# 前方文脈の埋め込みを利用した日本語述語項構造解析

東北大学工学部　乾・鈴木研究室

今野 颯人，松林 優一郎，大内 啓樹，清野 舜，乾 健太郎

# 述語項構造とは

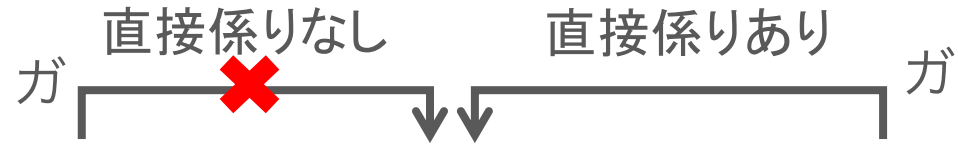
- 述語項構造
  - 文章内の**述語**とその**項**の間の関係を規定する構造


述語項構造を解析することで、文の理解に重要な
構造化された意味関係を獲得することができる

# ゼロ照応

日本語では、**述語**と意味的関係を持つ**項**が
直接の係り関係にない事例(ゼロ照応)がたびたび現れる

■文内ゼロ




(φ ガ) 風邪を ひいた ので、[私] は学校を 休んだ。

■文間ゼロ (項が他の文中に存在)


[彼女] は パン を 食べた。 (φ ガ) 牛乳 も 飲んだ。

■外界ゼロ (項が文書内にない)

(φ ガ) 映画館 に 行こう かな。　　( ガ: [一人称] )

# ゼロ照応

- ゼロ照応の特徴
  - **複雑な構文構造**をしていることが多く、**統語的な手がかりが少ない**
- 解析精度 (最高精度の既存研究[1])
  - $F_1$ 値で58%程度 (文内ゼロ)

(※直接係り関係ありは$F_1$ 値で91%程度)

[1] Matsubayashi+'18

- 事例数
  - 文内ゼロ：全体の**約20%**存在、無視できないほど存在

ゼロ照応解析の精度向上は
日本語の述語項構造解析における主要な課題

# 本研究の目的

- 本研究の目的：前方文脈埋め込みによる
  文内ゼロ照応の解析精度向上

  - ゼロ照応の事例は**統語的な手がかりが少ない**
  - **前方文脈を参照**することで、複雑な構文構造を持った文の**解析難易度を緩和**できないか

# 本研究の目的

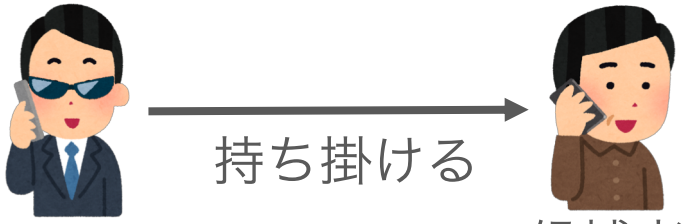
- 文内ゼロのみでなく、**文間ゼロ**、**外界ゼロ**も**同時に解く**ということも大きな課題
  - しかし、解析対象を**文内ゼロに限った場合**の方が同時に解くよりも**解析精度が良い**
  - 文内ゼロを解くに当たっても前方文脈を利用し、解析精度の向上を狙う
- 解析対象を文内ゼロに限り、前方文脈の埋め込みによる効果を検証する

# 解析難易度の高い事例

- 解析対象文

> …男性の声で電話がかかり、**[候補者 ニ格]**自身が出ると
> 「投票用紙を一枚五、六万円で買わないか」と持ち掛け、
> 候補者以外が電話に出ると、…
>
> (NAIST Text Corpus: 950113-0145-950113207.ntc)

- 述語：持ち掛ける
- 項：候補者 (ニ格)




- **ゼロ照応**
- **複雑な構文構造**や「電話をかけてきた人が電話に出た人に何かの用事がある」といった**常識的知識**を**理解しなければ導けない**

