

自然言語処理分野の 最前線

進藤 裕之

奈良先端科学技術大学院大学

2017-03-12

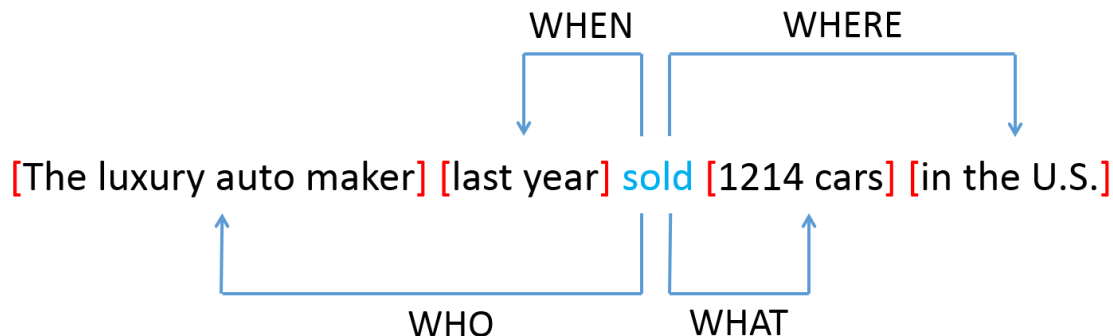
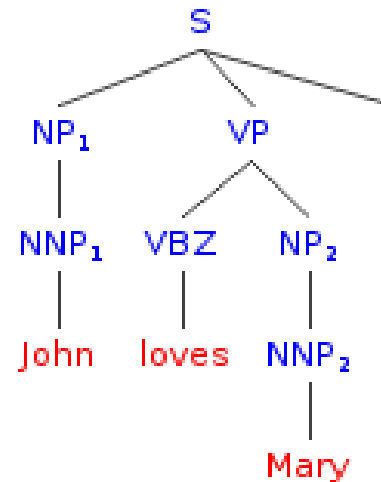
第五回ステアラボAIセミナー

- 進藤 裕之 (Hiroyuki Shindo)
- 所属： 奈良先端科学技術大学院大学
自然言語処理学研究室(松本研) 助教
- 専門： 構文解析, 意味解析
- @hshindo (Github)

これまでの取り組み

文の文法構造・意味構造の導出

- 構文解析
- 複単語表現解析
- 述語項構造解析



最近の取り組み(企業との共同研究)

国際会議WSDM 2017のコンペティション「WSDM Cup 2017」のTriple scoring taskにおいて、自然言語処理学研究室の佐藤元紀さん(博士前期課程1年)、進藤裕之助教と株式会社Studio Ousiaが共同で開発したシステムが準優勝しました。(2017/2/9)

2017年2月6日～10日、英ケンブリッジで開催された情報科学における著名な国際会議であるWSDM 2017のコンペティション「WSDM Cup 2017」のTriple scoring taskにおいて、自然言語処理学研究室の佐藤元紀君(博士前期課程1年)、進藤裕之助教と株式会社Studio Ousiaが共同で開発したシステムが準優勝しました(賞金\$750)。

この成果は、NAISTと株式会社Studio Ousiaとの共同研究における成果です。

本コンペティションには、世界中から21チームが参加し、提案した手法は、二位となりました。また、一位は中国の国立研究機関である中国科学院、三位は、情報科学の研究で著名な米イリノイ大学アーバナ・シャンペーン校が獲得しました。



最近の取り組み(企業との共同研究)

Triple scoring task [Bast+ SIGIR 2015] :

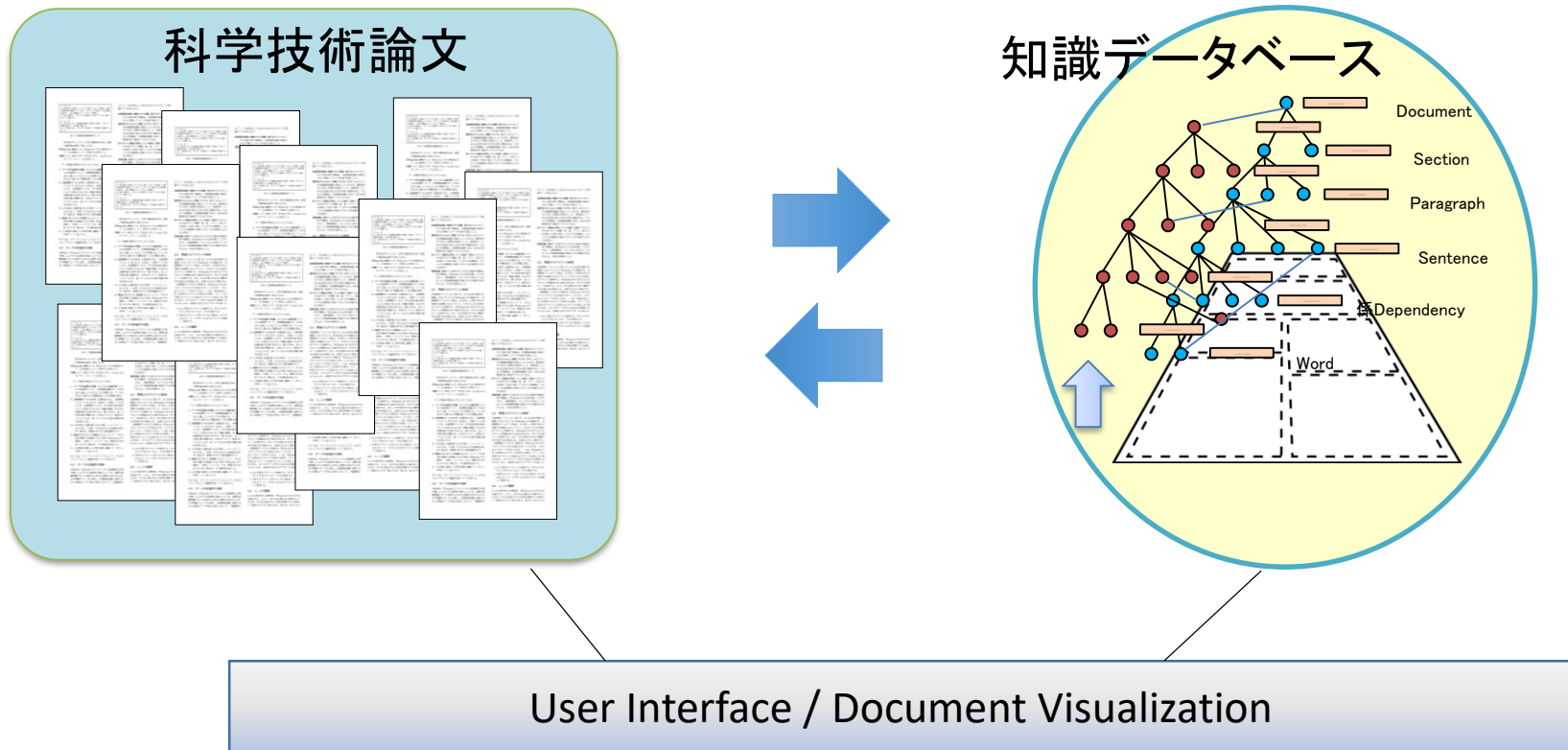
Wikipediaなどの知識ベース上にある人物の属性について、ユーザから見た妥当性を推定するタスク.

例えば「Barack Obama」は、Wikipedia上では政治家、作家、弁護士、教授などの様々な職業が付与されているが、多くのユーザは政治家として検索等の処理を行ってほしい.

このタスクでは、クラウドソーシングを使って作成された少量のアノテーションから、任意の人物に対する属性の妥当性を高精度に推定する.

最近の取り組み：論文解析

膨大な科学技術論文からの知識獲得・編集・検索



最近の取り組み：論文解析

1. PDFの解析

- PDF → XML(構造化テキスト)への自動変換
- 図表や数式の解析・意味理解

2. 論文からの情報抽出・知識獲得

- 計算機が論文を読んで理解する
- 得られた知識を自動でデータベース化する

3. 論文解析用のアノテーション・機械学習ツールの開発

ACL 2016の傾向

分野別採択数の上位

1. Semantics (意味)
2. IE, QA, Text Mining (情報抽出, 質問応答)
3. Tagging, Chunking, Parsing (解析系)
4. Machine Translation (機械翻訳)
5. Resources and Evaluation (データ構築と評価)

ACL 2016 Outstanding Papers

- A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task
- Learning Language Games through Interaction
- Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression
- Improving Hypernymy Detection with an Integrated Path-based and Distributional Method
- Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction
- Multimodal Pivots for Image Caption Translation
- Harnessing Deep Neural Networks with Logic Rules
- Case and Cause in Icelandic: Reconstructing Causal Networks of Cascaded Language Changes
- On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems
- Globally Normalized Transition-Based Neural Networks

Finding Non-Arbitrary Form-Meaning Systematicity

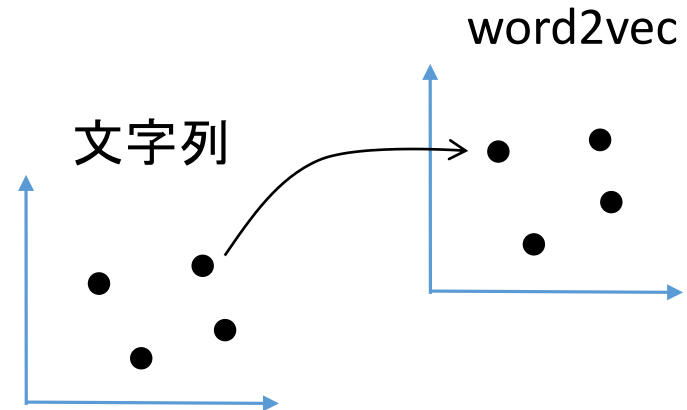
“Arbitrariness of the sign” [Saussure 1916]:
語形と意味は関係がない



