

Proc. ACL Volume 1: Long Papers, pages 7052–7087

# When and how to paraphrase for named entity recognition?

Saket Sharma<sup>¶</sup>, Aviral Joshi<sup>¶</sup>, Yiyun Zhao<sup>¶</sup>, Namrata Mukhija<sup>¶</sup>,  
Hanoz Bhathena<sup>¶</sup>, Prateek Singh<sup>¶</sup>, Sashank Santhanam<sup>||</sup>

<sup>¶</sup> Machine Learning Center of Excellence, JPMorgan Chase & Co.

<sup>||</sup> Apple

【一言まとめ】  
生成モデル (GPT3) で  
データ拡張するといいい感じ

紹介者：駒水 孝裕

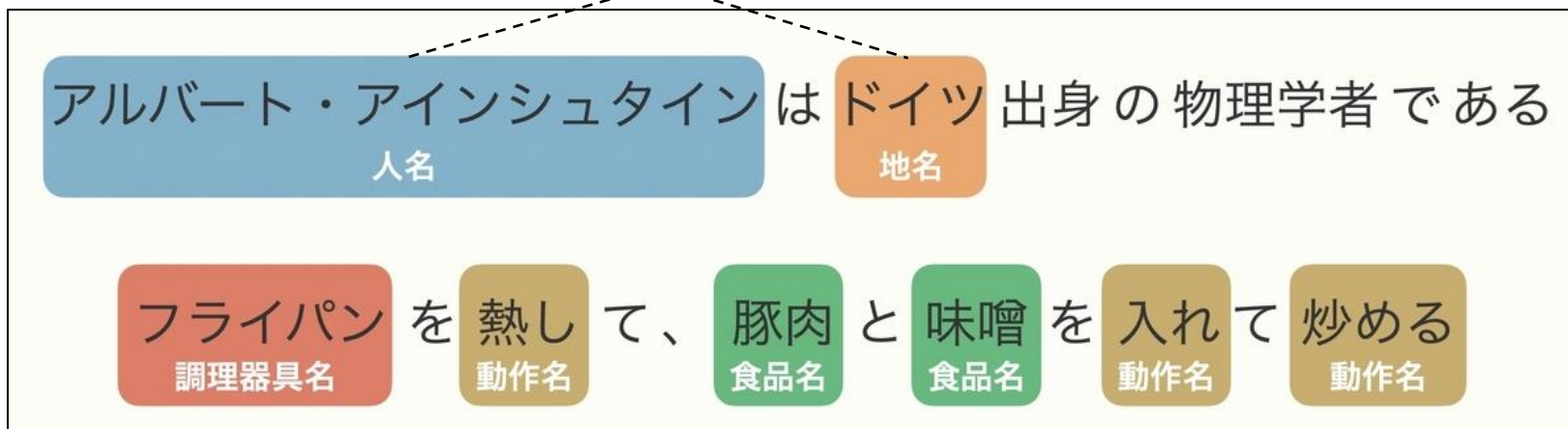


# NER (Name Entity Recognition; 固有表現認識)

2

- 文書から「固有表現 (Entity)」を検出・認識するタスク

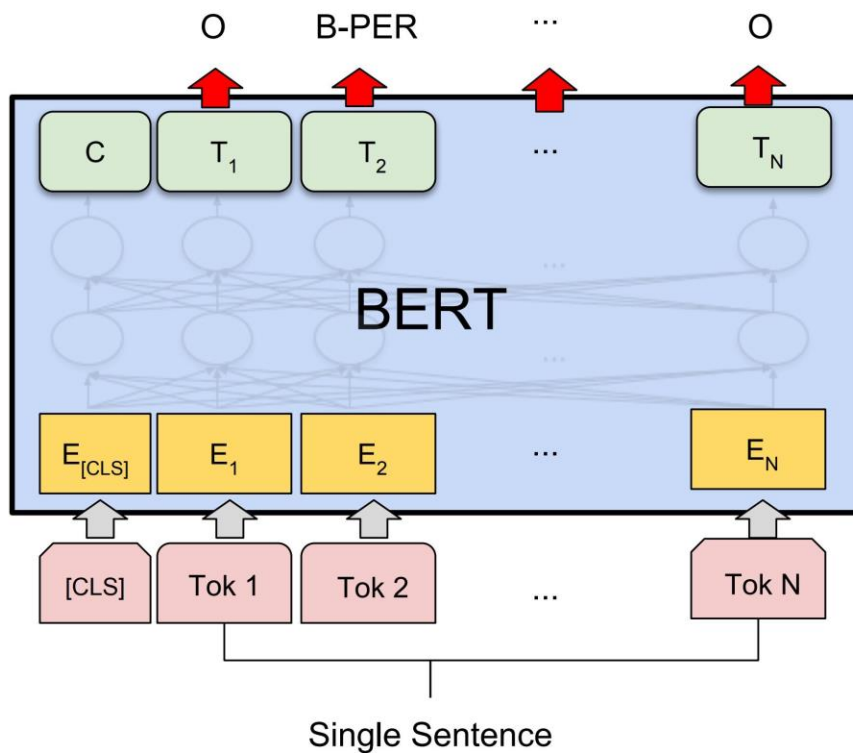
メンション (mention)



