国際会議 EMNLP2021 参加報告

東北大学 情報科学研究科 修士2年
小林 悟郎

2021年12月3日
情報処理学会 第251回 自然言語処理研究会

自己紹介

小林 悟郎 @goro_koba

- 東北大学 乾研究室 所属
- 研究の興味
  - 単語埋め込み・言語モデルの挙動や性質
- 主著論文が EMNLP 2021 に採択
  - 去年 (EMNLP 2020) に引き続き 2 年連続
  - NL研での参加報告も2度目

コロナ前は…

Incorporating Residual and Normalization Layers into Analysis of Masked Language Models

Goro Kobayashi[1] **Tatsuki Kuribayashi**[1,2] **Sho Yokoi**[1,3] **Kentaro Inui**[1,3]
[1] Tohoku University [2] Langsmith Inc. [3] RIKEN
goro.koba@dc.tohoku.ac.jp
{kuribayashi, yokoi, inui}@tohoku.ac.jp

本日の内容 (15分程度)

1. EMNLP2021 会議概要 (~4分)
   - 各種スタッツ
2. 会議の様子 (~4分)
   - 発表・聴講の様子
3. 論文紹介 (7分~)
   - 私の投稿論文を紹介 (宣伝)
   - 言語モデル周りで興味深かった論文をピックアップ

EMNLP2021 会議概要

EMNLP: ***E***mpirical ***M***ethods in ***N***atural ***L***anguage ***P***rocessing

- 自然言語処理分野の難関国際会議のひとつ
  - ACL, NAACL と共に分野の 1st Tier 会議として扱われる
- 初のハイブリッド開催
  - ドミニカ共和国 (~500人)
  - バーチャル (3000+人)

Google Scholar

Top publications

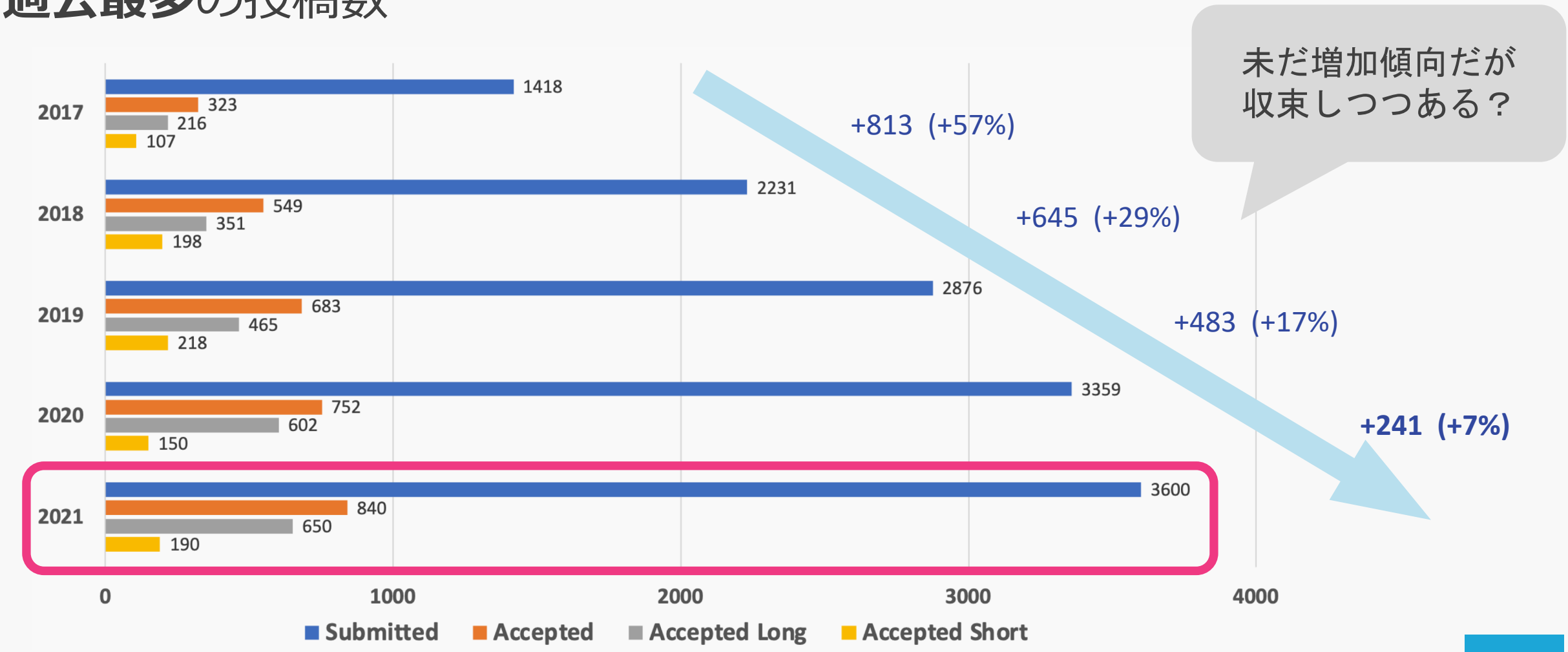
Categories > Engineering & Computer Science > **Computational Linguistics**

	Publication	h5-index	h5-median
1.	Meeting of the Association for Computational Linguistics (ACL)	157	265
2.	Conference on Empirical Methods in Natural Language Processing (EMNLP)	132	235
3.	Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)	105	195
4	International Conference on Computational Linguistics (COLING)	64	

https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computationallinguistics