本当かどうか
統計的に検証

手法：カーネル回帰(右図)

結果：語形と意味には高い相関がある(ものが存在する)ことを示した



ACL 2016 Outstanding Papers

- A Thorough Examination of the CNN/Daily Mail [Reading Comprehension Task](#)
- Learning Language Games through Interaction
- Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression
- Improving [Hypernymy Detection](#) with an Integrated Path-based and Distributional Method
- Integrating Distributional Lexical Contrast into Word Embeddings for [Antonym-Synonym Distinction](#)
- Multimodal Pivots for [Image Caption Translation](#)
- Harnessing Deep Neural Networks with [Logic Rules](#)
- Case and Cause in Icelandic: Reconstructing Causal Networks of Cascaded Language Changes
- On-line Active Reward Learning for Policy Optimisation in [Spoken Dialogue Systems](#)
- Globally Normalized Transition-Based Neural Networks

ACL 2016 Outstanding Papers

Reading Comprehension Task (文章読解)

Passage

(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

文章読解問題の分析

Question

characters in " @placeholder "
movies have gradually become
more diverse

Answer

@entity6

ACL 2016 Outstanding Papers

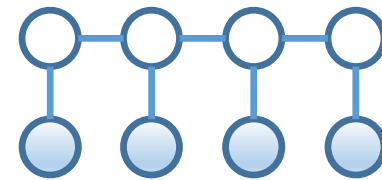
- Improving **Hypernymy Detection** with an Integrated Path-based and Distributional Method
 - 単語の上位・下位関係の予測
Ex. (pineapple, fruit), (green, color), (Obama, president)
- Integrating Distributional Lexical Contrast into Word Embeddings for **Antonym-Synonym Distinction**
 - 類義語・反意語の区別（どちらも同じ文脈で出現し得るので区別が難しい）

構造学習としての自然言語処理

1. 系列 → 系列

$$f_{\theta}(\text{seq} | \text{seq})$$

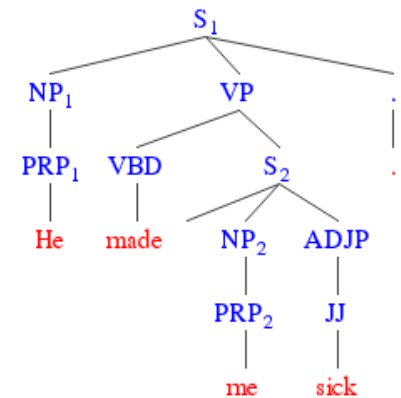
- 形態素解析・固有表現認識
- 機械翻訳, 自動要約
- 質問応答, 対話



2. 系列 → 木構造

$$f_{\theta}(\text{tree} | \text{seq})$$

- 構文解析



3. 系列 → グラフ構造

$$f_{\theta}(\text{graph} | \text{seq})$$

- 意味解析

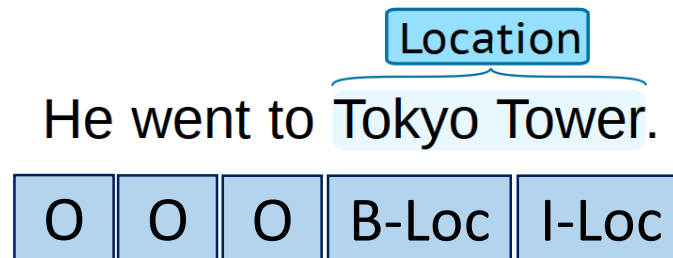
系列ラベリング

系列ラベリング

- 形態素解析(単語分割, 品詞タギング)



- 固有表現認識(人名, 会社名, 場所名, etc.)



BIOタギング

系列ラベリング

従来

人手で設計した
特徴量の抽出

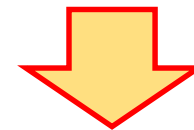
- 単語n-gram
- 文字n-gram
- それらの
組み合わせ



CRF

近年

ニューラルネットで
特徴量を計算(学習)

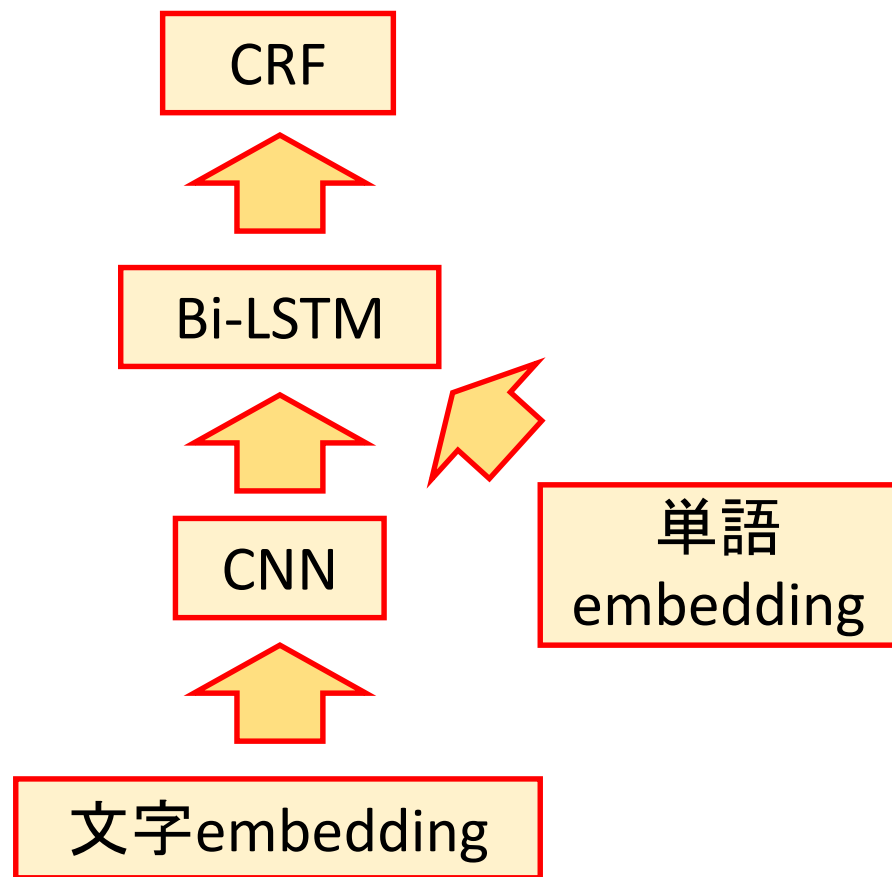
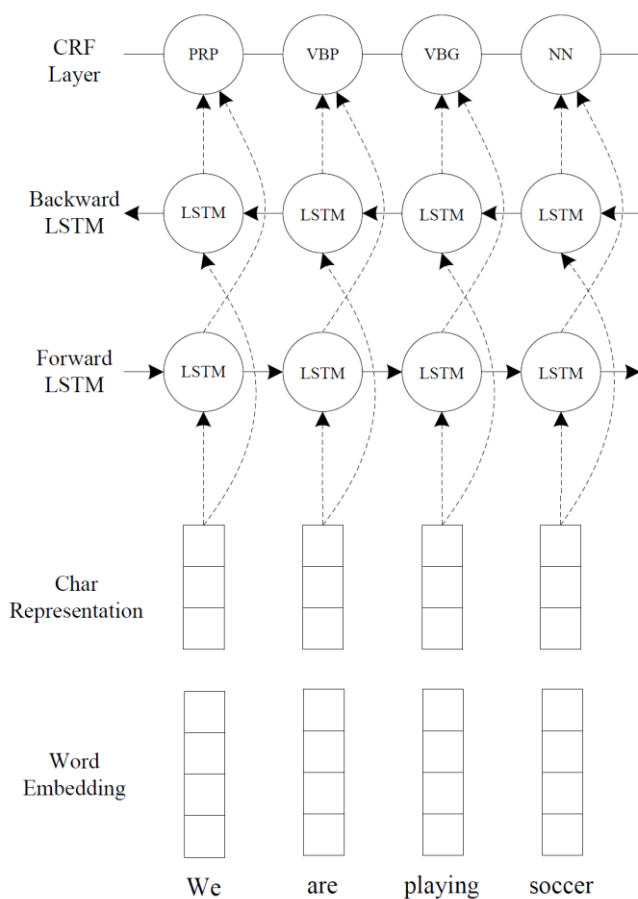


RNN

CRF

系列ラベリング

LSTM-CNNs-CRF

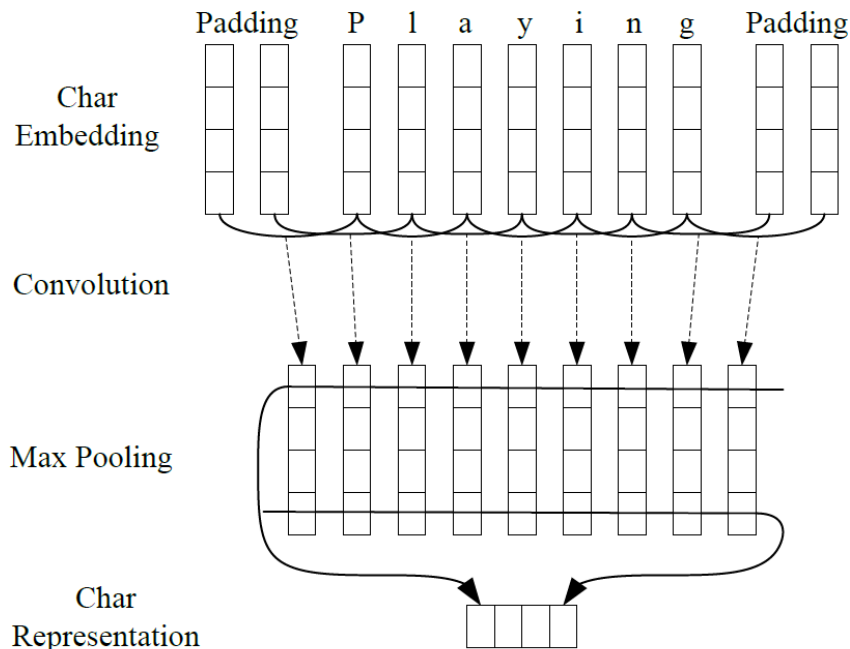


End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [Ma+2016]