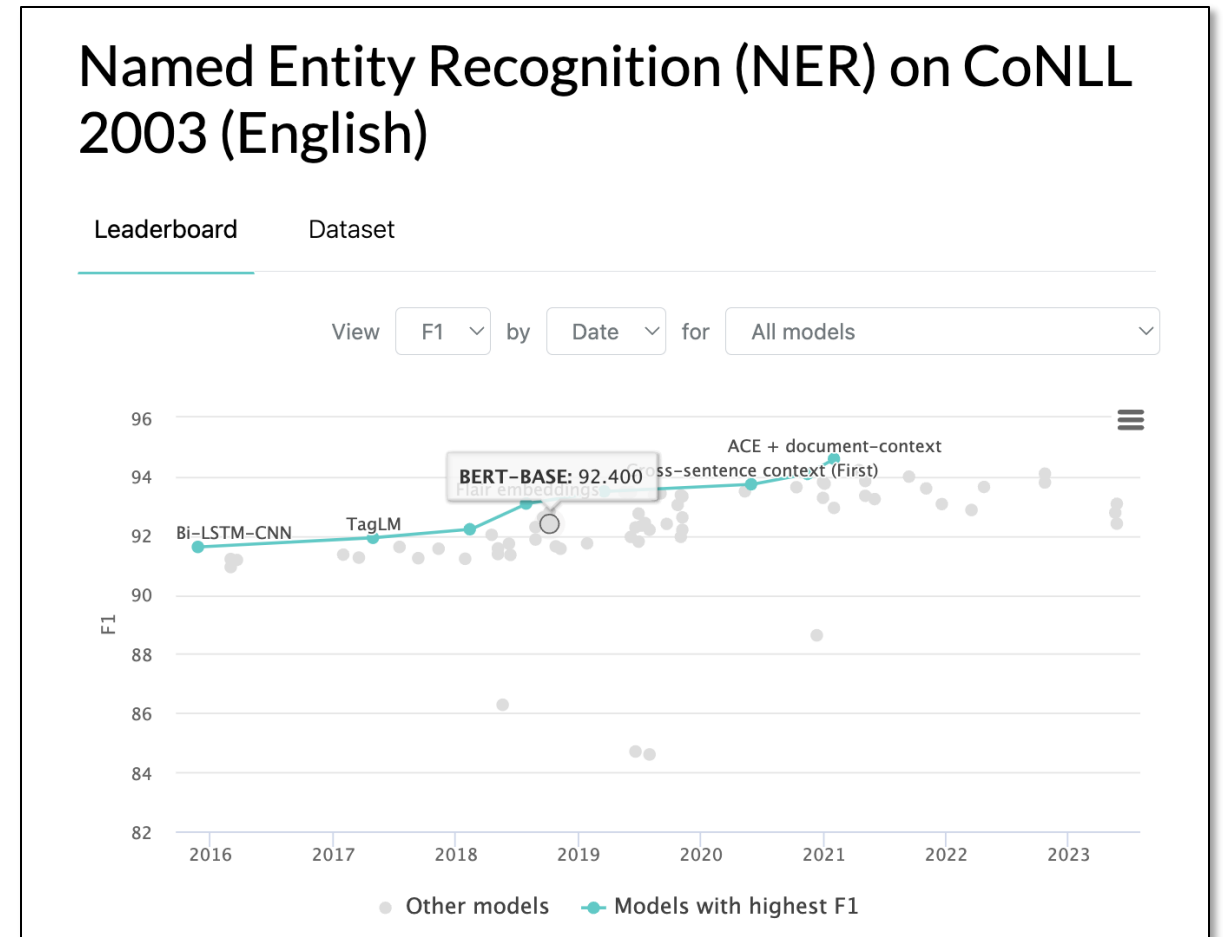
<https://techlife.cookpad.com/entry/2020/11/06/110000>

- 下流タスクのための事前処理
  - 検索, 推薦, 対話システム, etc.

