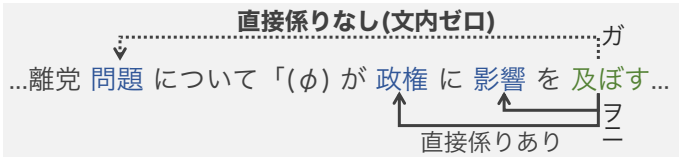
BERTによる擬似訓練データ生成に基づく述語項構造解析

今野颯人[1*◯], 松林優一郎[1,2], 清野舜[2,1], 高橋諒[1,2], 大内啓樹[2,1], 乾健太郎[1,2]
* ryuto@ecei.tohoku.ac.jp 1. 東北大学 2. 理化学研究所

1. 概要

述語項構造解析：述語とその項の間の関係を解析する



文内ゼロのF_1値 58%[1]

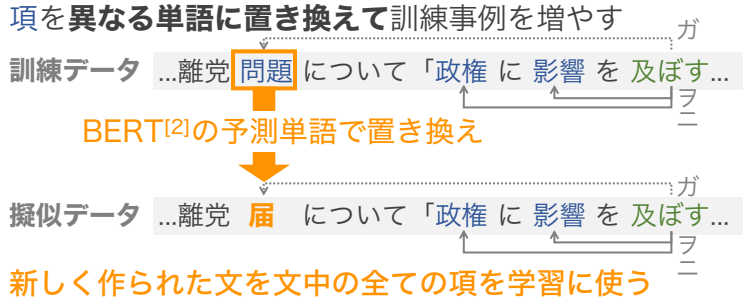
ゼロ照応解析の精度向上が主要な課題

仮説 **訓練事例数が足りていない**ため精度が低い
- ゼロの事例数は約1.6万 (全体の20%程度)

提案 **擬似訓練データ生成**による事例数の増加
- BERTにより自然な文を生成
- 項のバリエーションを増やす

2. 擬似訓練データ生成

項を**異なる単語に置き換えて**訓練事例を増やす



新しく作られた文を文中の全ての項を学習に使う

best	文中の項を1つずつBERTの予測確率が最も高い単語に置き換える
sampling	文中の項を1つずつBERTの予測した確率分布からサンプリングした単語に置き換える
sampling-multi	文中の全ての項を同時にsamplingで置き換える．これをn事例作る (n=5)

3. 実験

学習方法

mix	擬似訓練データと真の訓練データを混ぜる
pretrain	疑似訓練データで事前学習後，真の訓練データでfine-tune

実験設定

モデル	Matsubayashi+'18[1] ベースモデル (10層 Bi-GRU)
seed	3種類
学習率	{0.001, 0.0005, 0.0001}から探索
その他ハイパラ	Matsubayashi+'18[1] に準拠

3.1 擬似訓練データ生成による事例数の増加

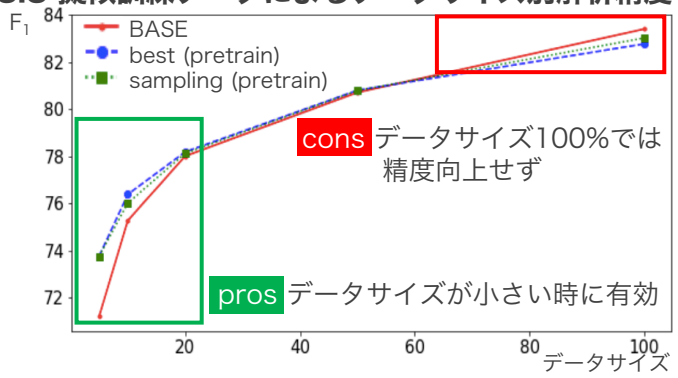
	ALL	DEP	ZERO
BASE	83,941	67,246	16,695
best	346,266	266,382	79,884
sampling	346,266	266,382	79,884
sampling-multi	419,705	336,230	83,475

約４倍増大

3.2 擬似訓練データによる解析精度

BASE F_1：83.48 ＞ ベストスコア F_1：83.13 精度向上せず

3.3 擬似訓練データによるデータサイズ別解析精度



		データサイズ5%			データサイズ100%		
		F_1	DEP	ZERO	F_1	DEP	ZERO
	BASE	70.78	79.21	37.76	**83.48**	**90.22**	**54.71**
mix	best	73.41	81.67	38.77	82.89	89.67	53.56
	samp	73.70	**81.98**	39.56	82.85	89.68	53.40
	samp-multi	**73.71**	81.95	38.63	82.16	89.17	50.95
pretrain	best	73.32	81.46	**39.74**	83.13	89.86	54.00
	samp	73.55	81.74	39.50	83.09	89.82	53.79
	samp-multi	72.53	80.88	38.15	82.93	89.78	52.78

4. 考察

Q. データサイズが上がるにつれて精度のゲインが小さくなるのはなぜか

A. データサイズが上がるにつれて，擬似訓練データによる**未知の項**に対するカバーが少なくなるため

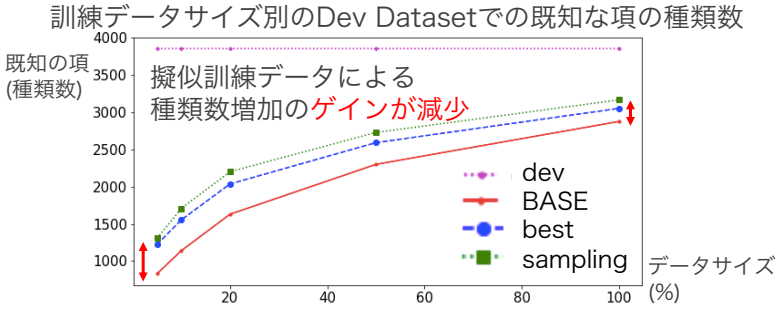
Dev Datasetにおける未知の項の解析精度

	F_1	DEP	DEP事例数	ZERO	ZERO事例数
BASE	82.23	88.36	13880	53.26	3391
BASE (既知の項)	82.68	88.95	12671	53.68	3157
BASE (未知の項)	77.29	82.15	1209	47.38	234

未知の項に対する解析精度は著しく低い

未知の項：開発データに出現するが訓練データに出現しない項

訓練データサイズ別のDev Datasetでの既知な項の種類数



5. Future work

未知語をカバーしたい
- トランスダクティブ学習
  - 評価データでBERTをfine-tune

項の置き換えでは増える情報量が少ない
- 述語と項を同時に置き換える
- chunk全体をBERTに予測させ，置き換える
- 生文から擬似データを生成
  - 直接係り受けありの解析結果からゼロ事例を作成

[1] Yuichiroh Matsubayashi, Kentaro Inui:Distance-Free Modeling of Multi-Predicate Interactions in End-to-End Japanese Predicate-Argument Structure Analysis. COLING 2018
[2] 柴田 知秀, 河原 大輔, 黒橋 禎夫: BERTによる日本語構文解析の精度向上, 言語処理学会 第25回年次大会, pp.205-208, 名古屋, (2019.3).