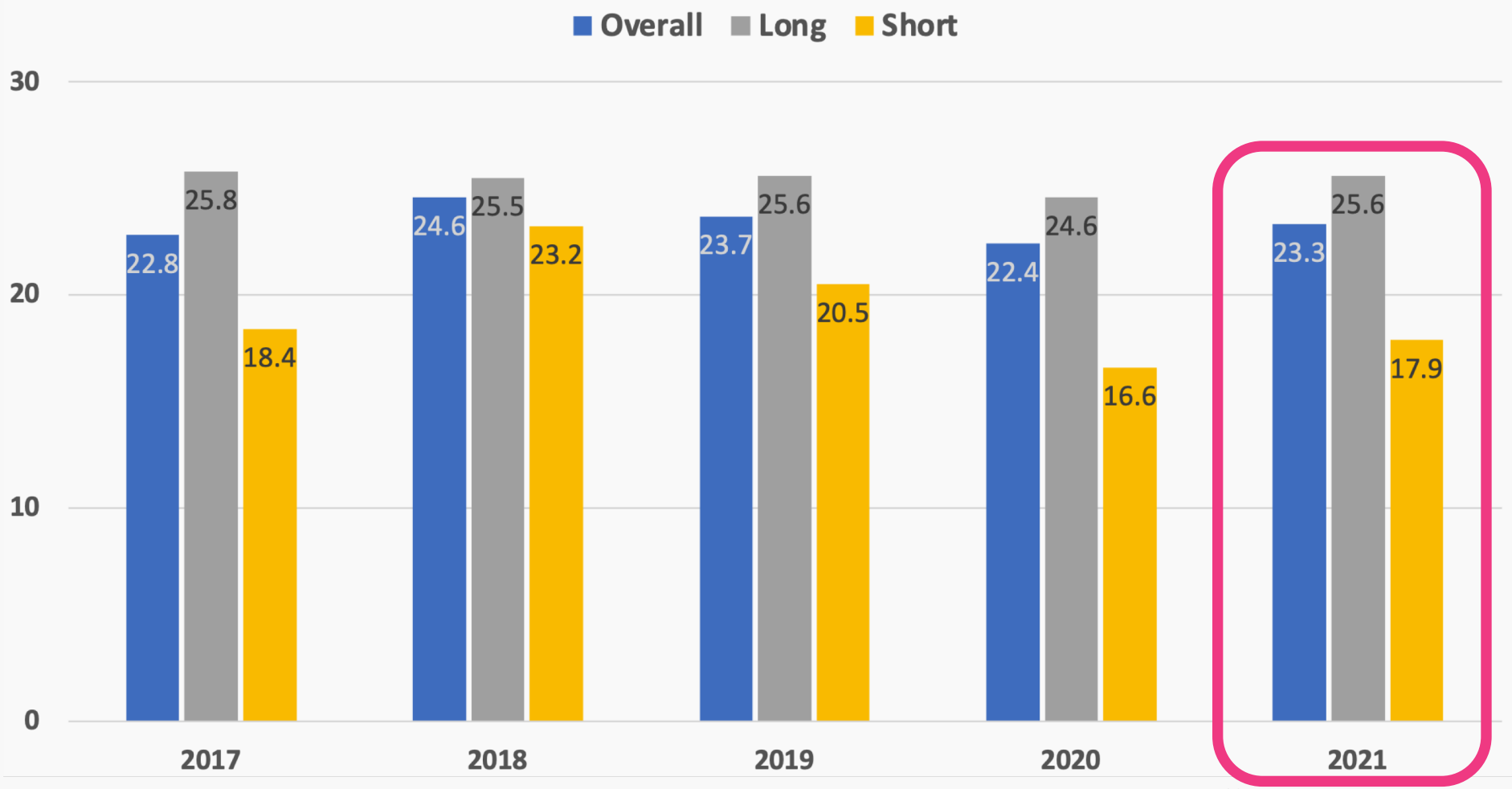
投稿数・採択数

過去最多の投稿数



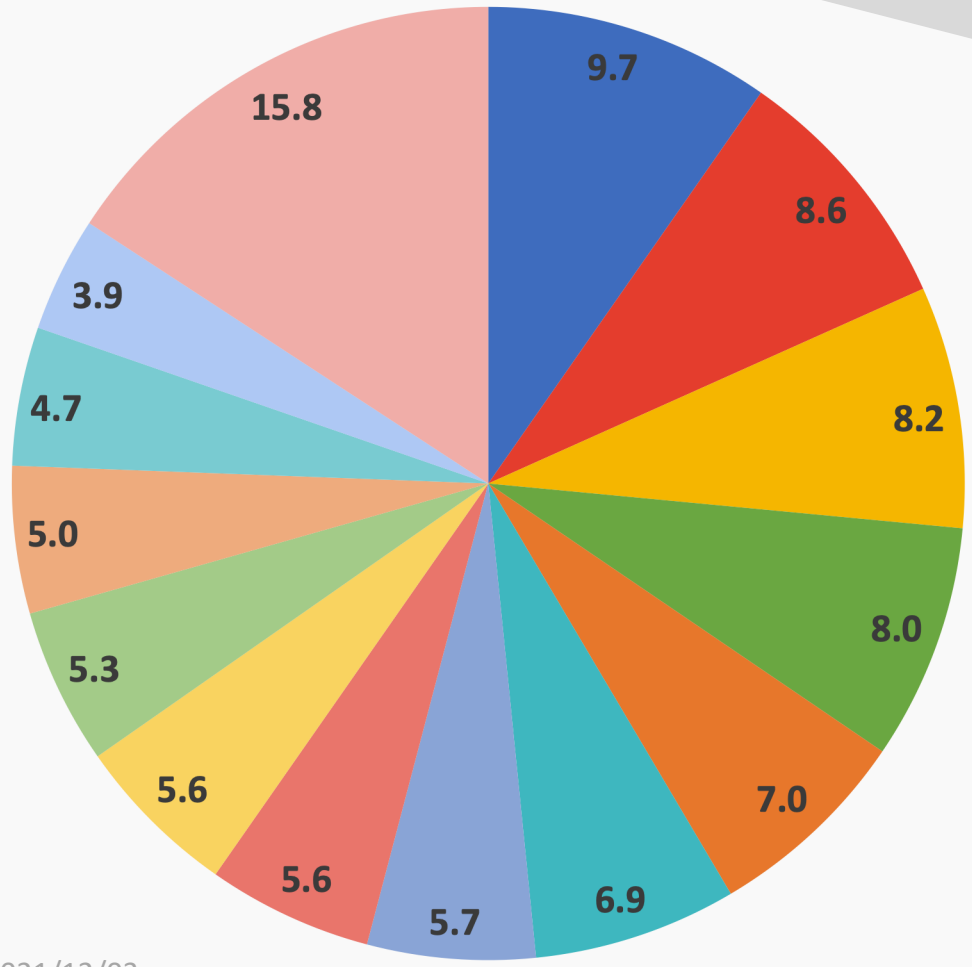
採択率

採択率は 22~25% で**キープ**



分野の傾向 ～トラック別投稿～

機械翻訳・情報抽出・対話・質問応答が人気



- NLP Applications
- Machine Learning for NLP
- Machine Translation and Multilinguality
- Information Extraction
- Dialogue and Interactive Systems
- Semantics: Lexical, Sentence level, Textual Inference and Other areas
- Interpretability and Analysis of Models for NLP
- Question Answering
- Resources and Evaluation
- Generation
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech, Vision, Robotics, Multimodal Grounding
- Summarization
- Others