- BERT + Linear layer
  - 文をトークナイズ
  - BERT に入力しエンコード
  - トークン単位でラベル推定



## Named Entity Recognition (NER) on CoNLL 2003 (English)



<https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003>

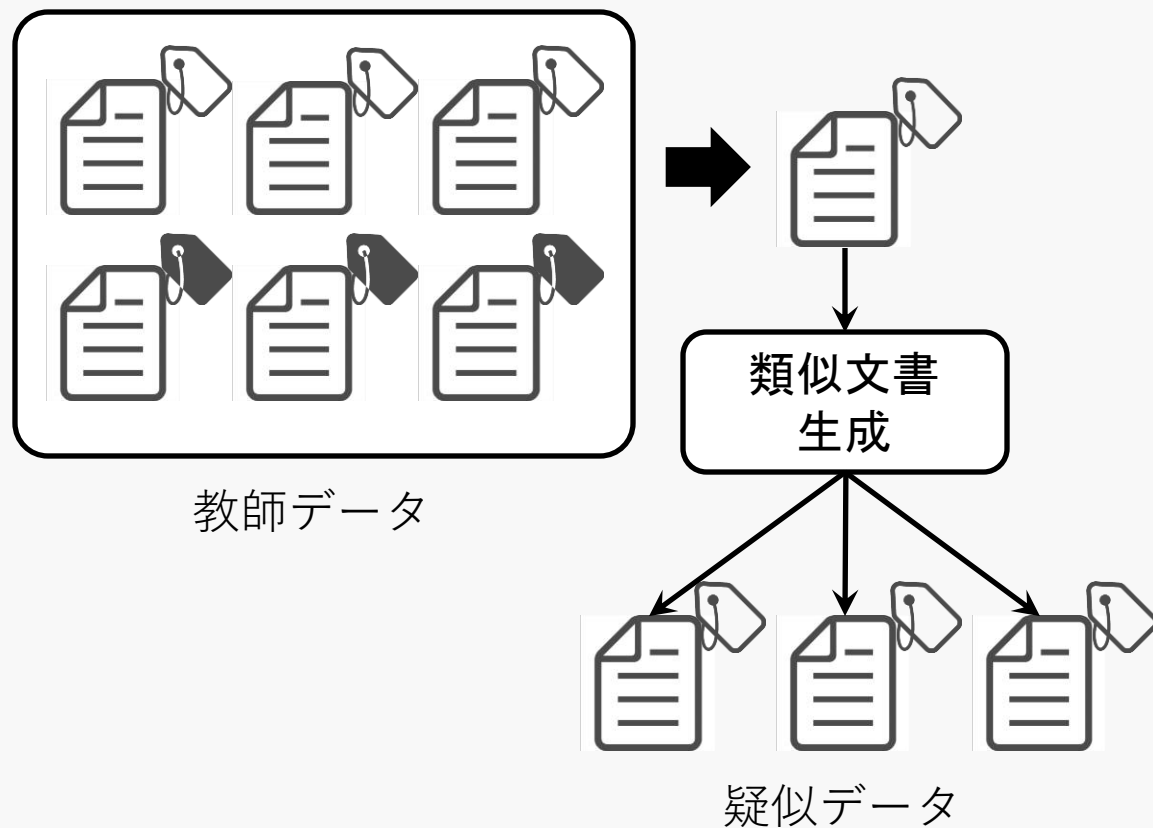
# 問題：データ不足

- 一般的である固有表現 → データ十分 → 十分な性能
  - クラウドソーシングや自動アノテーションが可能
- 一般的でない固有表現 → データ不足 → 不十分な性能
  - データ不足：アノテーションのコスト，専門知識の必要性
  - 不十分な性能：学習データのバリエーション不足
    - 特に，文脈のバリエーションが不十分