# 解析難易度の高い事例

- 解析対象文の前方1文

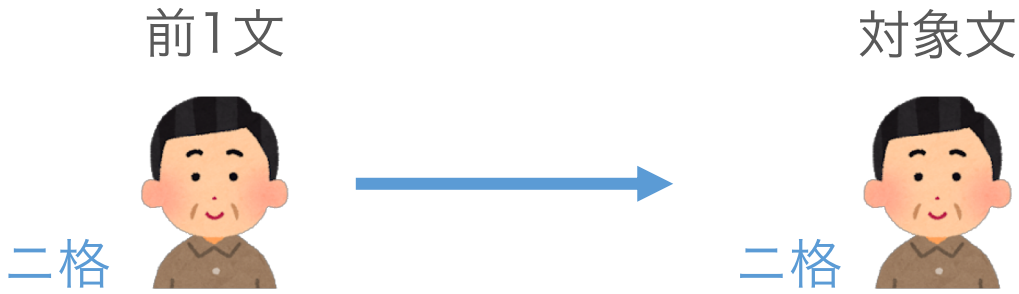
> なにものか**が**大量にコピーした偽造用紙**を**、
> 本物の投票用紙と偽って候補**[者 ニ格]に**売り付けようとしていたとみて調べている。

- 述語：売り付ける
- 項　：なにもの (ガ格)、偽造用紙 (ヲ格)、候補者 (ニ格)

■述語と項の**直接的な係り受け関係**をもって**明瞭に**書かれている

# 前方文脈の情報を解析に用いる理由

- 前方文脈で文の主役となっているエンティティは、後方の文でも項として言及されることが多い

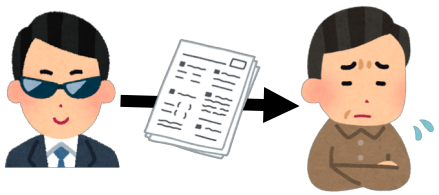
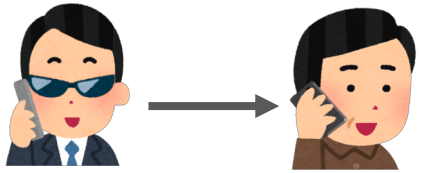



前一文: なにものかが大量にコピーした偽造用紙を、本物の投票用紙と偽って候補**[者 ニ格]**に売り付けようとしていたとみて調べている。

解析対象文: …男性の声で電話がかかり、候補**[者 ニ格]**自身が出ると「投票用紙を一枚五、六万円で買わないか」と持ち掛け、候補者以外が電話に出ると、…

# 前方文脈の情報を解析に用いる理由

- 前文で示される直接的な意味関係を用いて
  複雑なケースの解析を容易にできる可能性がある
  - 「候補者に売りつける」ために「候補者に持ち掛ける」
    といった**意味的なつながり**をとらえる