系列ラベリング

LSTM-CNNs-CRF

CNN



テキストデータに対する
CNNの使い方
[Santos+ ICML 2014]

※可変長の文字列を固定長
の特徴量に変換

End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [Ma+2016]

系列ラベリング

LSTM-CNNs-CRF

品詞タグ付けと固有表現認識の結果(英語)

Model	POS		NER					
	Dev	Test	Dev			Test		
	Acc.	Acc.	Prec.	Recall	F1	Prec.	Recall	F1
BRNN	96.56	96.76	92.04	89.13	90.56	87.05	83.88	85.44
BLSTM	96.88	96.93	92.31	90.85	91.57	87.77	86.23	87.00
BLSTM-CNN	97.34	97.33	92.52	93.64	93.07	88.53	90.21	89.36
BRNN-CNN-CRF	97.46	97.55	94.85	94.63	94.74	91.35	91.06	91.21

系列ラベリング

LSTM-CNNs-CRF

品詞タグ付け(左図)と固有表現認識(右図)の結果

Model	Acc.
Giménez and Màrquez (2004)	97.16
Toutanova et al. (2003)	97.27
Manning (2011)	97.28
Collobert et al. (2011) [‡]	97.29
Santos and Zadrozny (2014) [‡]	97.32
Shen et al. (2007)	97.33
Sun (2014)	97.36
Søgaard (2011)	97.50
This paper	97.55

Model	F1
Chieu and Ng (2002)	88.31
Florian et al. (2003)	88.76
Ando and Zhang (2005)	89.31
Collobert et al. (2011) [‡]	89.59
Huang et al. (2015) [‡]	90.10
Chiu and Nichols (2015) [‡]	90.77
Ratinov and Roth (2009)	90.80
Lin and Wu (2009)	90.90
Passos et al. (2014)	90.90
Lample et al. (2016) [‡]	90.94
Luo et al. (2015)	91.20
This paper	91.21

End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [Ma+2016]

構文解析

(系列 → 木構造)

構文解析

SyntaxNet (Google)

“The World’s Most Accurate Parser” (当時)

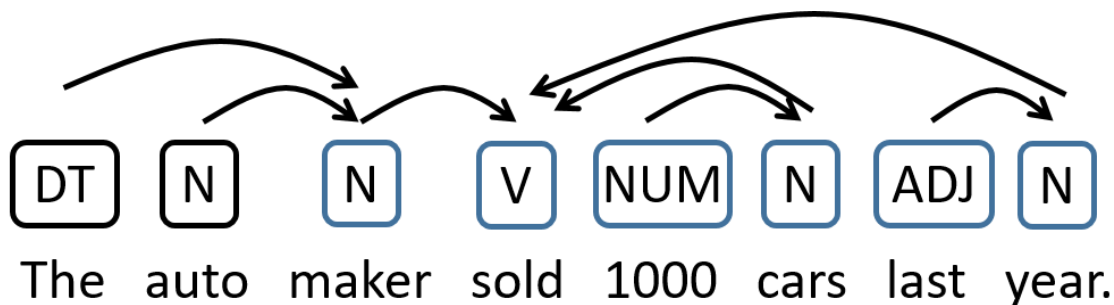
Globally Normalized Transition-Based Neural Networks [Andor+ ACL 2016]

構文解析

SyntaxNet (Google)

“The World’s Most Accurate Parser” (当時)

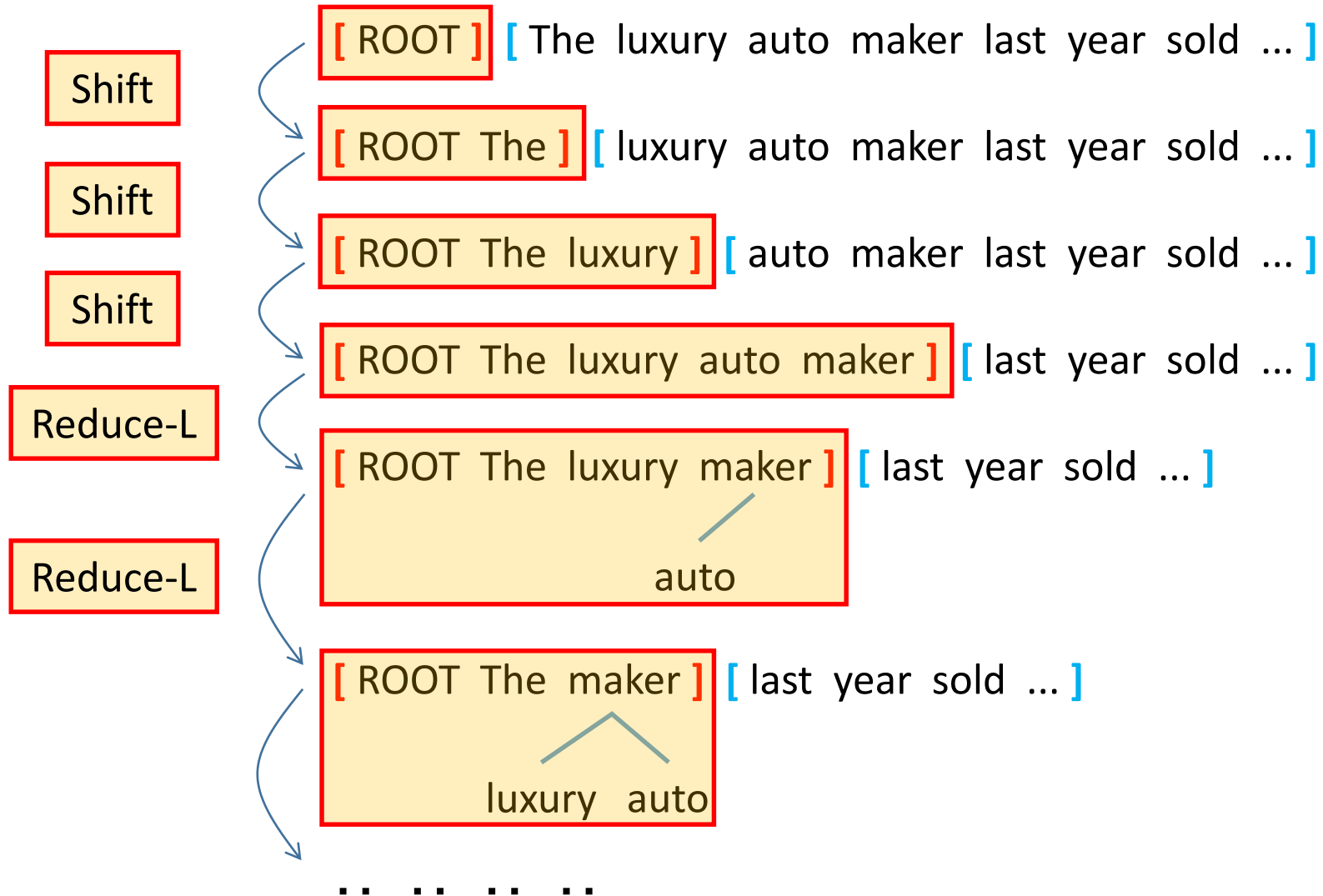
- 【前提知識】 依存構造解析 (係り受け解析)
 - 遷移型: 行動 (shift, reduce) 系列の出力によるデコード
 - グラフ型: 動的計画法によるデコード



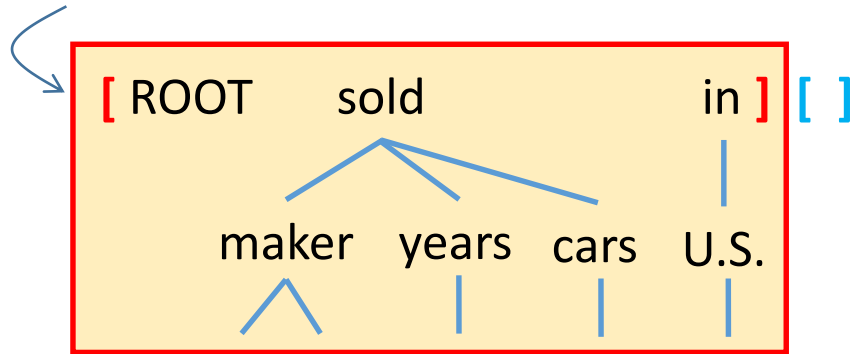
Globally Normalized Transition-Based Neural Networks [Andor+ 2016]

(参考) 遷移型依存構造解析

行動 (shift: 次の単語を見る, reduce: 木の一部を作る)

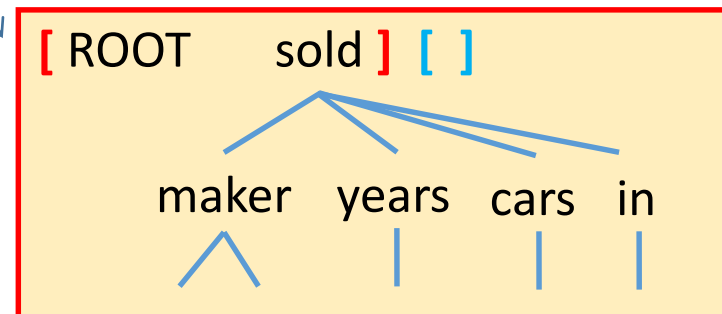
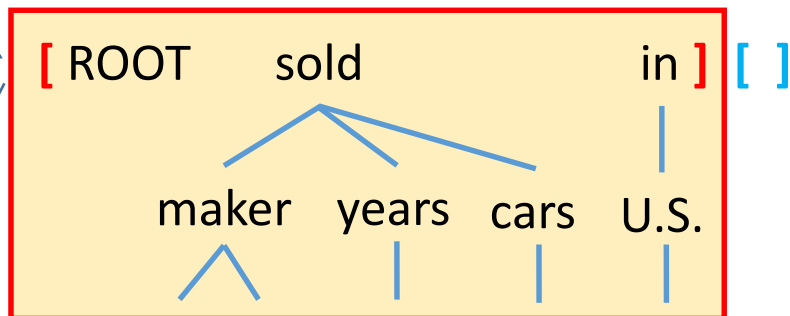