## データ拡張（Data Augmentation）で対応

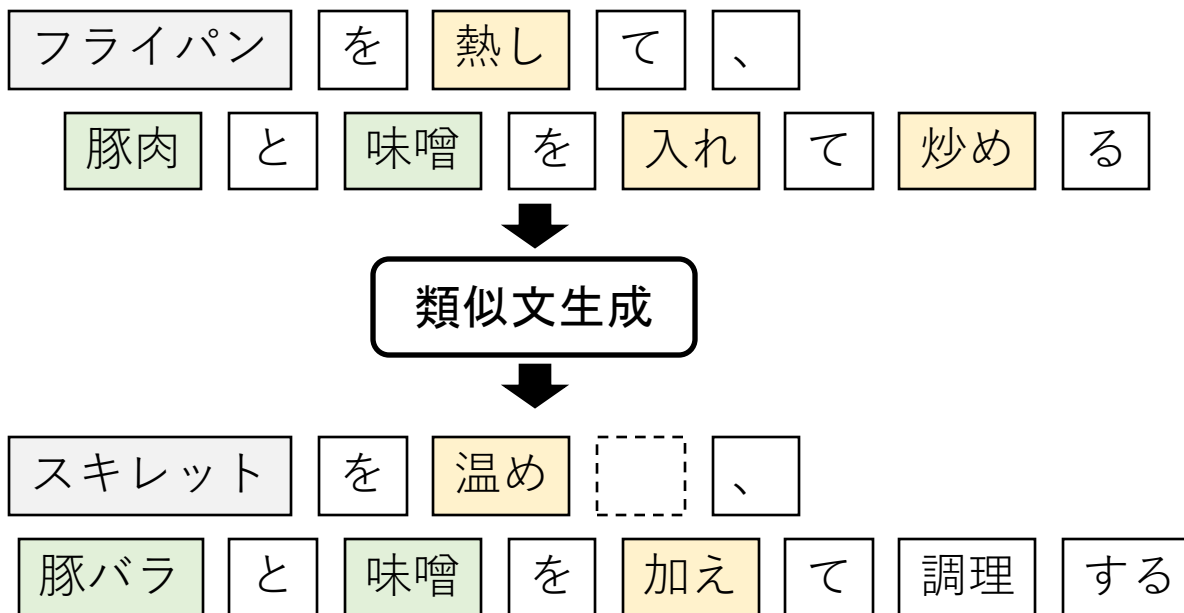


## 文書/文レベルタスクのデータ拡張



## NER のデータ拡張

- トークン単位のラベル

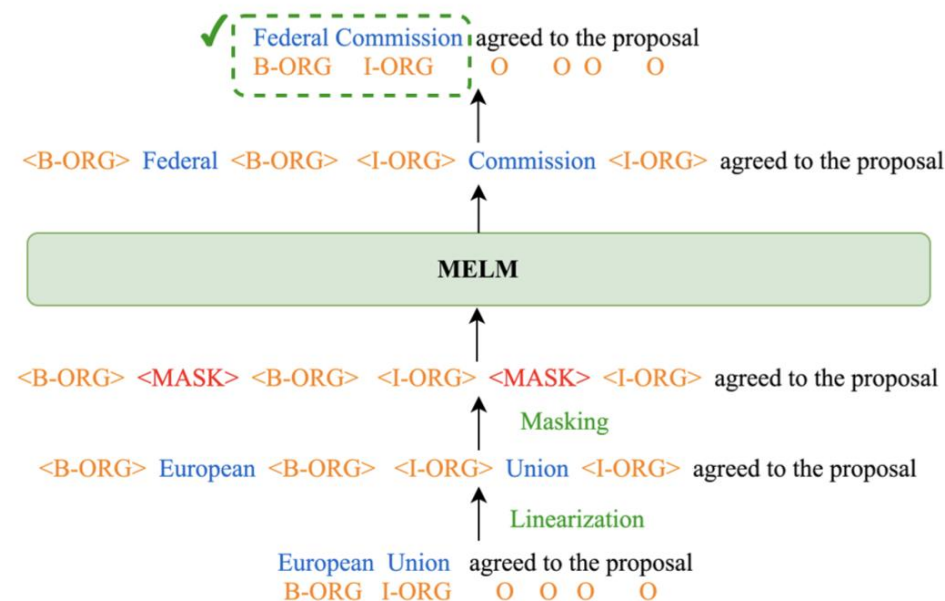


- 難しさ：疑似データの正確なラベル付け
  - 文書レベルよりも繊細
    - 文長の変化，トークンの妥当性

- Local Replacement : e.g., (Dai and Adel, COLING'20)
  - 学習データ内の同ラベルのトークンで置き換え (LwTR)
  - 課題 : メンションの整合性, 意味の通らない文脈

元文	She	did	not complain of	headache	or	any	other	neurological symptoms	.
	O	O	O O O	B-problem	O	B-problem	I-problem	I-problem	O
拡張	L.	One	not complain of	headache	he	any	interatrial	neurological	current
	O	O	O O O	B-problem	O	B-problem	I-problem	I-problem	O