言語モデル・Transformer 関連の話題を多く含む

私の論文の投稿先

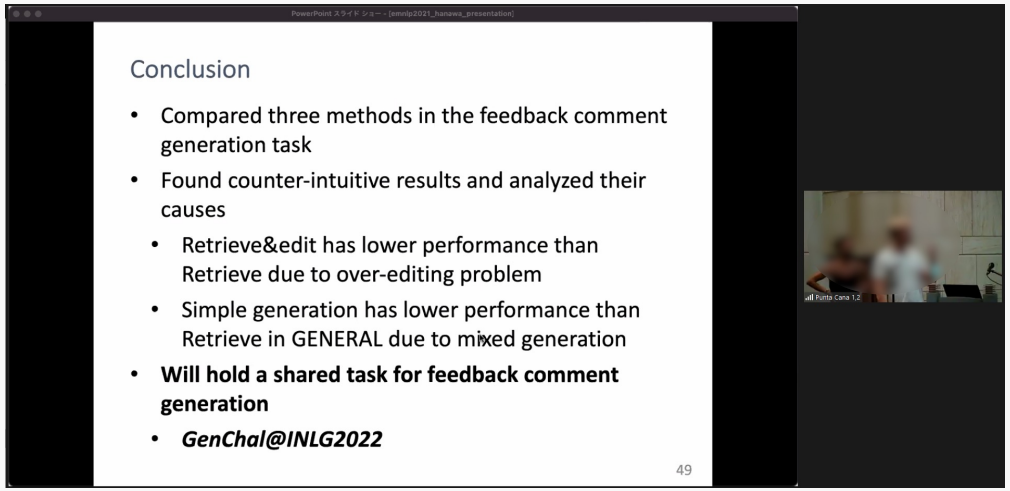
会議の様子

スケジュール

- 日本時間では **深夜～午前**(22:00~10:00) にイベントが集中
  - 現地タイムゾーンに合わせるハイブリッド開催の難しさ？

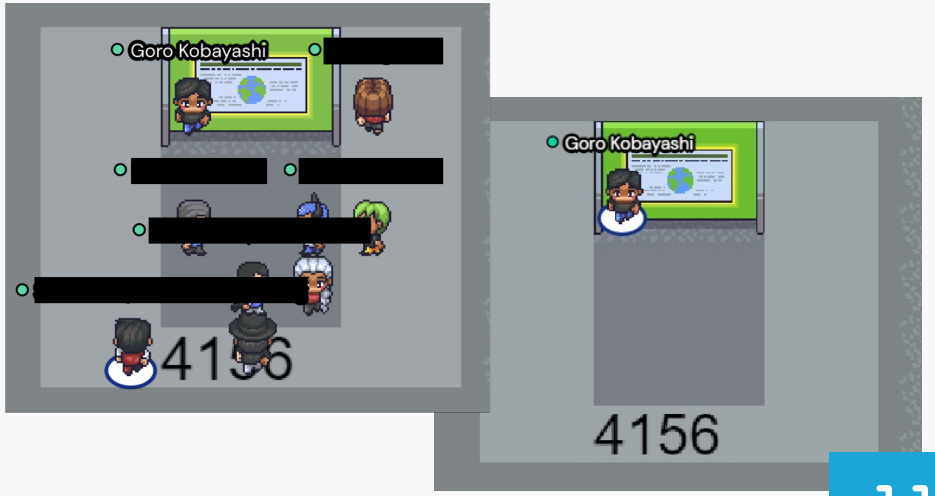
発表・聴講の様子

- オーラルセッション
  - 90分の Zoom セッション
  - 現地会場と連携



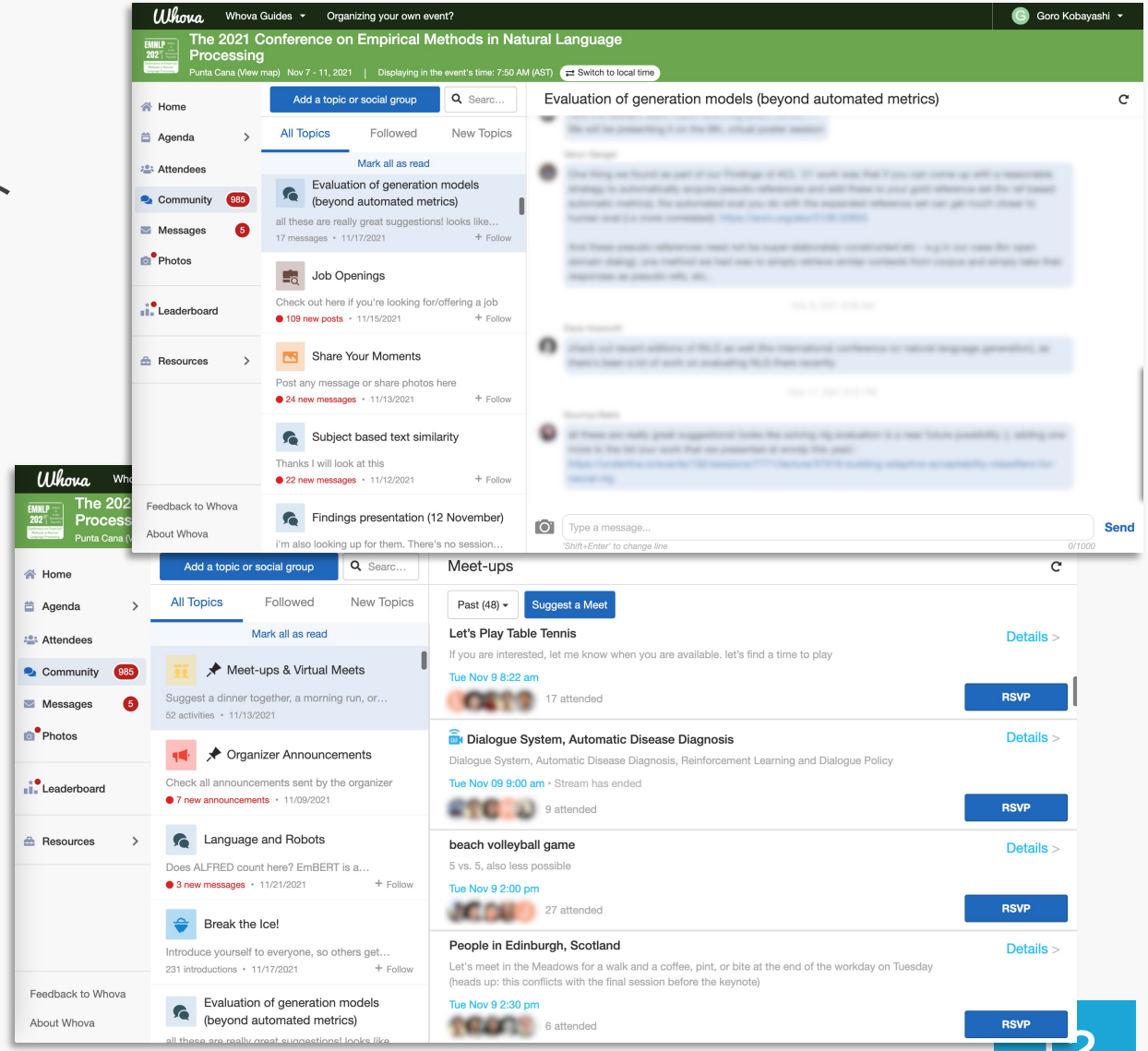
Hanawa+, Exploring Methods for Generating Feedback Comments for Writing Learning

- ポスターセッション
  - 2時間の Gather.town セッション
  - No show が続出… (半数以上という噂も)



交流

- 参加者間の交流用アプリ: ***Whova***
  - 研究の議論・雑談・就活関連のチャット
  - 現地のアクティビティ募集
  - 現地で撮影した写真共有

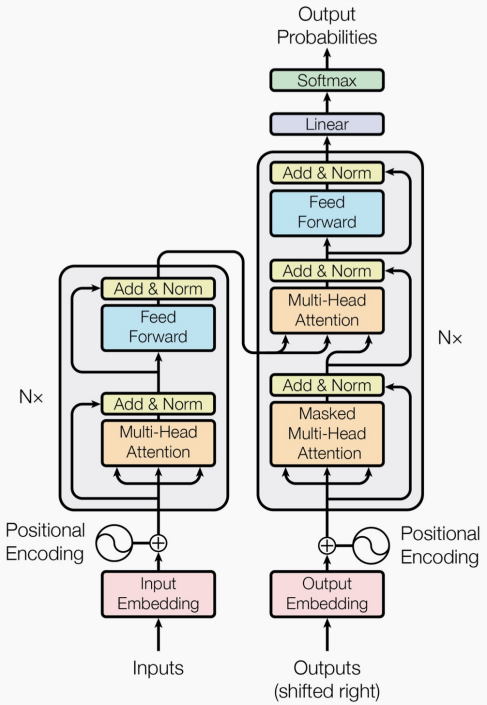


論文紹介

紹介する論文

- 主著として投稿した論文 (宣伝)
  - [Kobayashi+] Incorporating Residual and Normalization Layers into Analysis of Masked Language Models
- 言語モデル関連で興味深かった論文 × 4
  - Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little [Sinha+2021]
