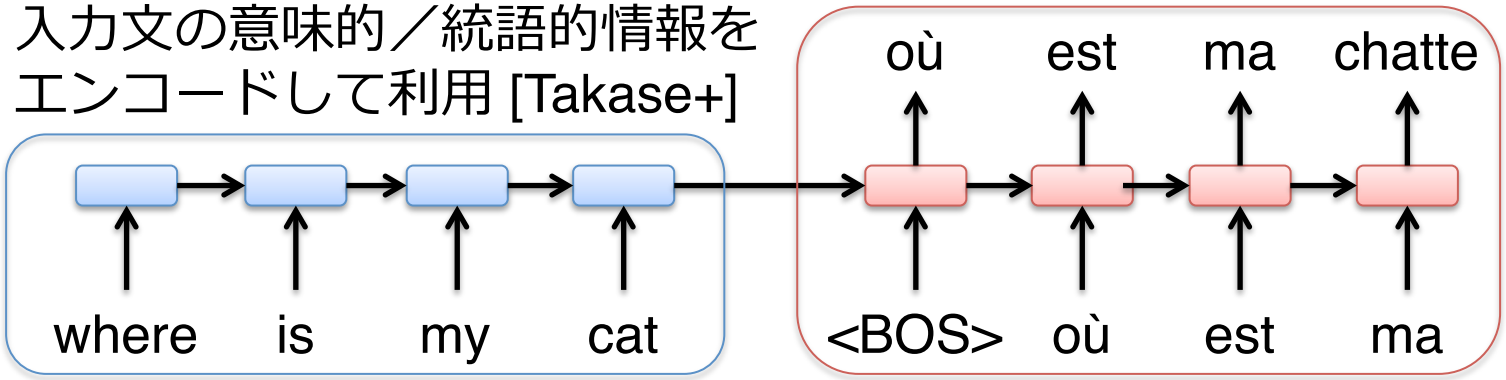
デコーダの改良

1 :
- 系列の学習を真面目に [Wiseman+]

2 :
- 外部メモリの利用 [Wang+]
- 出力長の制御 [Kikuchi+]

エンコーダの改良

- 入力文の意味的／統語的情報をエンコードして利用 [Takase+]



全体の改良

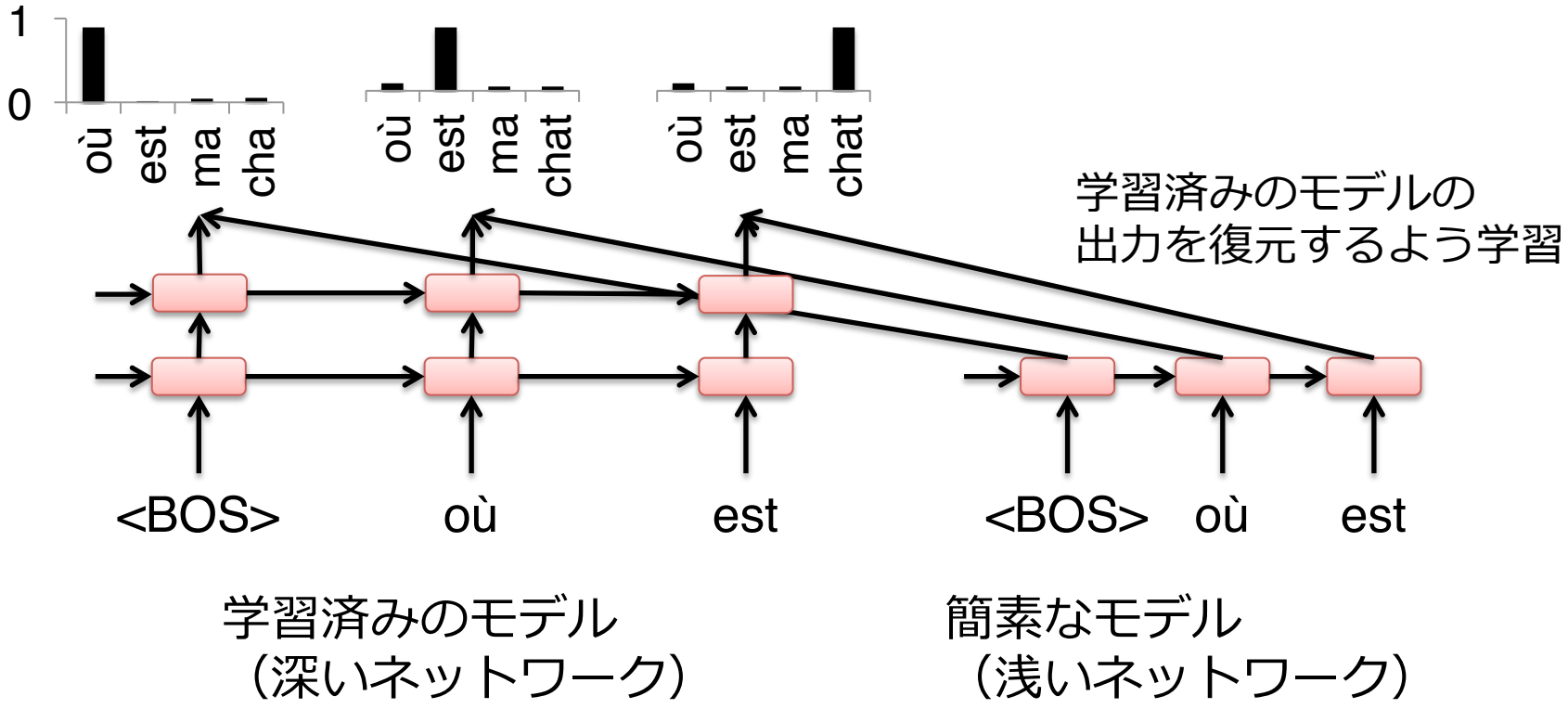
3 :
- モデルの簡素化 [Kim+]
- アライメントを陽に行う [Yu+]

