

An Empirical Study of Contextual Data Augmentation for Japanese Zero Anaphora Resolution

Ryuto Konno¹, Yuichiroh Matsubayashi^{1,2}, Shun Kiyono^{2,1}, Hiroki Ouchi², Ryo Takahashi^{1,2}, Kentaro Inui^{1,2} ¹Tohoku University ²RIKEN

Summary

- We proposed data augmentation (DA) for zero anaphora resolution (ZAR)
- We augmented labeled data by replacing tokens using language model (LM)
- We improved the performance of ZAR and analyzed the phenomenon in DA

Task: Zero Anaphora Resolution (ZAR)

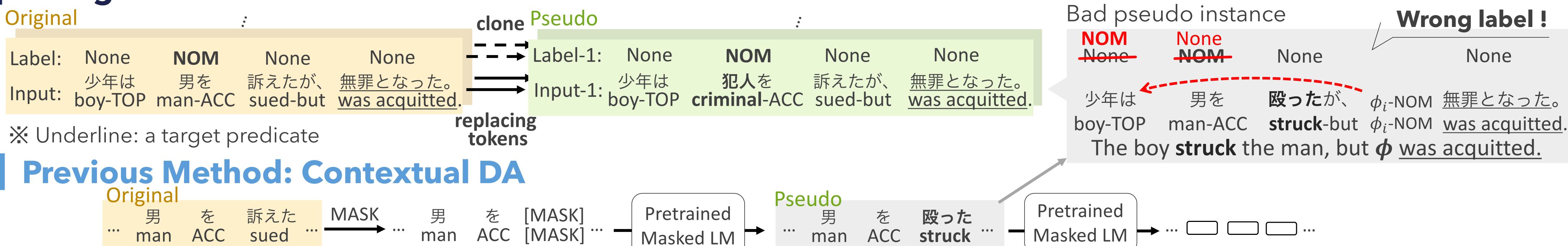
訴えたが、 ϕ_i -NOM 無罪となった。 少年は boy-TOP man-ACC sued-but ϕ_i -NOM was acquitted. The boy sued the man, but ϕ was acquitted.

Automatic recognition of omitted arguments of a predicate

A masked LM has to be **run twice**

Halving the computational cost

Background



Proposed Method: MASKING

1 MASKING improved the performance

Table1: F1 score

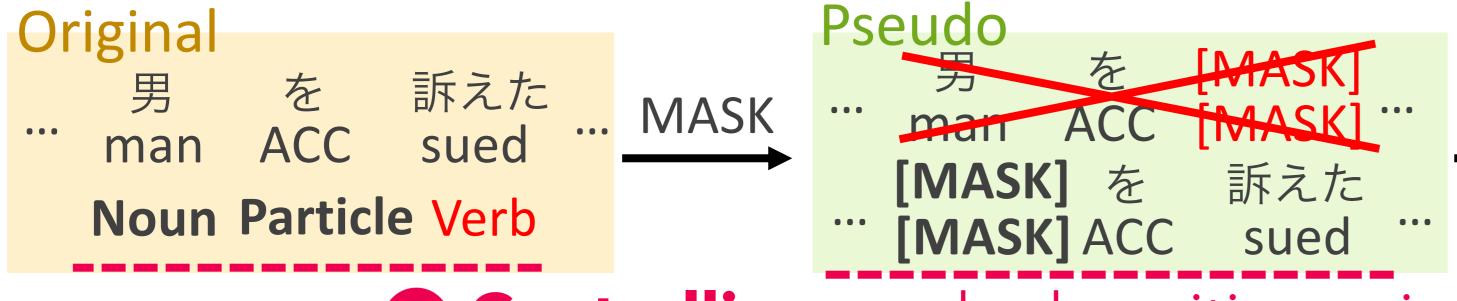
Matsubayashi&Inui'18

MASKING (all-but-verb)

Model

BASELINE

Contextual DA



Controlling masked position using POS tags

Results and analysis

Cannot control masked position

(2) Masking all POS categories except for verb is the best score

of ZAR on test set					Table2: F
	ALL	NOM	ACC	DAT	Model (Ma
	55.55	57.99	48.9	23	BASELINE
	63.89	66.45	57.2	27	All POS
	63.87	66.16	58.5	29	Only verb
	64.15	66.60	57.9	29	All POS exce

F1 score on dev set asking target) ZAR 64.08 64.89 64.15 cept for verb 65.02

(3) Masking verb may produce the bad instance

(one-pass running)

Pretrained

Masked LM

キズ゛は 直り ません が、 ϕ -NOM ほとんど <u>目立た</u> なくなって、… scratches-TOP be-fixed not but, ϕ -NOM almost noticeable not become,... The scratches can't be repaired, but ϕ are becoming barely <u>noticeable</u> ... null キズ は あり ませんが、 ϕ -NOM ほとんど <u>目立た</u> なくなって、… scratches-TOP exist not but, ϕ -NOM almost noticeable not become,... There are no scratches, but ϕ doesn't stand out anymore ...