  - Frequency Effects on Syntactic Rule Learning in Transformers [Wei+2021]
  - Frustratingly Simple Pretraining Alternatives to Masked Language Modeling [Yamaguchi+2021]
  - CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations [Li+2021]

分野内で HOT な技術: Transformer と マスク言語モデル



- **Transformer** [Vaswani+'17]
  - 再帰 (Recurrence)・畳み込み (Convolution) を使わずに，**注意機構** (Attention) を核としたネットワーク
  - 幅広いタスクに**大幅な性能向上**をもたらした
- **マスク言語モデル** [Devlin+'19]
  - 大量の生テキストデータを使った**単語の穴埋めタスク**で学習 (**事前学習**) させたモデル
  - タスク毎の教師ありデータで追加訓練 (**fine-tuning**) することで幅広い下流タスクで高性能
  - BERT から始まり，改善を施した後継モデルが多く提案 (RoBERTa, XLNet, ALBERT など)
- 成功要因の解明と更なる改善へ
  - **「なぜ上手くいくのか」「何が出来ていないのか」の分析が盛ん**
  - **アーキテクチャ・事前学習タスク・データセットの改良が盛ん**

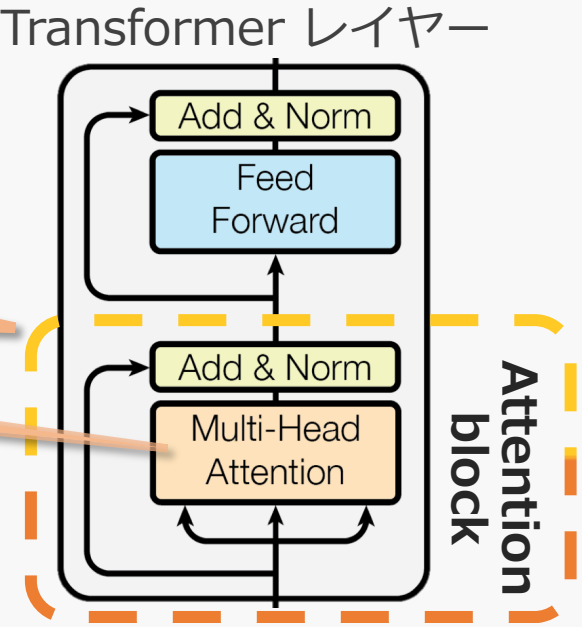
Incorporating Residual and Normalization Layers into Analysis of Masked Language Models

