

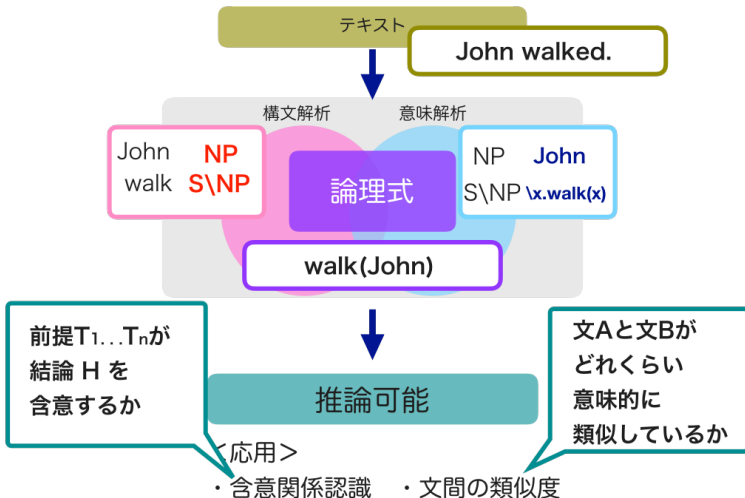
# RNN 変換モデル用いた高階論理からの文生成

馬目 華奈

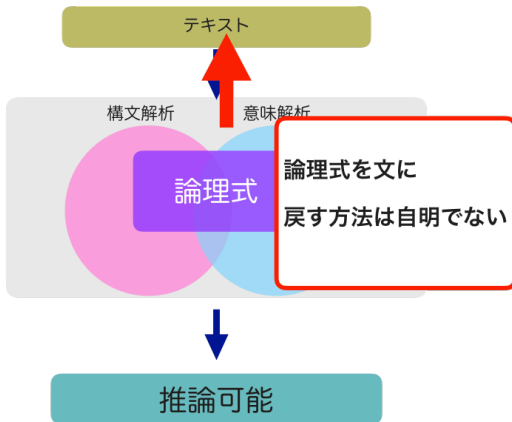
戸次研究室

卒業研究発表会  
February 6, 2018

# 研究背景



# 研究背景



<応用>

- ・ 含意関係認識
- ・ 文間の類似度

# 研究背景



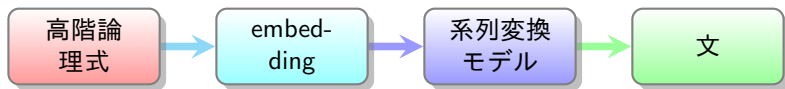
## 前提が足りない例

- All women ordered coffee or tea.
- Some woman ordered tea.

## 知識が足りない例

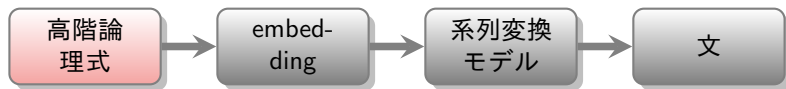
- All women ordered drink.
- Some woman ordered tea.

# 研究概要



- ニューラルネットによる系列変換モデルを用いて高階論理式から文を生成する手法を提案.
- 埋め込みの際, 4 種の手法 (記号, トークン, 木構造, グラフ) を検討する.

## 関連研究 1



- ニューラルネットによる系列変換モデルを用いて高階論理式から文を生成する手法を提案.
- 埋め込みの際, 4 種の手法 (記号, トークン, 木構造, グラフ) を検討する.