売りつける   持ち掛ける

前一文: なにものかが大量にコピーした偽造用紙を、
本物の投票用紙と偽って候補**[者 ニ格]**に売り付けようと
していたとみて調べている。

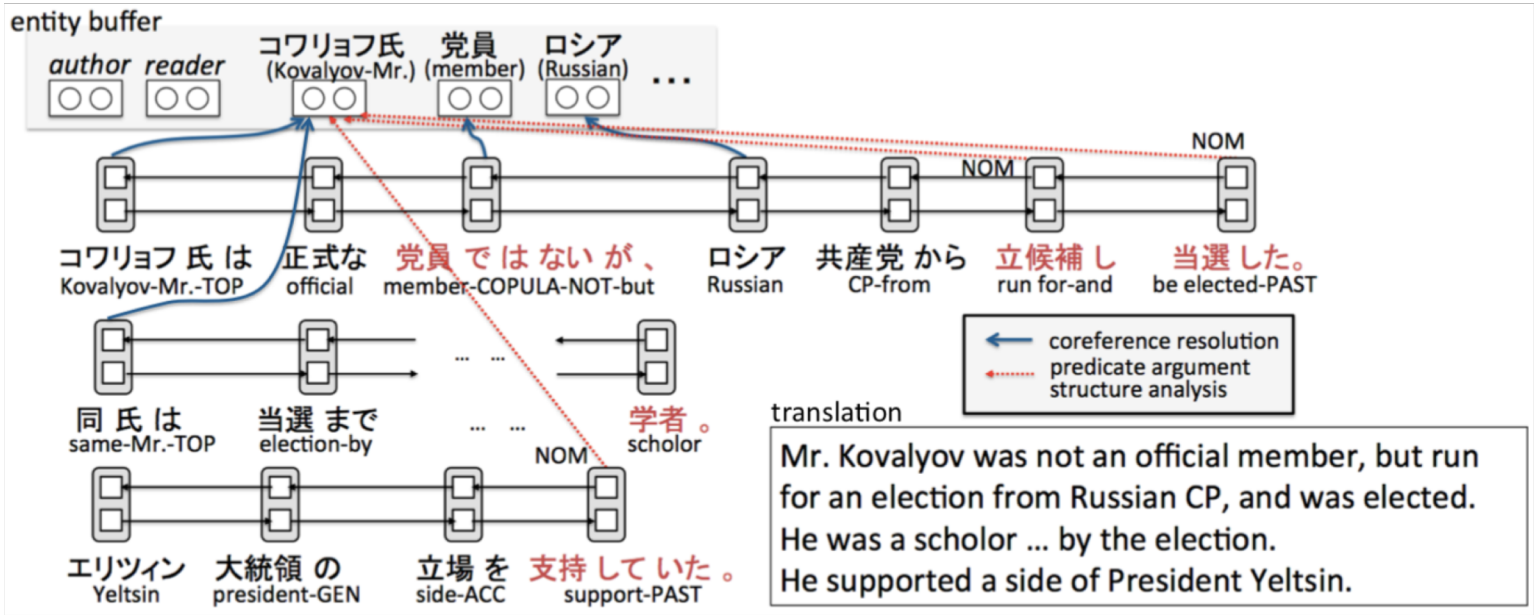
解析対象文: …男性の声で電話がかかり、候補**[者 ニ格]**自身が出ると
「投票用紙を一枚五、六万円で買わないか」と持ち掛け、
候補者以外が電話に出ると、…

# 既存手法

- 日本語述語項構造解析の直近の研究
  - 多層双方向RNNを用いたend-to-endの解析モデルが複数提案[1-3]
    - [1] Matsubayashi+'18
    - [2] Ouchi+'17
    - [3] Kurita+'18
  - これらのモデルは解析対象の述語が含まれる**1文の情報のみ**から解析を行う
  - 文を超えた文脈情報を捉えることはできない

# 前方文脈を利用する既存手法

- 前方文脈を利用する既存のニューラル解析モデル [4]
  - **エンティティベクトルを更新**していくことで文脈を考慮
  - [4] Shibata+'18