(参考) 遷移型依存構造解析



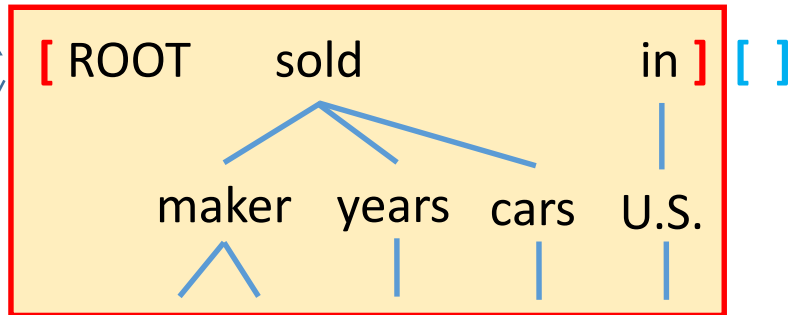
(参考) 遷移型依存構造解析

Reduce-R

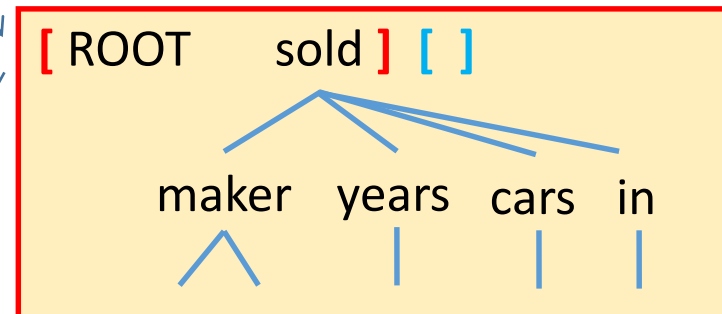


(参考) 遷移型依存構造解析

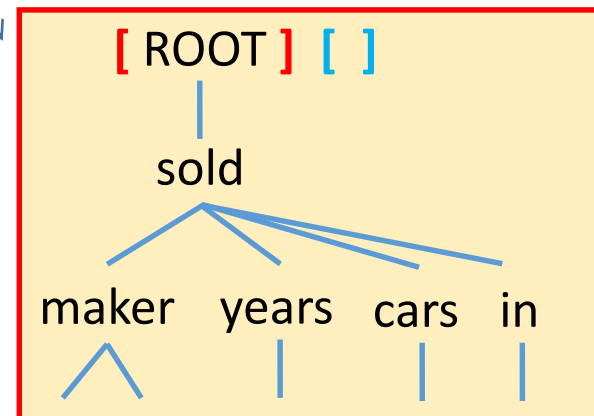
Reduce-R



Reduce-R



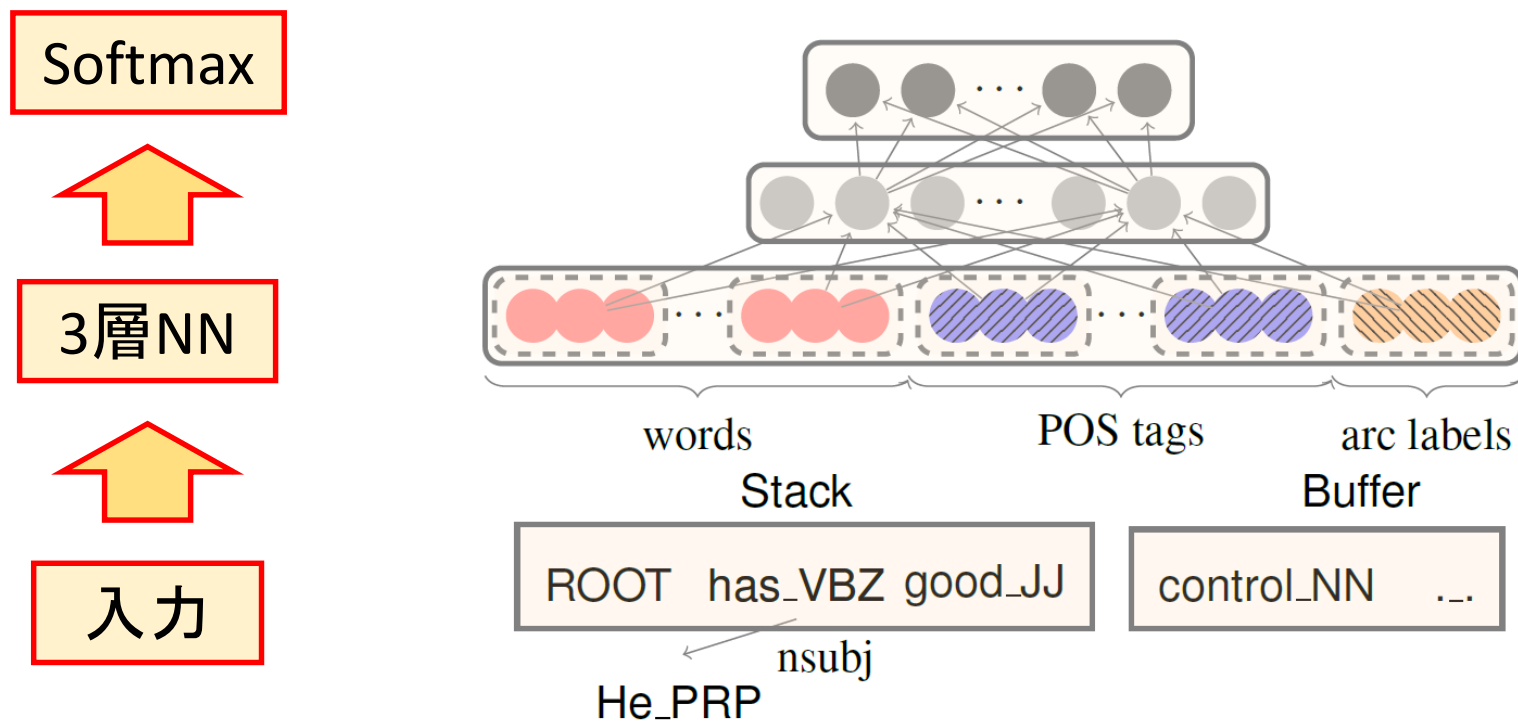
完成



構文解析

ニューラル遷移型依存構造解析 [Chen+ 2014]

次の行動を決定するためにニューラルネットでスコアリング

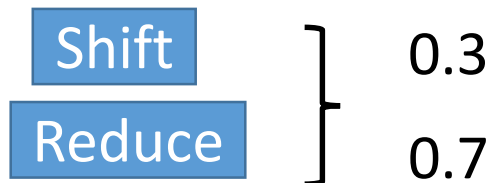


A Fast and Accurate Dependency Parser using Neural Networks [Chen+ 2014]

構文解析

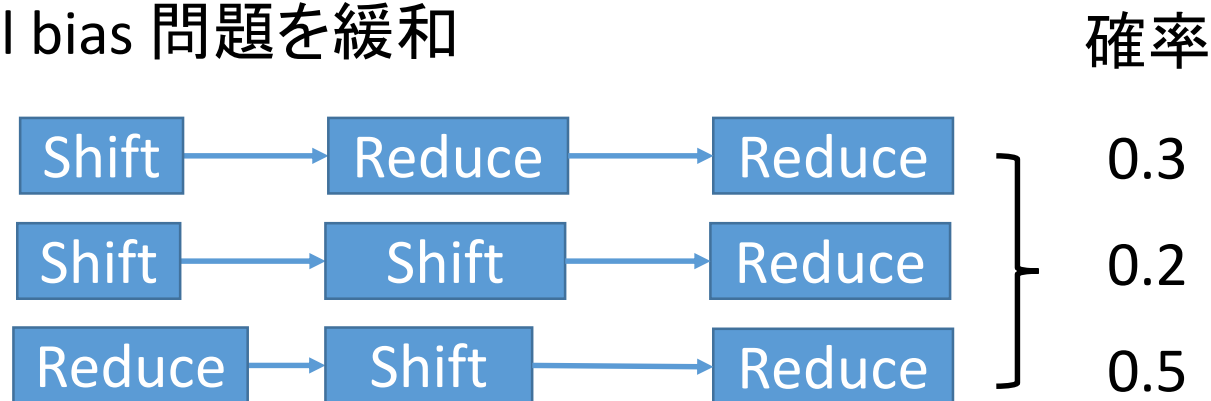
SyntaxNet (Google)

- [Chen+ 2014]では、各ステップで全**行動** (shift, reduce) の確率の和が1になる (local normalization)



- SyntaxNetでは、全**行動系列**の確率の和を1にする (global normalization)