Goro Kobayashi[1], Tatsuki Kuribayashi[1,2], Sho Yokoi[1,3], Kentaro Inui[1,3] ([1]Tohoku U., [2]Langsmith Inc., [3]RIKEN)

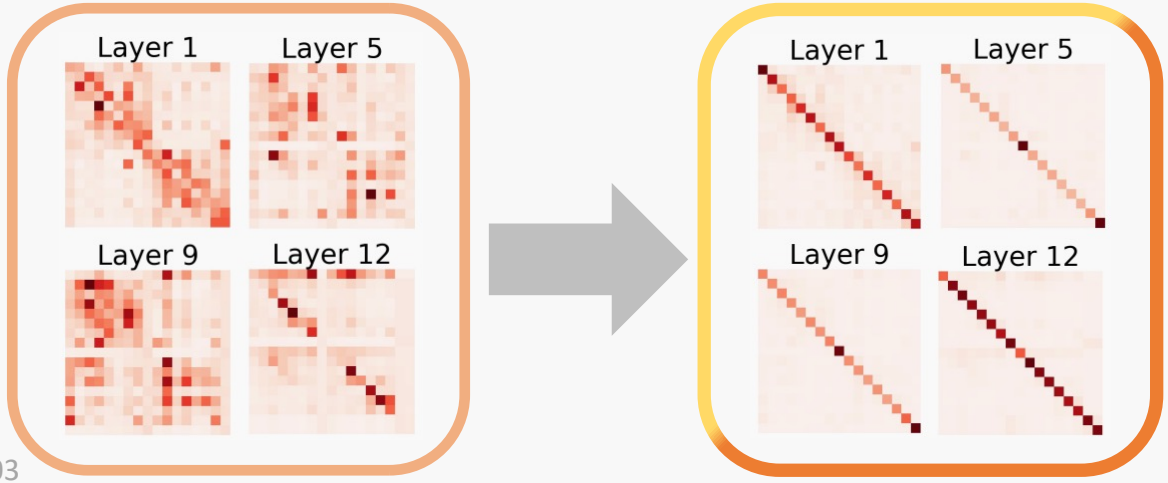
- Transformer の分析スコープを拡張することを提案

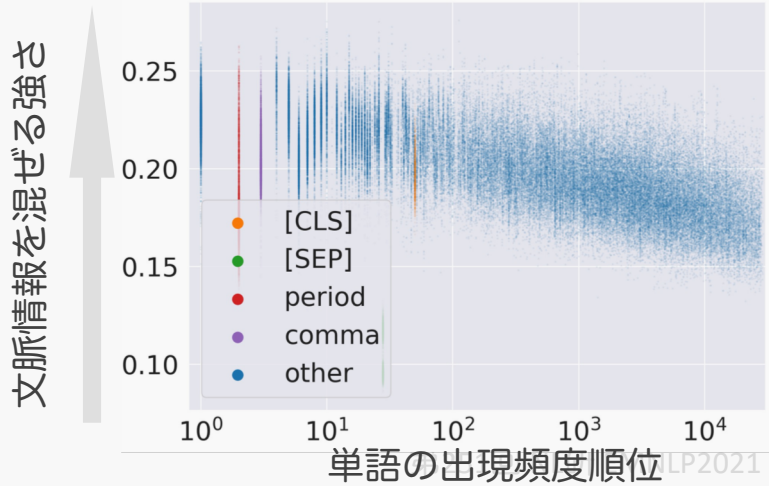
既存分析:
注意機構だけに注目した分析が典型的

提案:
残差結合と層正規化も考慮してアテンションブロックを分析



- マスク言語モデル (e.g., BERT) を分析してみると…
  - 注意機構での “周囲情報の混ぜ合わせ” はかなり弱められている
  - 高頻度語の表現ほど周囲情報を強く混ぜ合わせる傾向





Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, Douwe Kiela

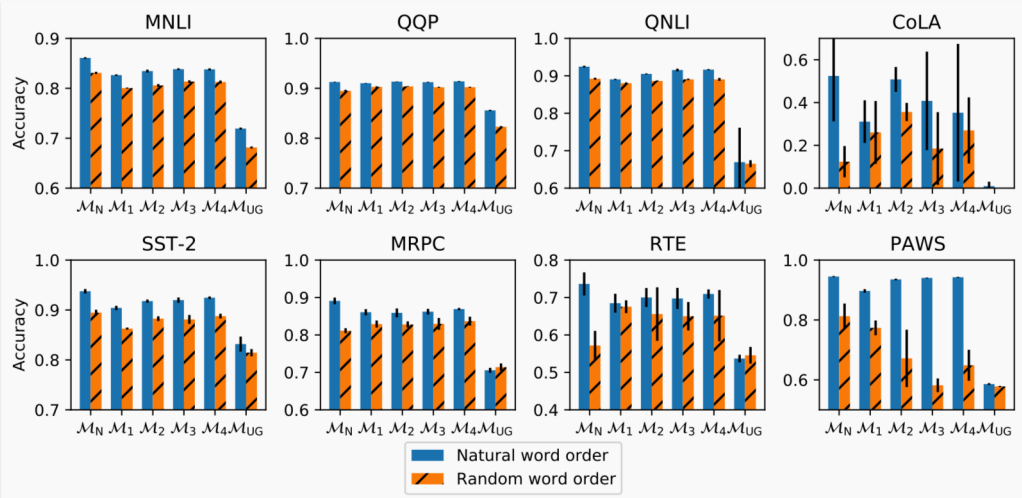
- マスク言語モデルの成功に“**事前学習時の語順情報**”は必要？
- 手法：**事前学習コーパスの各文で語順をランダムに並べ替え**て使う
  - 4種類で検証： unigram ~ 4-gram の塊で並べ替え (ローカルな分布情報は保持させる)