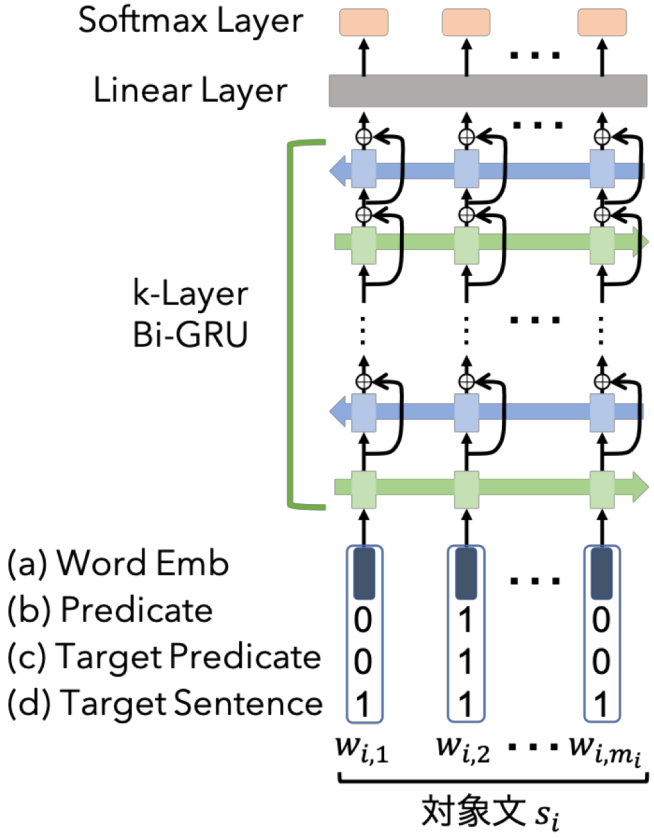
# 提案手法

- 既存手法 [4]
  - [4] Shibata+'18
  - 前方の文の共参照解析や項構造解析を陽に解き、**エンティティベクトルを更新**することで文脈を考慮

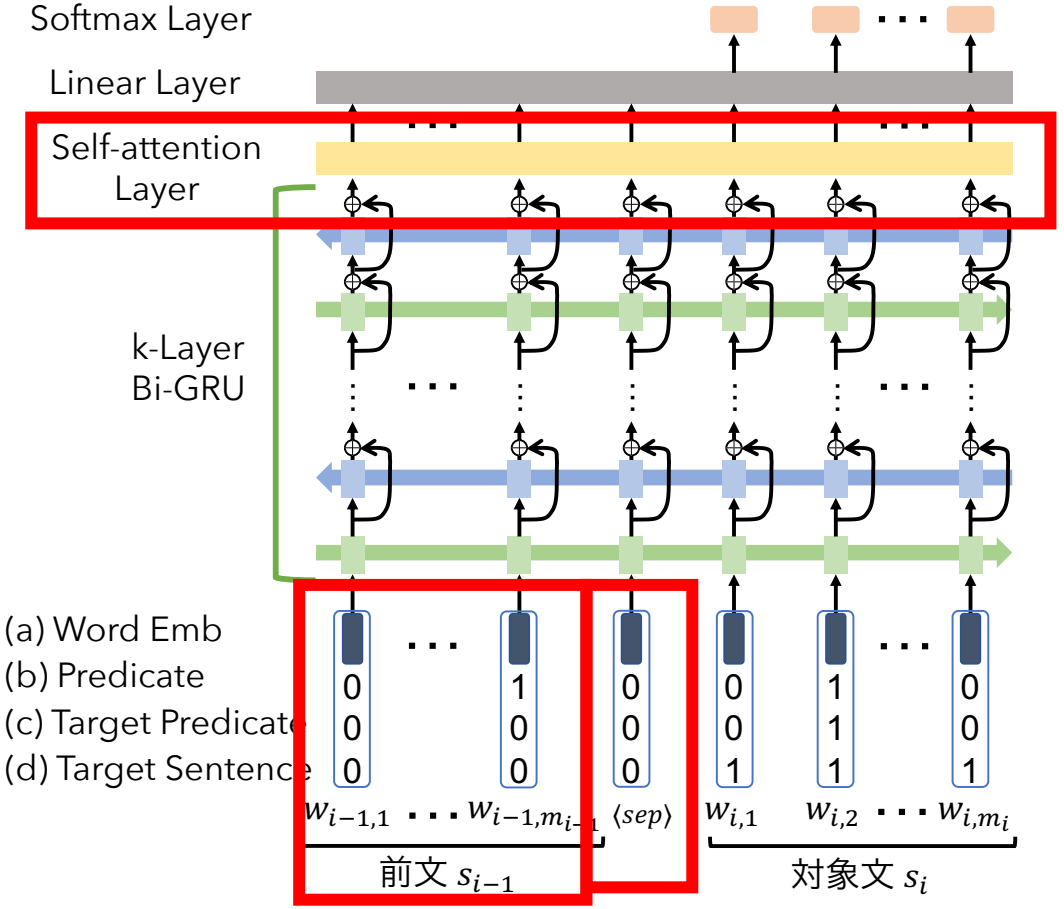
- 提案手法
  - これらの関係をモデルが陽に解かず、**前方のn文のすべての情報をRNNでエンコード**