→ label bias 問題を緩和

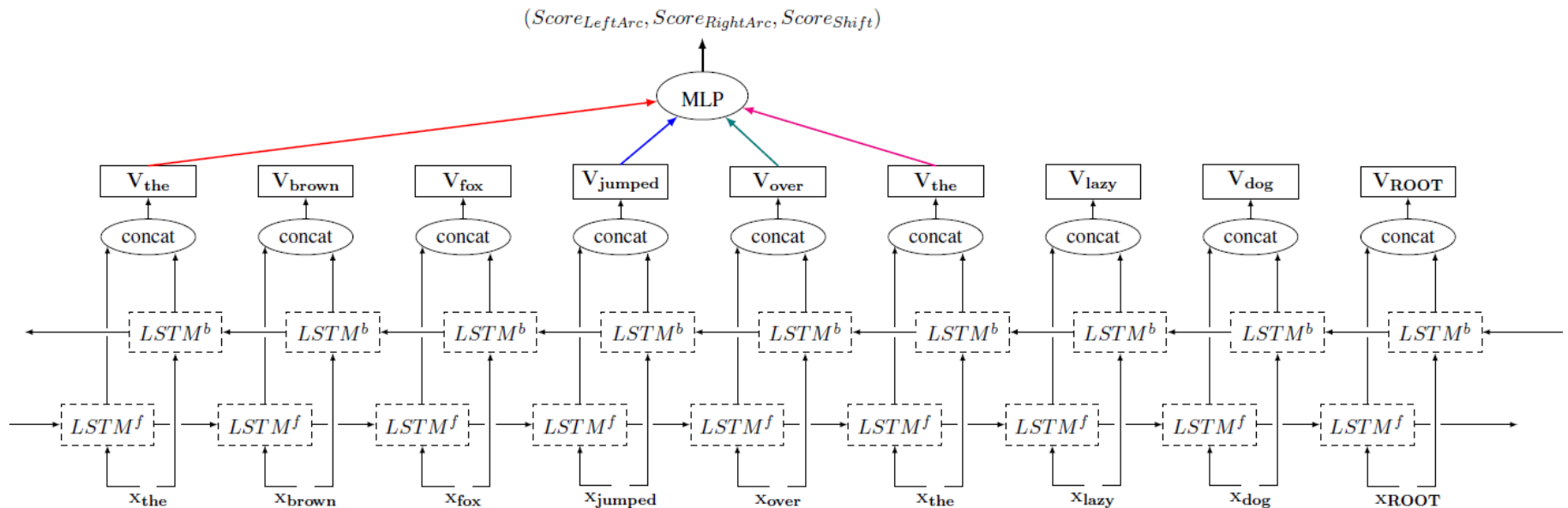


Globally Normalized Transition-Based Neural Networks [Andor+ 2016]

構文解析

Bi-LSTM Feature Representation

- 入力文全体から大域的な特徴量を学習して, 依存構造解析に用いる



Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations
[Kiperwasser+ 2016]

構文解析

Bi-LSTM Feature Representation

System	Method	Representation	Emb	PTB-YM	PTB-SD		CTB	
				UAS	UAS	LAS	UAS	LAS
This work	graph, 1st order	2 BiLSTM vectors	–	–	93.1	91.0	86.6	85.1
This work	transition (greedy, dyn-oracle)	4 BiLSTM vectors	–	–	93.1	91.0	86.2	85.0
This work	transition (greedy, dyn-oracle)	11 BiLSTM vectors	–	–	93.2	91.2	86.5	84.9
ZhangNivre11	transition (beam)	large feature set (sparse)	–	92.9	–	–	86.0	84.4
Martins13 (TurboParser)	graph, 3rd order+	large feature set (sparse)	–	92.8	93.1	–	–	–
Pei15	graph, 2nd order	large feature set (dense)	–	93.0	–	–	–	–
Dyer15	transition (greedy)	Stack-LSTM + composition	–	–	92.4	90.0	85.7	84.1
Ballesteros16	transition (greedy, dyn-oracle)	Stack-LSTM + composition	–	–	92.7	90.6	86.1	84.5

入力文全体からBi-LSTMで単語の特徴量を学習する：
単純だが、依存構造解析に対して効果が高い。

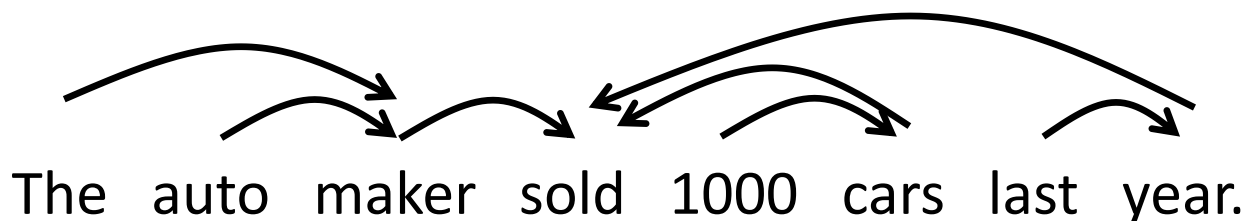
Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations
[Kiperwasser+ TACL 2016]

構文解析

Dependency Parsing as Head Selection

- 文全体から大域的な特徴量を学習する [Kiperwasser+ 2016]
- デコードはさらに単純化して、**各単語ごとに独立に**依存先 (head) の単語を選ぶ(！！)

※ 出力が木構造になる保証はない

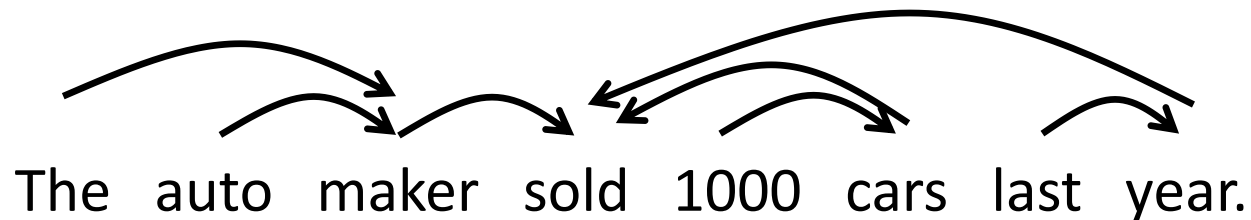


Dependency Parsing as Head Selection [Zhang+ 2016]

構文解析

Dependency Parsing as Head Selection

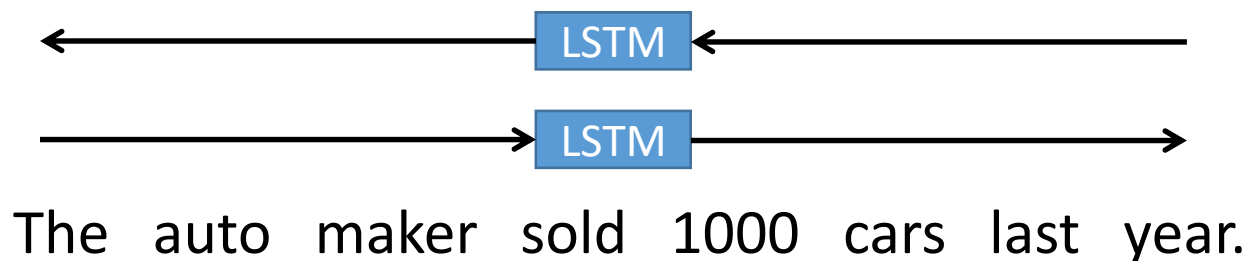
- 遷移型やグラフ型の依存構造解析は、ボトムアップに木を組み立てていく



構文解析

Dependency Parsing as Head Selection

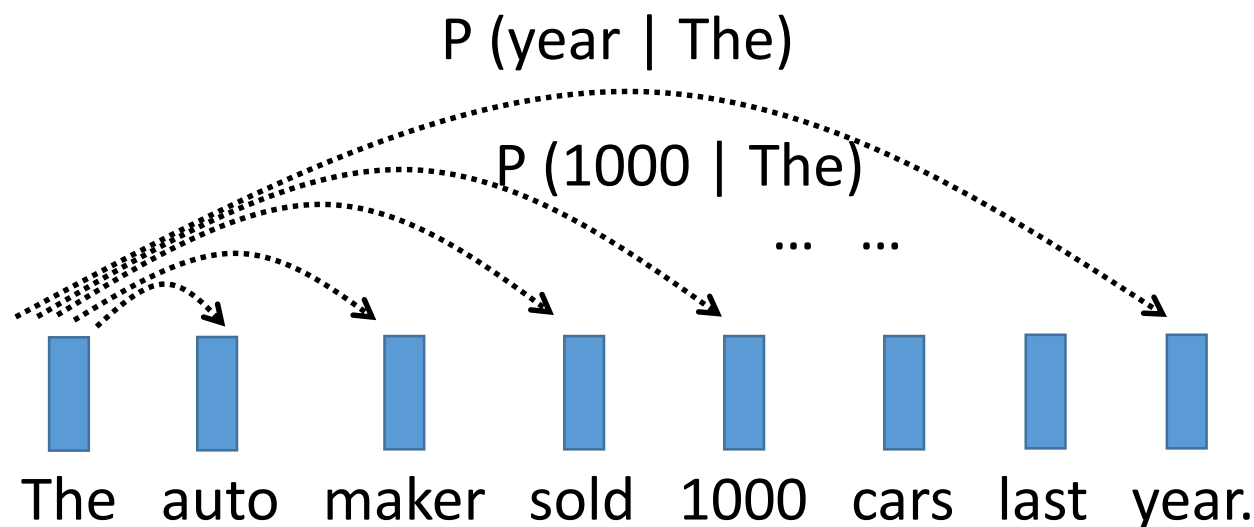
- 遷移型やグラフ型の依存構造解析は、ボトムアップに木を組み立てていく
- Head selectionでは、単語ごとに依存先を独立に決定する



構文解析

Dependency Parsing as Head Selection

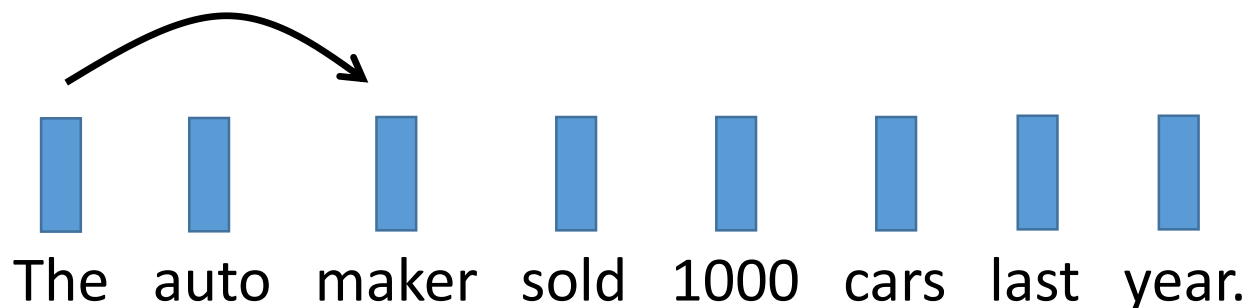
- 遷移型やグラフ型の依存構造解析は、ボトムアップに木を組み立てていく
- Head selectionでは、単語ごとに依存先を独立に決定する