下流タスクでの性能が通常学習モデルと僅かしか変わらない！

➡ **語順情報は重要でない**

ローカルな分布情報を与えるほど性能向上

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\mathrm{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21
$\mathcal{M}_4$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32
$\mathcal{M}_3$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23
$\mathcal{M}_2$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.8
$\mathcal{M}_1$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10
$\mathcal{M}_{\mathrm{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.2
$\mathcal{M}_{\mathrm{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.1
$\mathcal{M}_{\mathrm{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.6



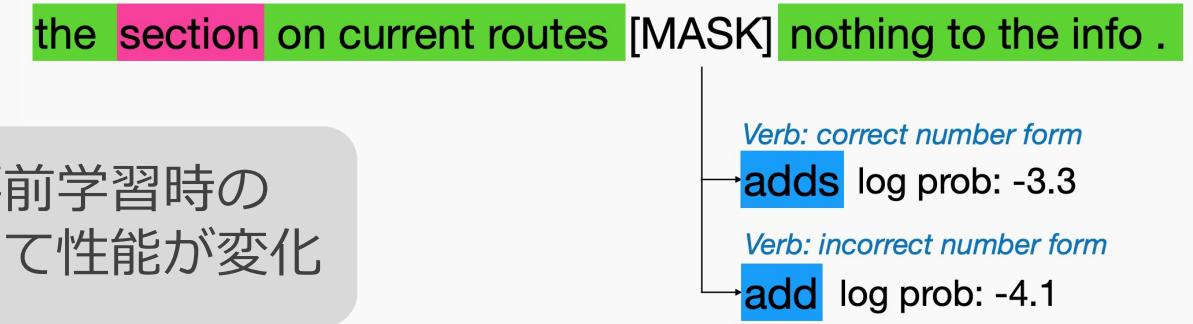
fine-tuning 時も語順シャッフルした場合の性能変化

語順情報の使い方は，下流タスクへの fine-tuning 時に学習可能（必要なら）

Frequency Effects on Syntactic Rule Learning in Transformers (1/2)

Jason Wei, Dan Garrette, Tal Linzen, Eellie Pavlick

- マスク言語モデル (BERT) の “主語と動詞の数の一致” に関するケーススタディ
- 前提：BERT は穴埋めされた動詞を予測する際に，主語と数が一致する正しい形を優先できることが知られている [Goldberg’19]
- キモ：事前学習との関係を調べる

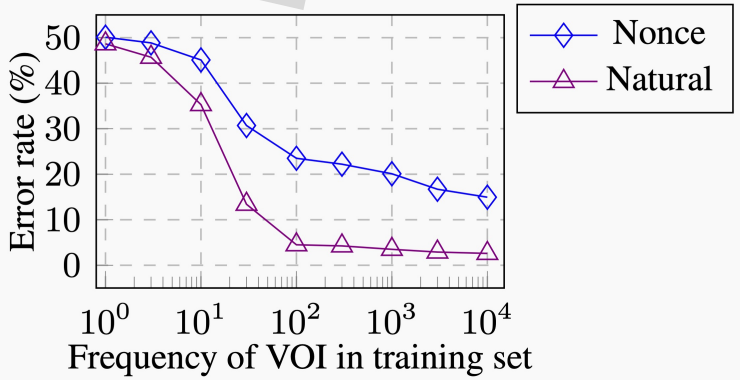


事前学習時に見ていない
主語-動詞ペアも上手く扱える
➡ **単語間の共起で判断していない**

	Natural		Nonce	
	Seen	Unseen	Seen	Unseen
argmax_V P(V)	39.1	39.0	50.0	50.0
argmax_V P(SV)	22.9	50.0	41.2	50.0
BERT	3.3	8.8	15.6	17.6

動詞予測時の数の一致でのエラー率

該当動詞の事前学習時の
出現回数に応じて性能が変化



該当動詞の出現回数に対するエラー率

Frequency Effects on Syntactic Rule Learning in Transformers (2/2)

Jason Wei, Dan Garrette, Tal Linzen, Eellie Pavlick

- マスク言語モデル (BERT) の “主語と動詞の数の一致” に関するケーススタディ
- 前提：BERT は穴埋めされた動詞を予測する際に，主語と数が一致する正しい形を優先できることが知られている [Goldberg’19]
- キモ：事前学習との関係を調べる

“主語と動詞の数は同じ” という **構文ルールを正しく適用するシステムと振る舞いが一致**

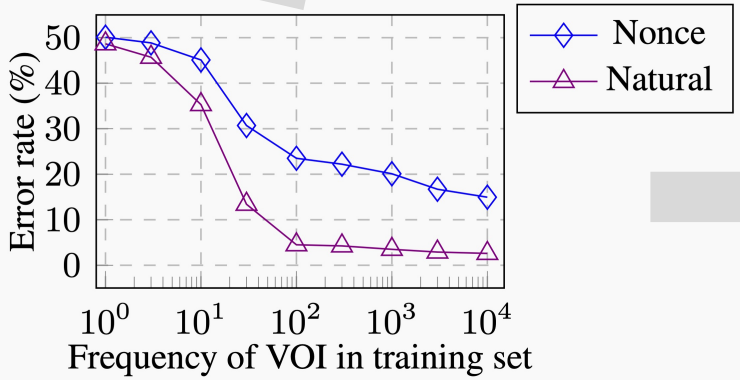
事前学習時に見ていない主語-動詞ペアも上手く扱える

➡ **単語間の共起で判断していない**

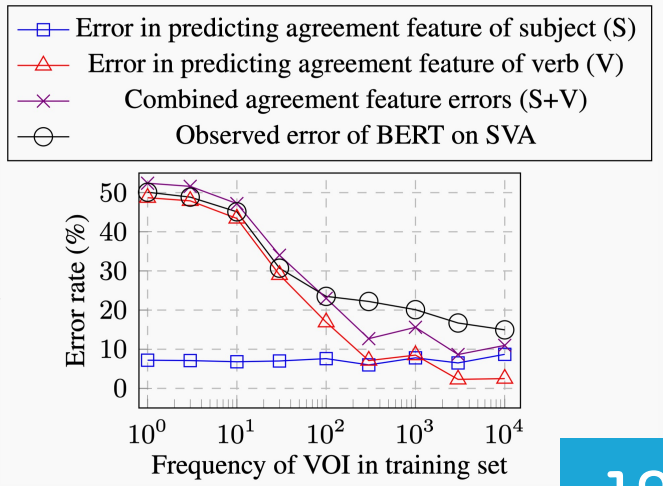
	Natural		Nonce	
	Seen	Unseen	Seen	Unseen
argmax_V P(V)	39.1	39.0	50.0	50.0
argmax_V P(SV)	22.9	50.0	41.2	50.0
BERT	3.3	8.8	15.6	17.6