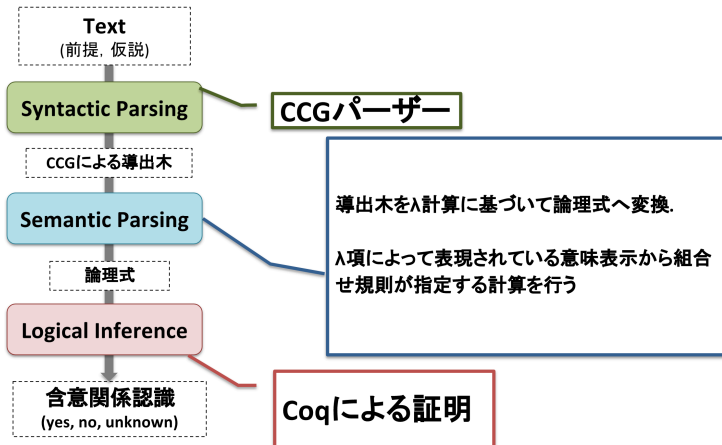
# 関連研究：CCG に基づく論理式による文の意味表現

## cgc2lambda

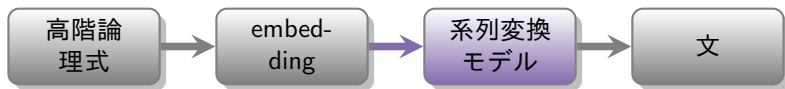
<https://github.com/mynlp/cgc2lambda>

Mineshima+ [EMNLP2015]

Martínez-Gómez+ [ACL2016]



## 関連研究 2

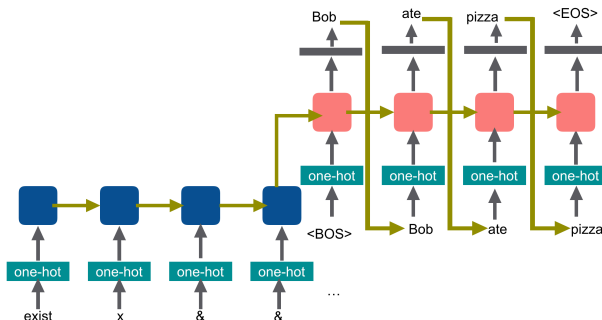


- ニューラルネットによる系列変換モデルを用いて高階論理式から文を生成する手法を提案.[?]
- 埋め込みの際, 4 種の手法 (記号, トークン, 木構造, グラフ) を検討する.

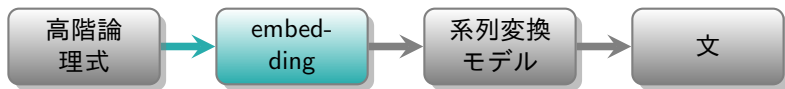


## 関連研究：系列変換モデル

- 入出力がシーケンスとなるニューラルネットのモデル
- エンコータ：入力列を再帰型 NN により隠れ状態ベクトルに変換
- デコーダ：隠れ状態ベクトルを初期値とし、隠れ状態と自身のこれまでの出力結果を基に次のトークンを生成



# 提案手法



- ニューラルネットによる系列変換モデルを用いて高階論理式から文を生成する手法を提案.
- 埋め込みの際, 4 種の手法  
(記号, トークン, 木構造, グラフ) を検討する.

## 提案手法：論理式埋め込み 1

Bob walked の論理式：

$\text{exists } x.((x = \text{Bob}) \ \& \ \text{exists } e.(\text{walk}(e) \ \& \ (\text{Subj}(e) = x)))$

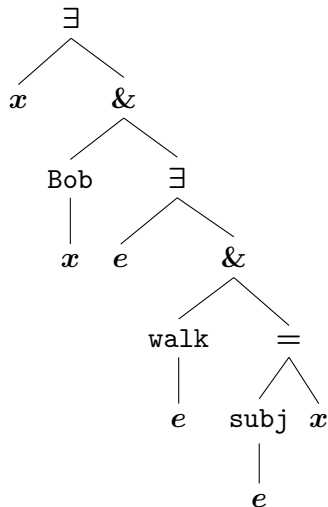
記号ごとに区切る

$[e, x, i, s, t, s, \_, x, \_, ., \_, (, (, x, \_, =, \dots]$

トークンごとに区切る

$[\text{exists}, x, (, (, x, =, \text{Bob}, ), \&, \text{exists}, \dots]$

## 提案手法：論理式埋め込み 2

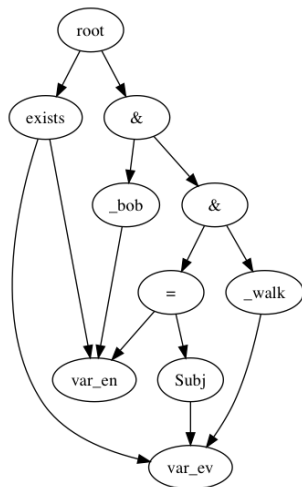


### 木構造

[exists,  $x$ , &, =  
, Bob,  $x$ , exists,  $e$ , &, ...]