- 置き換え生成 : e.g., (Zhou et al., ACL'22)
  - [MASK] 推定で置き換えトークンを生成
  - Linearization : 固有表現をラベルで挟み込み一文に変換
  - 課題 : 学習データがさらに必要



## 言い換え (paraphrasing) を用いたデータ拡張の可能性検証

- 言い換え生成：バリエーションに富んだ文が生成可能
- 言い換え生成モデルの選択が NER の性能に与える影響の調査
- データ拡張におけるメンション記憶問題に対するメンション置き換えの提案

## この研究で明らかになったこと

- LLM (GPT-3 Davinci) + inline annotation が経験則ベースの手法よりも良い性能
- (低頻度の固有表現に効果的)
- LLM (GPT-3 Davinci) がメンション記憶問題に頑健
  - メンション置き換えを導入するとさらに改善

- データセット (評価: F1-score)
  - BioCreative V CDR (BC5CDR) (Wei et al., 2016)
  - Ontonotes5 (Hovy et al., 2006)
  - MIT Restaurant NER dataset (MIT-R) (Liu et al., 2013)
  - Tweebank (Jiang et al., 2022)
  - WNUT 2017 (Derczynski et al., 2017)
- 言い換えモデル
  - 逆翻訳 (Back Translation; BT) : BART (Lewis et al., 2019)
    - 言語: 英語とドイツ語
  - 文書要約: Pegasus (Zhang et al., 2020)
    - 言い換え用にファインチューニングされたものを利用
  - LLM (GPT-3; OpenAI API): Ada (~350M parameters), Davinci (~175B parameters)
    - プロンプト: 2種類 (後述)

	Train	Dev	Test
BC5CDR	5,228	5330	5,865
Ontonotes	59,924	8,528	8262
MIT-R	6,900	760	1,521
Tweebank	1,639	710	1,201
WNUT-17	2,394	1,009	1,287

NER モデル:  
Distil BERT + Linear Layer



## Prompt A (言い換え生成)

Create a paraphrase for inputs like the following example:

Input: Japanese band The Altruists is releasing their hit single this fall.

Paraphrases:

1. The Altruists, a Japanese band is releasing their hit single this fall

## Prompt B (inline annotation)

Create a paraphrase for inputs like the following example. *Preserve the annotations in the [] and ():*

Input: Japanese band *[The Altruists](ORG)* is releasing their hit single this fall.

Paraphrases:

1. *[The Altruists](ORG)*, a Japanese band is releasing their hit single this fall

学習データの文を入力

Input: **BLANK**

Paraphrases:

- 1."