構文解析

Dependency Parsing as Head Selection

- 遷移型やグラフ型の依存構造解析は、ボトムアップに木を組み立てていく
- Head selectionでは、単語ごとに依存先を独立に決定する



Dependency Parsing as Head Selection [Zhang+ 2016]

構文解析

Dependency Parsing as Head Selection

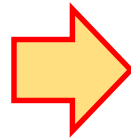
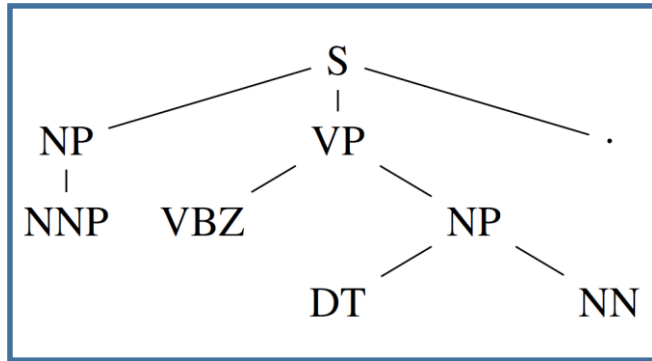
英語の依存構造解析の結果

Parser	Dev		Test	
	UAS	LAS	UAS	LAS
Bohnet10	—	—	92.88	90.71
Martins13	—	—	92.89	90.55
Z&M14	—	—	93.22	91.02
Z&N11	—	—	93.00	90.95
C&M14	92.00	89.70	91.80	89.60
Dyer15	93.20	90.90	93.10	90.90
Weiss15	—	—	93.99	92.05
Andor16	—	—	94.61	92.79
K&G16 <i>graph</i>	—	—	93.10	91.00
K&G16 <i>trans</i>	—	—	93.90	91.90
DENSE-Pei	90.77	88.35	90.39	88.05
DENSE-Pei+E	91.39	88.94	91.00	88.61
DENSE	94.17	91.82	94.02	91.84
DENSE+E	94.30	91.95	94.10	91.90

- 高精度
- 文長が長くなったときにどの程度の性能か要検証

構文解析(句構造)

木構造の線形化(linearization)



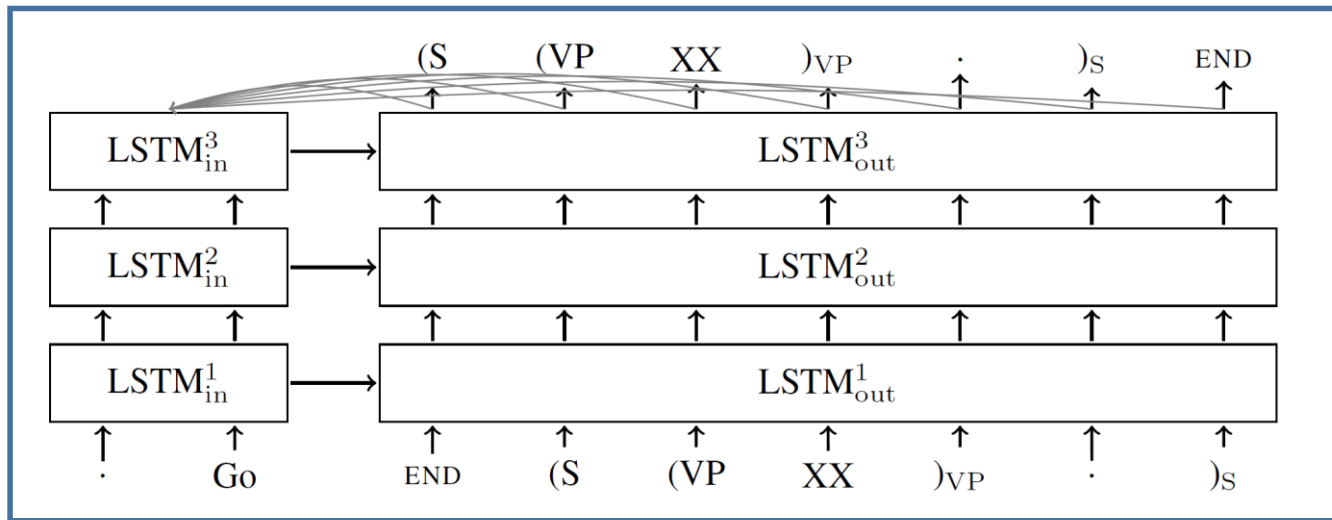
(S (NP NNP)_{NP} (VP VBZ (NP DT NN)_{NP})_{VP} .)_S

- 木構造を推定する問題を系列モデリング(3層LSTM)で解く

Vinyals et al., “Grammar as a Foreign Language”, Arxiv, 2015

構文解析(句構造)

木構造の線形化(linearization)



- モデルが不正な木構造を出力する割合は1.5%(意外と少ない)
- Attentionを入れないと精度が大きく低下
- 最終的に従来手法とほぼ同等の結果

Vinyals et al., “Grammar as a Foreign Language”, Arxiv, 2015

構文解析

それ以外にも

- Span-Based Constituency Parsing with a Structure-Label System and Provably Optimal Dynamic Oracles [Cross+ ACL 2016 Outstanding Paper]
- Global Neural CCG Parsing with Optimality Guarantees [Lee+ EMNLP 2016 Best Paper]

A*探索で最適な木構造を出力

系列から系列の生成 (Sequence-to-Sequence Learning)

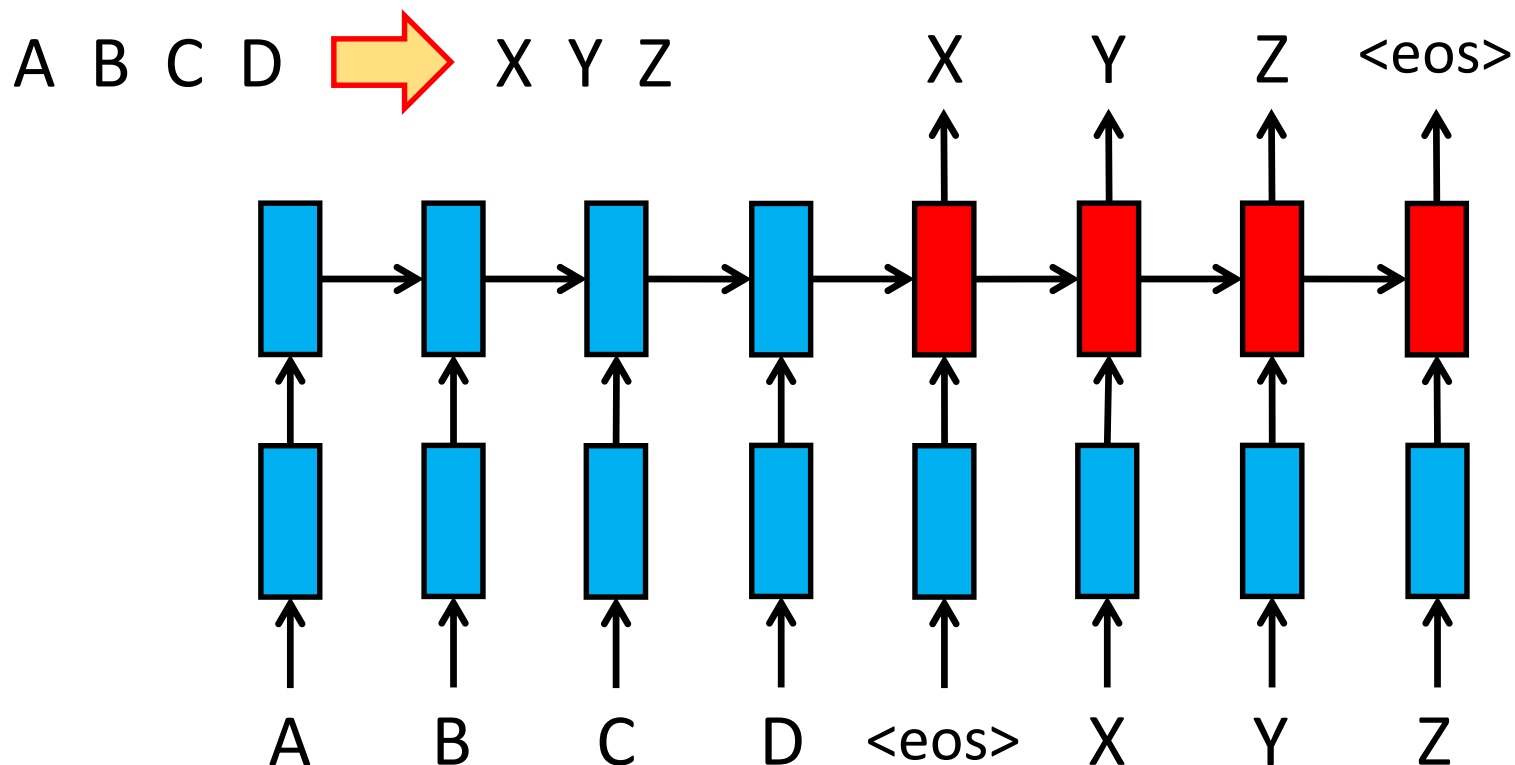
Seq2Seq Learning

応用例:

- 機械翻訳
- 自動要約
- 質問応答
- 対話
- 文法誤り訂正

機械翻訳

RNNによる機械翻訳のモデル化

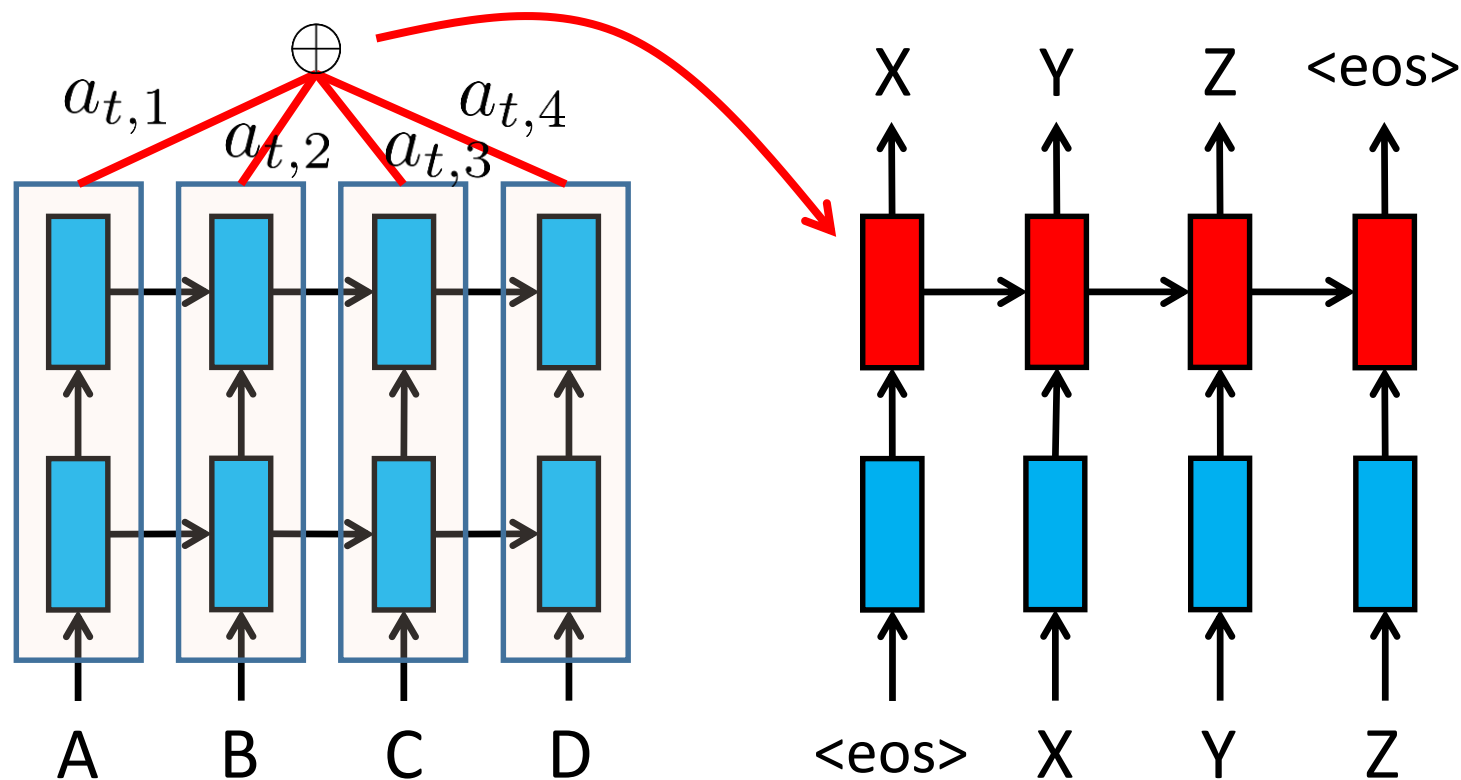


Sutskever et al., “Sequence to Sequence Learning with Neural Networks”, Arxiv, 2014