- 論理式をポーランド記法に変換する  
(論理演算子を前にもってくる)
- pre-order の深さ優先探索でたどる

## 提案手法：論理式埋め込み 3



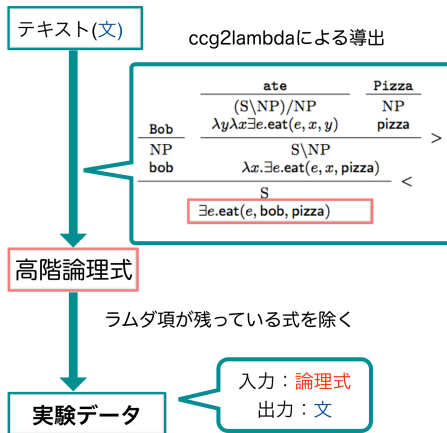
### グラフ構造

[exists,  $x$ ,  $x$ , &, Bob, &, =  
, Subj, walk...]

- 論理式をポーランド記法に変換する  
(論理演算子を前にもってくる)
- 同じ変数をさす様に edge を変更
- pre-order の深さ優先探索でたどる

# 提案手法：データセット

- SNLI を用い論理式と文のペアを作成
- 60 単語以内の文例を対象 train:9140/dev:2285/test:1500



## 実験：実験設定

- 系列変換モデルによる文生成（入力：論理式，出力：文）
- トークンベースの LSTM の出力次元数を LSTM の出力次元数を 256

|       | 記号    | トークン  | 木構造   | グラフ   |
|-------|-------|-------|-------|-------|
| 入力語彙数 | 70    | 5,118 | 5,107 | 4,991 |
| 出力語彙数 | 78    | 7,214 | 7,214 | 7,214 |
| 入力列最長 | 2,097 | 699   | 451   | 259   |
| 出力列最長 | 270   | 55    | 53    | 53    |

### 環境，ライブラリ

- tsubame サーバ（メモリ 240GiB, GPU × 4）
- ニューラルネットのモデル：Keras
- logic 関係：nltk

## 実験：評価方法

### BLEU による評価

$$score = BP \exp \left( \sum_{i=1}^N \frac{1}{N} \log P_n \right)$$

$$BP = \begin{cases} 1 & (c \geq r) \\ \exp \left( 1 - \frac{r}{c} \right) & (c < r) \end{cases}$$

$$P_n = \frac{\sum_{i=0} \text{出力文 } i \text{ 中と解答文 } i \text{ 中で一致した } n\text{-gram 数}}{\sum_{i=0} \text{出力文 } i \text{ 中の全 } n\text{-gram 数}}$$



# 実験：実験結果

## BLEU 評価

| 指標   | 記号   | トークン | 木構造  | グラフ  |
|------|------|------|------|------|
| BLEU | 34.9 | 39.7 | 41.8 | 44.7 |

|        |                                      |
|--------|--------------------------------------|
| 文      | Two surgeons are having lunch.       |
| 記号単位   | Two children are playing basketball. |
| トークン単位 | Two entertainers are having fun.     |
| 木構造    | Two teams are having a brawl.        |
| グラフ    | Two brothers are having a picnic.    |

|        |                                     |
|--------|-------------------------------------|
| 文      | The towel is pink and blue striped. |
| 記号単位   | A horse is talking to each other.   |
| トークン単位 | A guy snipping a ladys hair.        |
| 木構造    | The blue is blue and blue.          |
| グラフ    | A blue tractor is wearing blue.     |

## まとめ

- 系列変換モデルを用いて高階論理式から文を生成する手法を提案した.
- 評価をする際, ccg2lambda を用いてデータセットを作成した.
- 提案手法の評価を行った結果, BLEU スコアは, トークン単位で区切り, 論理式の計算の順序を考慮することで高くなった.

## 今後の課題

- 他の意味表現からの文生成との比較を行う.
- 他のデータセット（英語，日本語）でも行う.
- 逆変換（文→論理式）を行い，モデルを評価する.
- 評価方法に，文類似度を使用するなど文生成における評価方法を工夫する.
- アテンション付き系列変換モデルやコピー機構を用いるなどモデルの改良に取り組む.

## 参考文献