# 提案手法

## ベースモデル [1]

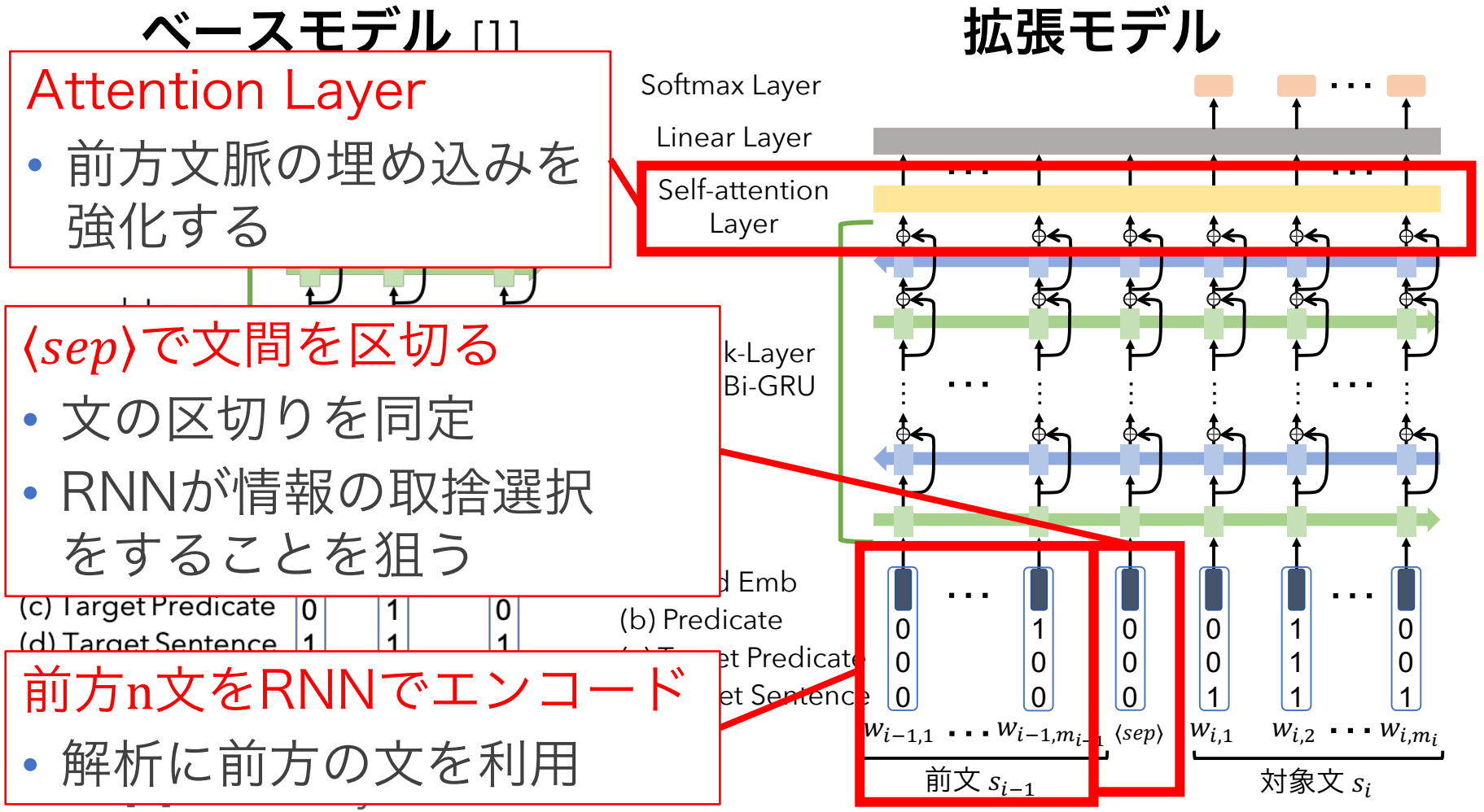



## 拡張モデル



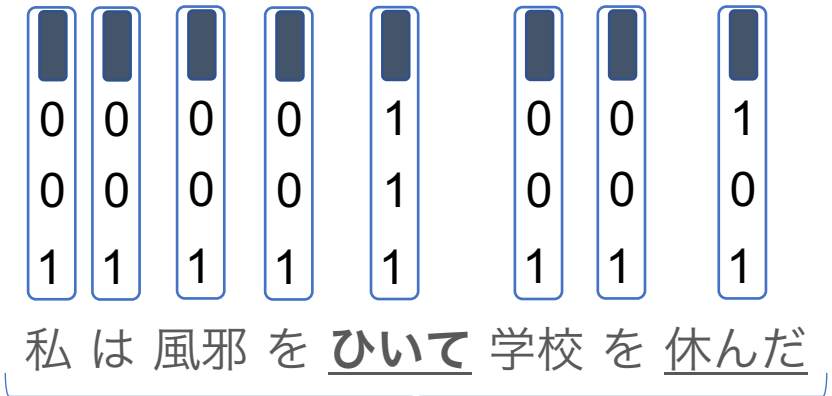

[1] Matsubayashi+'18

# 提案手法

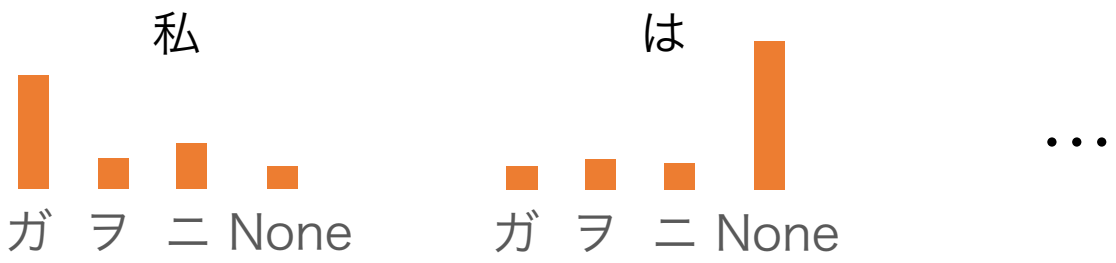

ベースモデル [1]
拡張モデル
Attention Layer
• 前方文脈の埋め込みを強化する
⟨sep⟩で文間を区切る
• 文の区切りを同定
• RNNが情報の取捨選択をすることを狙う
前方n文をRNNでエンコード
• 解析に前方の文を利用
Softmax Layer
Linear Layer
Self-attention Layer
k-Layer Bi-GRU
d Emb
(b) Predicate
et Predicate
et Sentence
(c) Target Predicate
(d) Target Sentence
$w_{i-1,1}$ ・・・ $w_{i-1,m_{i-1}}$ ⟨sep⟩ $w_{i,1}$ $w_{i,2}$ ・・・ $w_{i,m_i}$
前文 $s_{i-1}$
対象文 $s_i$

# 提案手法

- 入力:
  - Word Embedding
  - 述語の位置
  - 対象となる述語の位置
  - 対象文中の単語かどうか




- 出力:
  - 各単語に対する「ガ格」「ヲ格」「ニ格」「None」の確率分布



# 実験設定

- データセット
  - NAIST Text Corpus 1.5
  - 計 4万文 -> Train、Dev、Testに分ける
- 評価方法
  - 適合率、再現率、$F_1$値を求める
- 学習方法
  - Dev dataset の $F_1$ が5回改善されない場合に学習を終了
- Embedding
  - 日本語Wikipedia2016年9月1日のダンプデータより得られたEmbeddingを初期値として使用