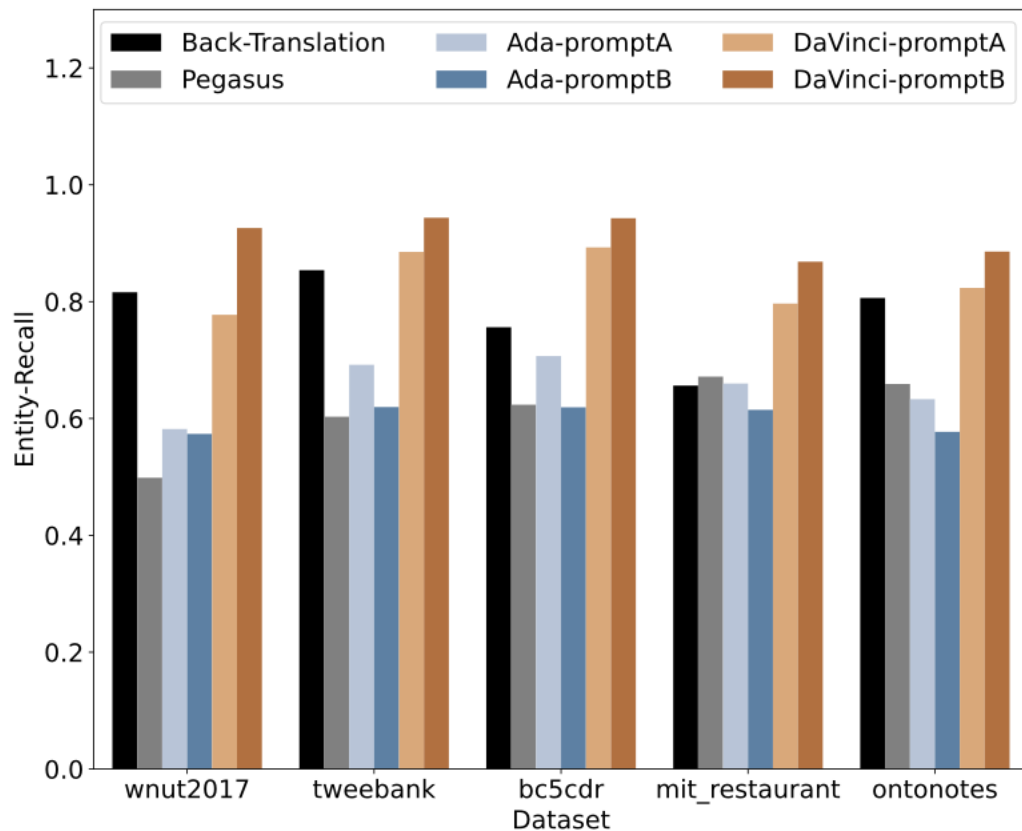
Input: **BLANK**

Paraphrases:

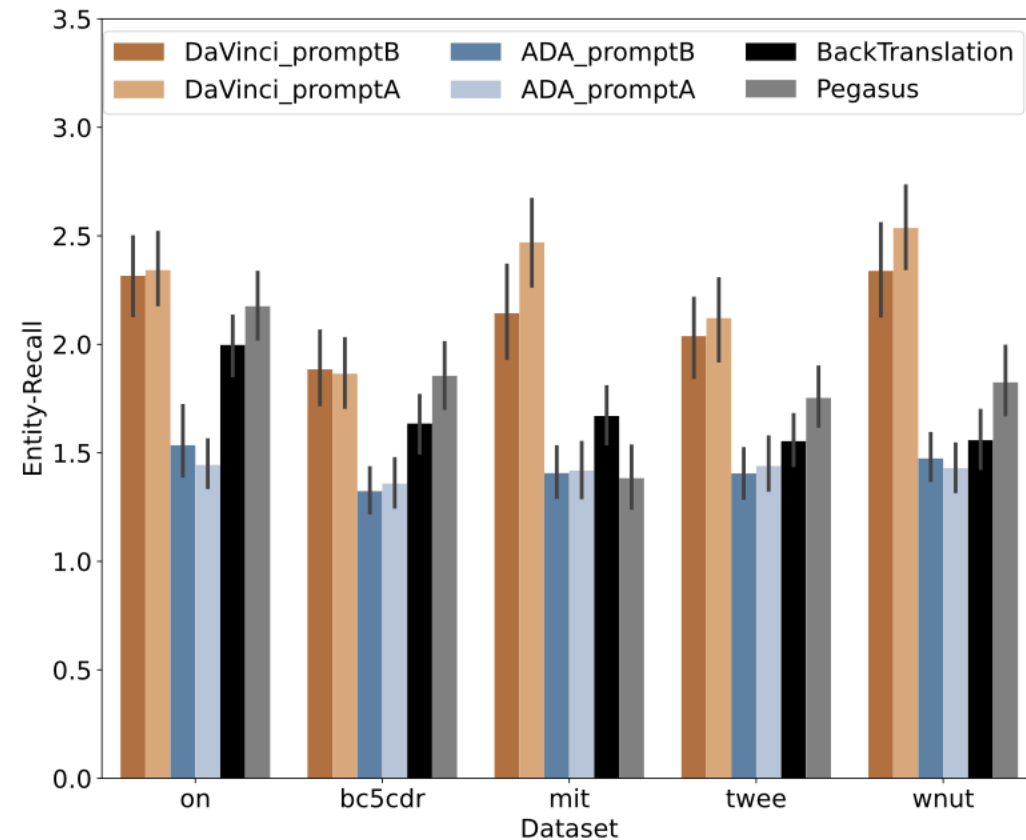
- 1.

- 完全一致するメンションにアノテーション付与
  - 言い換えモデルに与えた文中のメンションとの完全一致
  - inline annotation の場合, メンションが変化していないかをチェック
- フィルタリング
  - 15文字以下の文を削除
  - 重複削除
  - 固有表現のメンションがない文を削除
  - 生成された文のうち, 最初のものを採用
    - 酷似した文が生成されがちのため