動詞予測時の数の一致でのエラー率

該当動詞の事前学習時の出現回数に応じて性能が変化



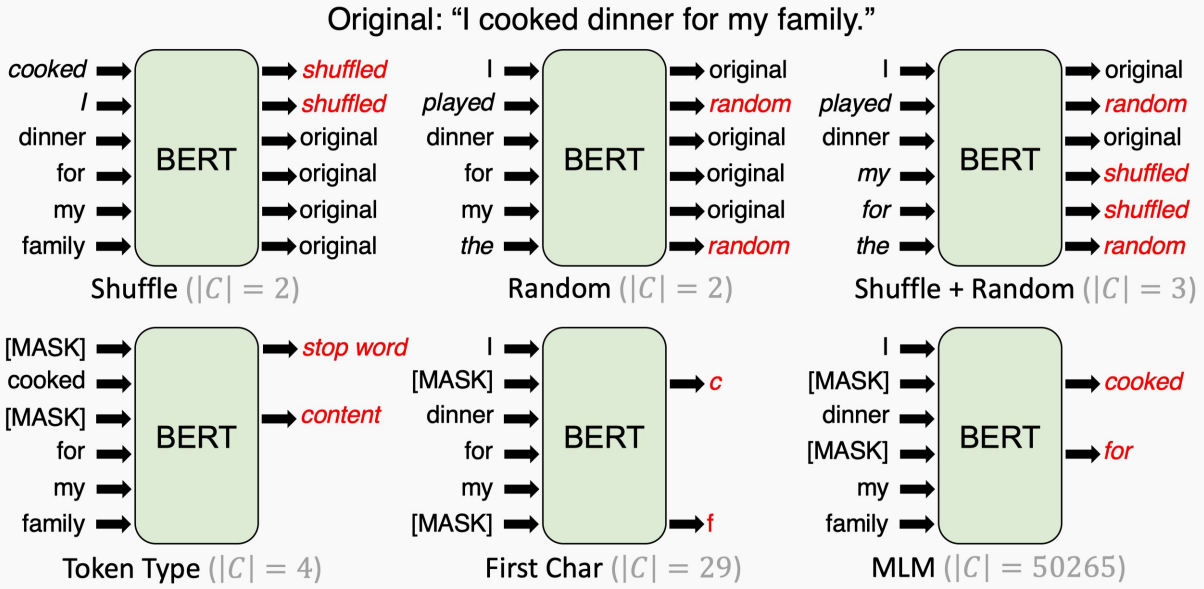
該当動詞の出現回数に対するエラー率



Frustratingly Simple Pretraining Alternatives to Masked Language Modeling

Atsuki Yamaguchi, George Chrysostomou, Katarina Margatina, Nikolaos Aletras

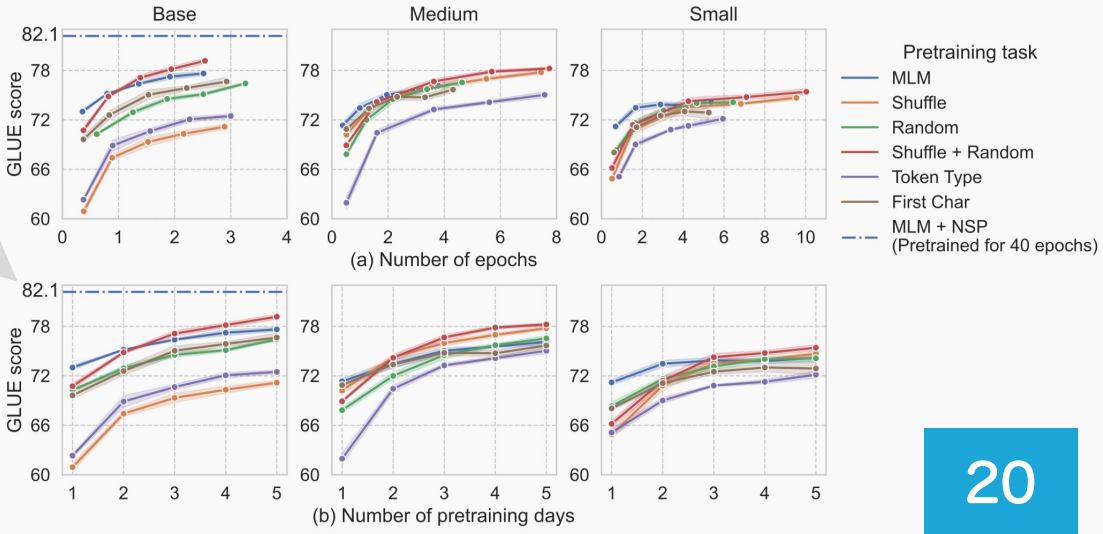
- 単語穴埋めタスク (MLM) に代わる事前学習タスクを調査
- 5種類のシンプルなタスクを提案



Shuffle+Random が MLM と同等以上の性能を発揮

Model	GLUE Average	SQuAD v 1.1
MLM	77.6 (0.2)	**84.8 (0.2)**
Shuffle	71.2 (0.3)	74.8 (0.2)
Random	76.4 (0.2)	81.6 (0.4)
Shuffle + Random	**79.2 (0.3)**	83.5 (0.2)
Token Type	72.5 (0.2)	78.6 (0.7)
First Char	76.7 (0.5)	82.0 (0.1)