# 実験結果

- 各拡張要素の効果

| Model | ALL | DEP | ZERO |
|---|---|---|---|
| BASE | 81.89 | 88.86 | 50.48 |
| BASE + 前1文 | **82.21** | 89.21 | **51.38** |
| BASE + 前2文 | 82.17 | 89.25 | 50.30 |
| ATTN | 81.47 | 88.60 | 49.51 |
| ATTN + 前1文 | 81.94 | **89.62** | 50.25 |
| ATTN + 前2文 | **82.21** | 89.32 | 51.12 |

# 実験結果

- 各拡張要素の効果

| Model | ALL | DEP | ZERO |
|---|---|---|---|
| BASE | 81.89 | 88.86 | 50.48 |
| BASE + 前1文 | **82.21** | 89.21 | **51.38** |
| BASE + 前2文 | 82.17 | 89.25 | 50.30 |

■**前方1文を追加**
- DEP、ZERO共に精度向上

■**前方2文を追加**
- DEPは精度向上
- ZEROは精度の向上は見られず
  - シーケンスが長い？

# 実験結果

- 各拡張要素の効果

| Model | ALL | DEP | ZERO |
|---|---|---|---|
| BASE | 81.89 | 88.86 | 50.48 |
| BASE + 前1文 | **82.21** | 89.21 | **51.38** |
| BASE + 前2文 | 82.17 | 89.25 | 50.30 |
| ATTN | 81.47 | 88.60 | 49.51 |
| ATTN + 前1文 | 81.94 | **89.62** | 50.25 |
| ATTN + 前2文 | **82.21** | 89.32 | 51.12 |

**ATTN+前2文**
全体的に精度向上

**ATTN、ATTN+前1文**
精度の向上は見られず

# +前1文で改善されたゼロ事例

対象の述語の項が、前方1文においても述語の項となっている

- 事例1 (ガ格)
  - 前方1文：核兵器を造り出したのは科学者だった。
  - 対象文：今度は知識と知恵を結集して核兵器の廃絶を目指すのが科学者の社会的責任であろう。
- 事例2 (ヲ格)
  - 前方1文：役に立つ白書を目指すならば、タイミングが必要ではないか。
  - 対象文：法務省・法務総合研究所が作成するこの白書は、わが国の犯罪研究に欠かせないばかりか、英訳され国際的にも高い評価を受けてきた。

# +前1文で改善されなかったゼロ事例

赤：予測
青：正解

前方1文に出現した項に
引っ張られ、間違えて予測

- 事例1 (ガ格)
  - 前方1文：広告会社社長は、同社から約八億三千万円を受け取ったという。
  - 対象文：中川容疑者は、同社が優良企業であるように装うため、事業報告書など投資家に配布する書類を偽造。
- 事例2 (ガ格)
  - 前方1文：目標が大きかったぶん魁皇も思い切って飛び込めたのだろう。
  - 対象文：武蔵丸も負けじと踏み込んで魁皇の当たりを止め、一気に黒房下に走った。

# 本研究のまとめ

- モデル
  - 多層双方向RNNを用いた既存の日本語述語項構造解析モデルを拡張し、**入力に前方のn文を加え**て解析を行うモデルを構築
- 結果
  - **前方1文**を追加したモデルで**ゼロ照応の精度向上**
  - **Attention Layer**を追加した拡張モデルは**シーケンスが長いほど精度が向上**したが、ゼロ照応においてはBASE+前方1文モデルの方が精度が良かった
- 今後
  - 文脈情報をよりよく取捨選択し、取り込むことができる機構を検討したい