## 固有表現の再現率



## 言い換えの人的評価



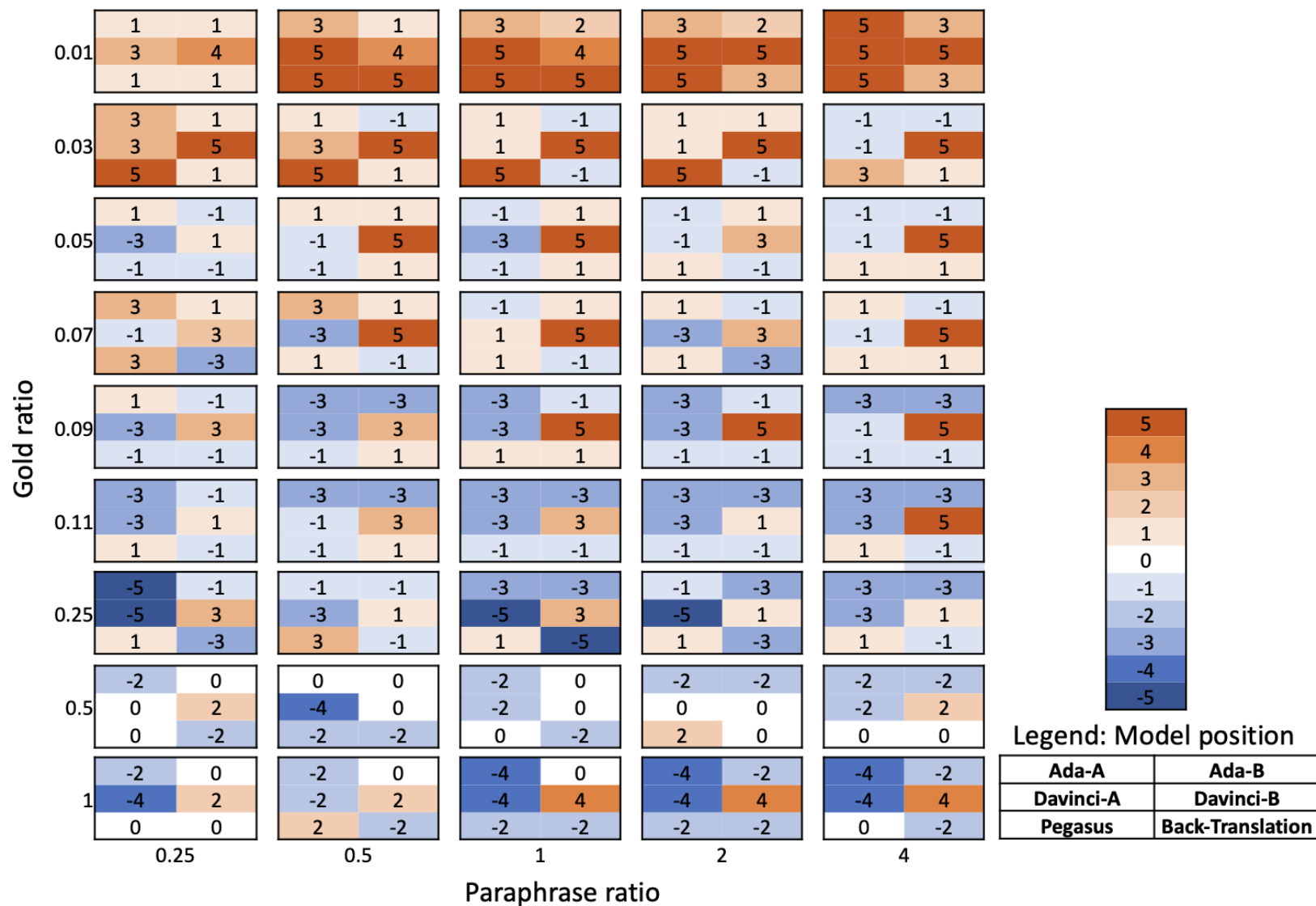
3人の5段階評価の平均  
(縦軸ラベルは間違い)

# 実験結果：拡張して良くなったかどうか

(図がスーパー見にくい)

- 各セル：各データセットで F1 が改善したかどうか (+1/-1) の総和
- Gold ratio：  
学習データの割合
- Paraphrase ratio：  
拡張データの割合
- Davinci + Prompt B が  
コンスタントに改善

Ada-A	Ada-B
Davinci-A	<b>Davinci-B</b>
Pegasus	Back-Translation



- メンションを維持したデータ拡張 → 同じメンションが異なる文脈で出現  
→ モデルがメンション自体を記憶してしまう問題
- 対策：メンション置き換え (Mention Replacement; MR)
  - LLM (GPT-3 Davinci) で類義メンションを生成  
(ENTITY\_TYPE と ENTITY\_VALUE は具体的な値で置き換え)