Shuffle+Random は MLM よりもデータ効率 & 計算効率が良い



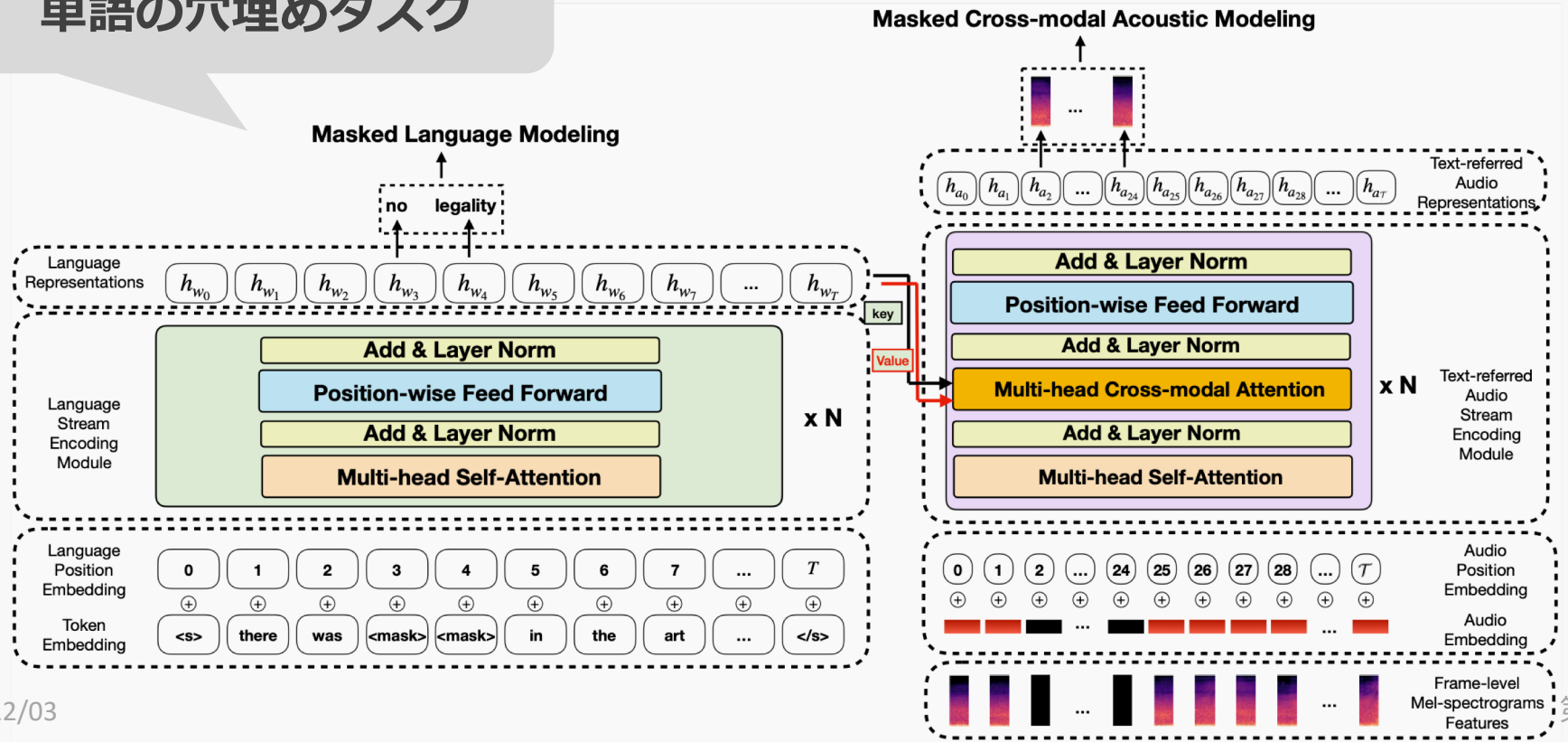
CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations (1/2)

Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, Zitao Liu

- 言語と音声のマルチモーダル処理の事前学習を提案 (初の試みらしい)

言語側は言語情報のみで単語の穴埋めタスク

音声側は言語情報も取り込みながら音声フレームの穴埋めタスク



Transformer を改良

デコーダの Attention mask 除去

未来の情報は参照できない

未来の情報も使える

Cross attention でエンコーダの最終埋め込みを参照

CTAL: Pre-training Cross-modal Transformer for Audio-and-Language Representations (2/2)

Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, Zitao Liu

- 事前学習
  - 2種類のモデルサイズ：**CTAL$_{BASE}$** (3層) と **CTAL$_{LARGE}$** (6層)
  - データセット：LibriSpeech [Panayotov+'15] (960時間の英語音声と280K発話の書き起こし)
- 下流タスクに fine-tuning (音声とその書き起こしが与えられる設定)
  - Emotion Classification: 2人の会話に対して4値の感情分類 (怒, 喜, 中立, 悲)
  - Sentiment Analysis: 映画レビューに対してポジティブ・ネガティブの度合いを予測
  - Speaker Verification: 発話が与えられて音声の特徴から発話者を特定する

Methods	WA ↑	UA ↑
LSTM_alignment (Xu et al., 2019)	0.6900	0.7014
MRDE (Yoon et al., 2018)	0.6702	0.6764
MHA (Yoon et al., 2019)	0.6780	0.6880
CTAL$_{BASE}$	0.7286	0.7370
CTAL$_{LARGE}$	**0.7395**	**0.7463**

Emotion Classification

Methods	Acc$_2$ ↑	F1 ↑	MAE ↓	Corr ↑
MulT	0.7966	0.8008	0.6367	0.6292
CTAL$_{BASE}$	0.8036	0.8055	0.6061	**0.6828**
CTAL$_{LARGE}$	**0.8077**	**0.8101**	**0.6027**	0.6809

Sentiment Analysis

Methods	EER ↓
GE2E (Wan et al., 2018)	0.0379
RawNet (Jung et al., 2019)	0.0253
CTAL$_{BASE}$	0.0194
CTAL$_{LARGE}$	**0.0155**

Speaker Verification

3タスクでSOTAを更新！

ご清聴
ありがとうございました！

ぜひ論文もお読みください！

Incorporating Residual and Normalization Layers into Analysis of Masked Language Models

Goro Kobayashi[1] **Tatsuki Kuribayashi**[1,2] **Sho Yokoi**[1,3] **Kentaro Inui**[1,3]
[1] Tohoku University [2] Langsmith Inc. [3] RIKEN

https://aclanthology.org/2021.emnlp-main.373/