機械翻訳

アテンションに基づくRNN

どこに「注意」して翻訳するかを学習する

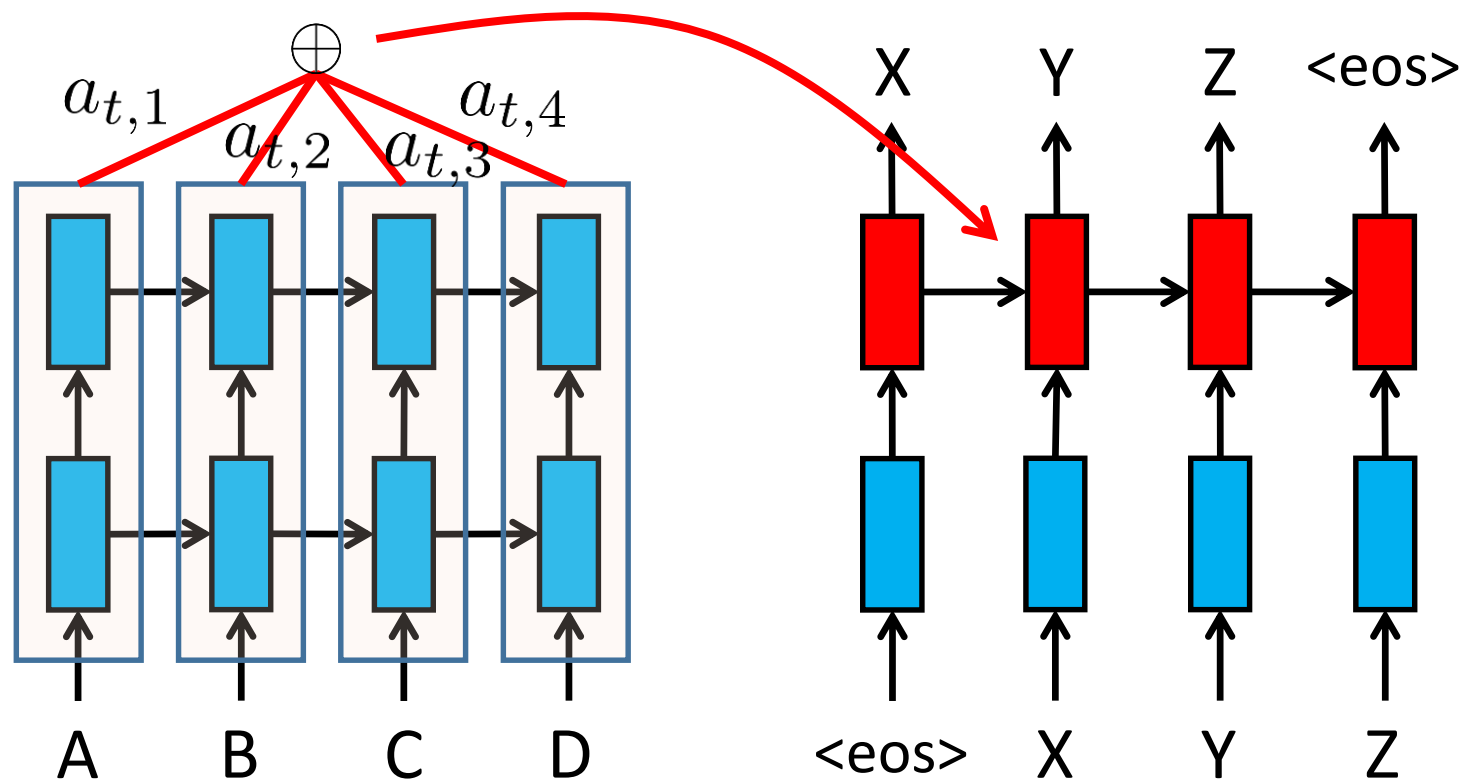


Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR, 2015

機械翻訳

アテンションに基づくRNN

どこに「注意」して翻訳するかを学習する

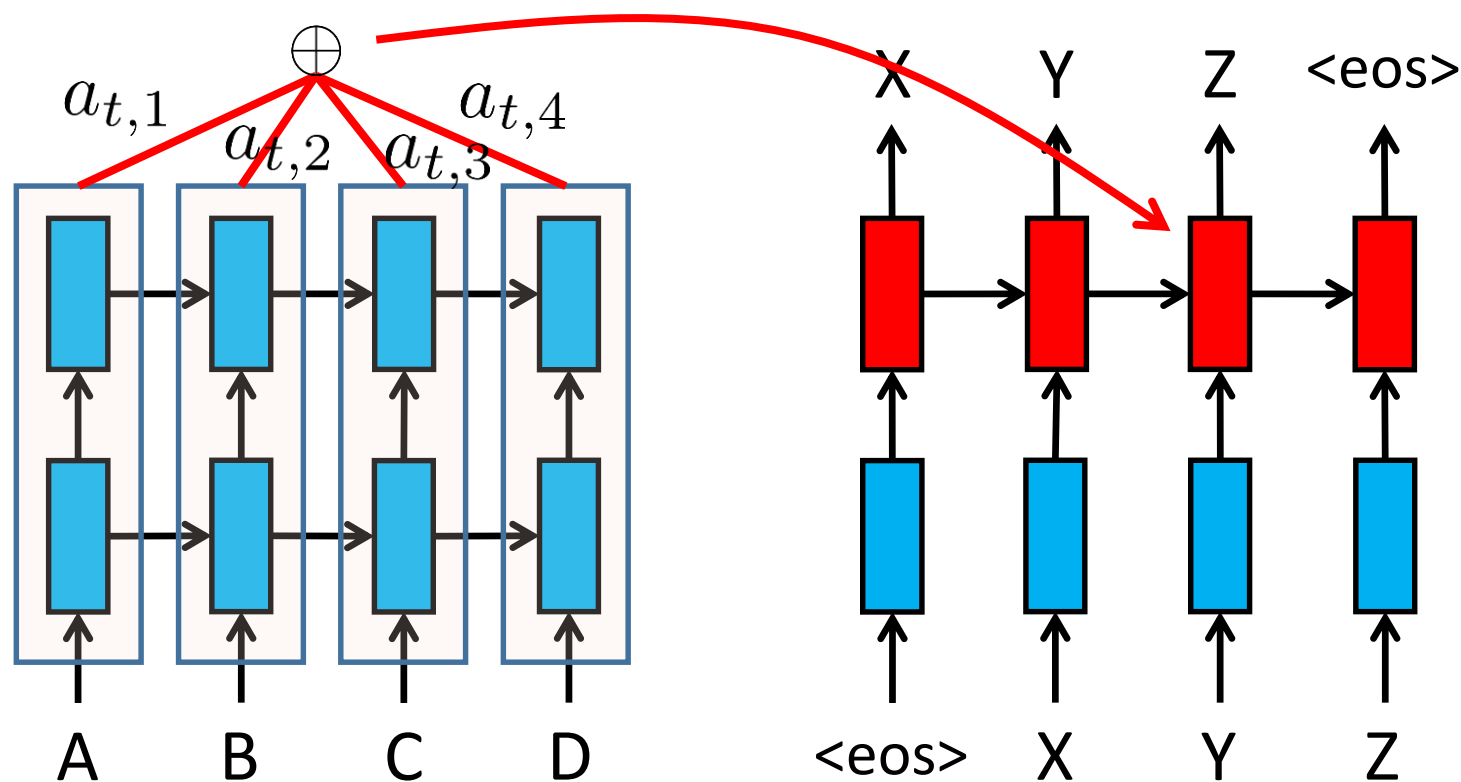


Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR, 2015

機械翻訳

アテンションに基づくRNN

どこに「注意」して翻訳するかを学習する

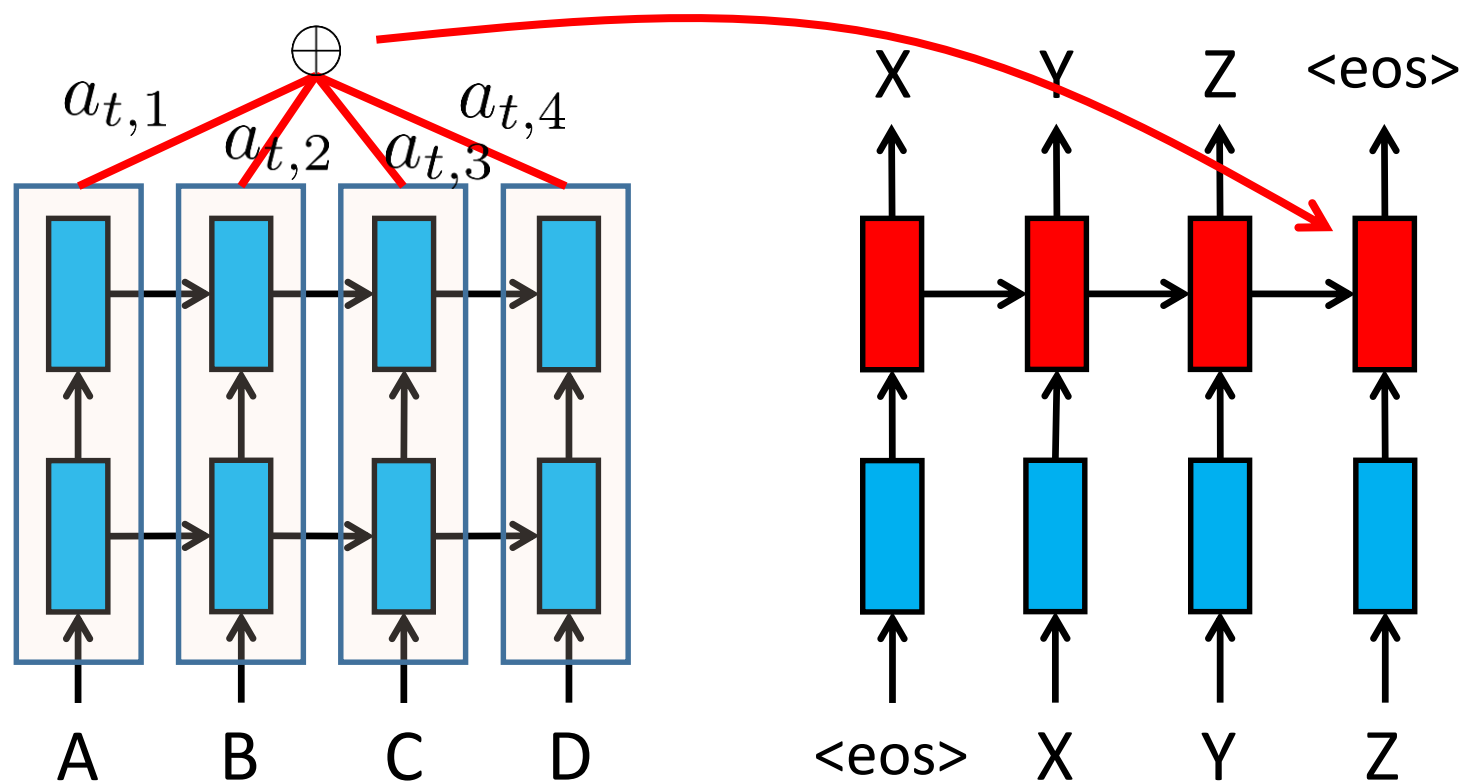


Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR, 2015

機械翻訳

アテンションに基づくRNN

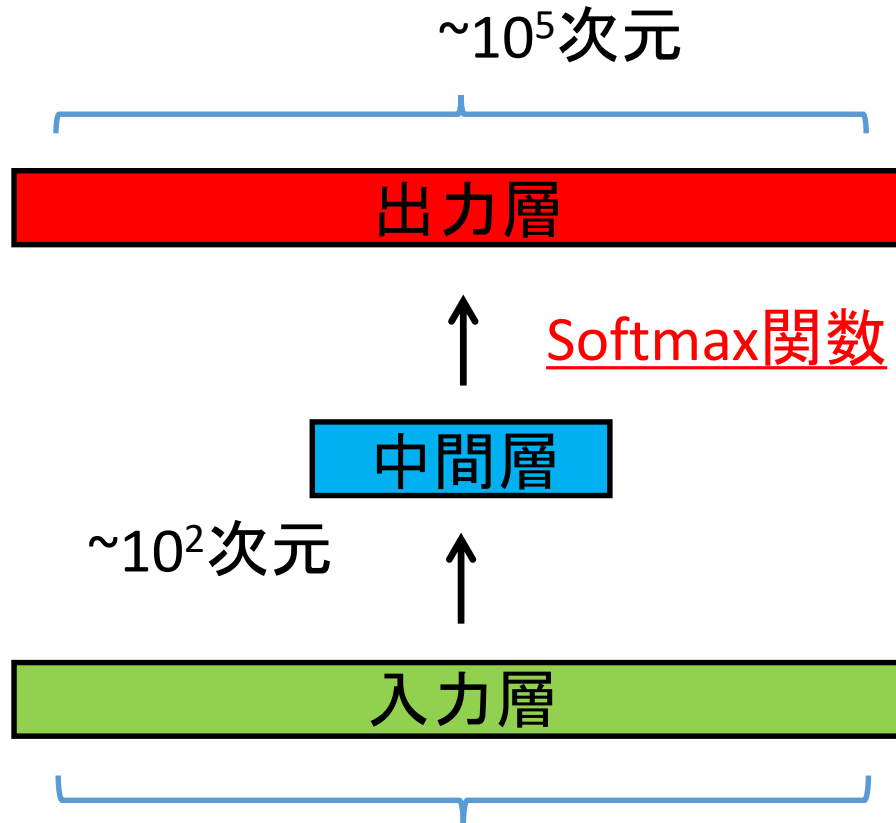
どこに「注意」して翻訳するかを学習する



Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR, 2015

単語ベース生成モデルの問題

単語を出力する系列モデルは出力層の計算が大変
未知語に弱い



$$\text{softmax}_i(\mathbf{z}) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$


次元数: $\sim 10^5$ (= 語彙数)

系列-系列の学習

サブ単語ベースの機械翻訳

- Byte pair encoding (BPE) [Gage 1994]を用いて単語分割を行う

出現頻度が高い2文字を, 別の1文字に置き換えていくことを繰り返して圧縮する

a b c d e f  ab c d e f

機械翻訳では, 人間と同じ基準の単語分割を行う必要はない

系列-系列の学習

サブ単語ベースの機械翻訳

英語↔ドイツ語の機械翻訳

name	segmentation	shortlist	vocabulary		BLEU		CHRF3		unigram F ₁ (%)		
			source	target	single	ens-8	single	ens-8	all	rare	OOV
syntax-based (Sennrich and Haddow, 2015)					24.4	-	55.3	-	59.1	46.0	37.7
WUnk	-	-	300 000	500 000	20.6	22.8	47.2	48.9	56.7	20.4	0.0
WDict	-	-	300 000	500 000	22.0	24.2	50.5	52.4	58.1	36.8	36.8
C2-50k	char-bigram	50 000	60 000	60 000	22.8	25.3	51.9	53.5	58.4	40.5	30.9
BPE-60k	BPE	-	60 000	60 000	21.5	24.5	52.0	53.9	58.4	40.9	29.3
BPE-J90k	BPE (joint)	-	90 000	90 000	22.8	24.7	51.7	54.1	58.5	41.8	33.6

まとめ

- 系列ラベリング
 - LSTM-CNNs-CRF
- 系列 → 木構造（主に構文解析）
 - 入力系列から大域的に特徴量を学習
→ デコードの方法を大幅に簡略化しても高精度
（動的計画法よりも, greedy探索, A*探索, pointwise）
 - 木構造を系列に変換して系列モデリングとして解く
- 系列の生成モデル(seq2seq learning)
 - 単語分割は教師なしで決める
（人間と同じでなくても良い）