*Please list 10 examples of ENTITY\_TYPE such as 'ENTITY\_VALUE':  
1.*

- 未知の固有表現 (Unseen Entity; UE) だけで評価

	MIT-R	Onto -notes	BC5 -CDR	Twee -bank	Wnut -17
DaV-B	15	<b>35</b>	28	14	15
DaV-B +MR	<b>30</b>	0	17	<b>31</b>	<b>30</b>

## 言い換え (paraphrasing) を用いたデータ拡張の可能性検証

- 言い換え生成：バリエーションに富んだ文が生成可能
- 言い換え生成モデルの選択が NER の性能に与える影響の調査
- データ拡張におけるメンション記憶問題に対するメンション置き換えの提案

## この研究で明らかになったこと

- LLM (GPT-3 Davinci) + inline annotation が経験則ベースの手法よりも良い性能
- 低頻度の固有表現に効果的
- LLM (GPT-3 Davinci) がメンション記憶問題に頑健
  - メンション置き換えを導入するとさらに改善

- 除外：各コンポーネントの貢献を線形回帰モデルで分析

factor	$\hat{\beta}$	t	p
<b>P</b>	0.0148	9.004	1e-15
<b>G</b>	-0.0048	-1.393	0.164
<b>DaV-B</b>	0.0106	3.986	1e-15
<b>support</b>	-0.0031	-5.815	1e-15
cap	-0.0011	-0.473	0.636
<b>number</b>	-0.0154	-5.031	1e-15
P:G	-0.0032	-1.937	0.053
<b>P:DaV-B</b>	0.0078	6.121	1e-15
<b>P:support</b>	-0.0028	-10.925	1e-15
P:cap	0.0008	0.715	0.475
<b>P:number</b>	-0.0165	-11.095	1e-15

Table 4: Coefficients of the model with entity level analysis (See the Appendix A.8 for the coefficients of the full model)  $\Delta_{F1} \sim \text{Paraphrase} * (\text{Gold} + \text{model} + \text{support} + \text{capitalize} + \text{number})$ .

factor	$\hat{\beta}$	t	p
<b>P</b>	-0.01	-5.769	1e-15
<b>G</b>	0.0012	0.323	0.746
<b>DaV-B</b>	-0.0138	-4.937	1e-15
<b>support</b>	0.0041	7.257	1e-15
cap	-0.0018	-0.759	0.448
<b>number</b>	0.0162	5.028	1e-15
<b>P:G</b>	0.0045	2.543	0.011
<b>P:DaV-B</b>	-0.0113	-8.354	1e-15
<b>P:support</b>	0.0031	11.456	1e-15
<b>P:cap</b>	-0.0044	-3.874	1e-15
<b>P:number</b>	0.0142	9.062	1e-15

Table 5: Coefficients of the Linear model for Memorization:  $\text{memorization} \sim \text{Paraphrase} * (\text{Gold} + \text{model} + \text{support} + \text{capitalize} + \text{number})$ . Full results Appendix A.9.

- NER は古典的な問題だが依然として取り組まれている。  
以下 ACL 2023 で発表された論文の一部
  - Ghosh et al., ACLM: A Selective-Denoising based Generative Data Augmentation Approach for Low-Resource Complex NER
    - MLM に基づいてデータ拡張する新しい方法論
  - Shen et al., PromptNER: Prompt Locating and Typing for Named Entity Recognition
    - BERT への入力 (Prompt) を工夫して効率的に学習
    - 認識性能は限定的
  - Liu and Ritter, Do CoNLL-2003 Named Entity Taggers Still Work Well in 2023?
    - CoNLL-2003 で学習した NER tagger は使えるか? → Yes
- データ不足は一般的だが、トークン単位 (e.g., NER) のデータ作成は容易でない
  - LLM の言語生成能力は強力、どのように利用するかがポイント



Proc. ACL Volume 1: Long Papers, pages 7052–7087

# When and how to paraphrase for named entity recognition?

Saket Sharma<sup>¶</sup>, Aviral Joshi<sup>¶</sup>, Yiyun Zhao<sup>¶</sup>, Namrata Mukhija<sup>¶</sup>,  
Hanoz Bhathena<sup>¶</sup>, Prateek Singh<sup>¶</sup>, Sashank Santhanam<sup>||</sup>

<sup>¶</sup> Machine Learning Center of Excellence, JPMorgan Chase & Co.

<sup>||</sup> Apple

【一言まとめ】  
生成モデル (GPT3) で  
データ拡張するとうい感じ

# Question?

A magnifying glass with a black handle and a silver ring is positioned over an open book. The book is open to two pages of text, and a yellow pencil with a silver eraser and a black tip lies on the right page. The entire scene is set against a dark wooden surface.