3：Sequence-Level Knowledge Distillation [Kim+]

- 目的：Seq2seqのモデルを小さくする
  - 翻訳機をオフラインの状況やスマートフォンでも動かしたい
- 手法：学習済みのモデルを圧縮
  1. 不要なパラメータの除去
  2. knowledge distillation
- Knowledge distillationを利用

Knowledge distillation

- 学習済みのモデルの出力を簡素なモデルで復元



実験結果

Knowledge distillationの結果
パラメータを減らしつつ
同程度のBLEUを達成

不要なパラメータを
削除することで
さらに圧縮が可能

Model	Prune %	Params	BLEU	Ratio
4×1000	0%	221 m	19.5	1×
2×500	0%	84 m	19.3	3×
2×500	50%	42 m	19.3	5×
2×500	80%	17 m	19.1	13×
2×500	85%	13 m	18.8	18×
2×500	90%	8 m	18.5	26×

- 同程度のBLEUを達成しつつ圧縮に成功
- 実装：https://github.com/harvardnlp/seq2seq-attn

Seq2seqモデルの改良まとめ

- EMNLP2016において提案された改良手法
  - 正確な単語列を出力できるよう学習 [Wiseman+]
  - デコーダのネットワークを改良 [Wang+]
  - モデルの圧縮 [Kim+]

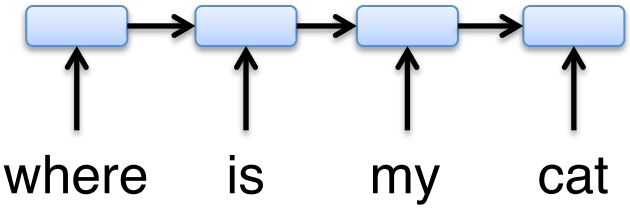
- これからどのような発展があるか？
  - 既存のアイデアをニューラル上で実装
    - [Wiseman+] のように既存のアイデアを取り込む
  - モデルの簡素化，計算の高速化
    - モデルの圧縮 [Kim+] を突き詰める
    - Convolutional NNなどを利用して計算を並列化
      - RNNとCNNの利点を取ったモデルも [Bradbury+ arXiv]

RNN vs. CNN

- LSTMなど，RNN系手法は並列化できない
- CNNなど，並列計算可能な方が高速なハズ
  - RNN系と比べたときの性能差は？

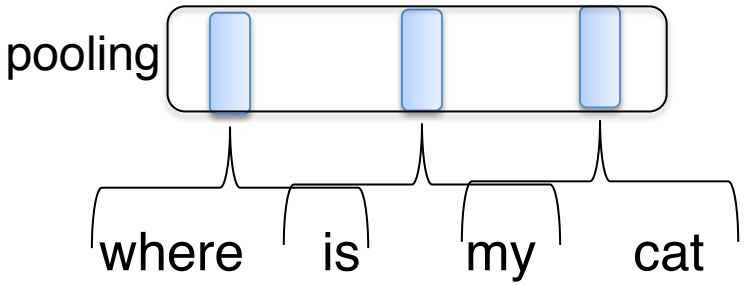
RNN系

前のステップの出力が
次のステップの入力なので
順に計算する必要がある



CNN

各 n-gram の計算は独立
（並列に計算可能）



CNNはエンコーダ／デコーダに使える？

- Convolutional Neural Network Language Models [Pham+]
  - 言語モデル（デコーダ部分）をCNNで実装，LSTMと比較

CNN

Model	k	w	Penn Treebank		
			val	test	#p
FFNN (Bengio et al., 2006)	128	-	156	147	4.5
Baseline FFNN	128	-	114	109	4.5
+CNN	128	3	108	102	4.5
+MLPConv	128	3	102	97	4.5
+MLPConv+COM	128	3+5	96	**92**	8
RNN (Mikolov et al., 2014)	300	1	133	129	6
LSTM (Mikolov et al., 2014)	300	1	120	115	6.3
LSTM (Zaremba et al., 2014)	1500	2	82	**78**	48

Penn TreebankでのperplexityPenn Treebankでのperplexity

- 使えるが，性能面ではLSTMに劣る（かもしれない）

まとめ

- EMNLP2016において，ニューラルネットに関連する研究は35%程度
  - ニューラルネットは言語処理でも，よく使われる道具のひとつになっている印象
  - 言語生成において強力なモデル，Seq2seqに様々な改良が見られた

- これからの方向性
  - 既存のアイデアをニューラル上で実装
  - モデルの簡素化，計算の高速化

参考文献

- Sequence to Sequence Learning with Neural Networks [Sutskever+ NIPS 14]
- Sequence-to-Sequence Learning as Beam-Search Optimization [Wiseman+ EMNLP 16]
- Memory-enhanced Decoder for Neural Machine Translation [Wang+ EMNLP 16]
- Controlling Output Length in Neural Encoder-Decoders [Kikuchi+ EMNLP 16]
- Neural Headline Generation on Abstract Meaning Representation [Takase+ EMNLP 16]
- Sequence-Level Knowledge Distillation [Kim+ EMNLP 16]
- Online Segment to Segment Neural Transduction [Yu+ EMNLP 16]
- Convolutional Neural Network Language Models [Pham+ EMNLP 16]
- Quasi-Recurrent Neural Networks [Bradbury+, arXiv]