国際会議報告

第12回対話システムシンポジウム

2021/11/30

報告する会議

- SIGDIAL　田中翔平（奈良先端科学技術大学院大学）
- ACL　塚越駿（名古屋大学大学院）
- INTERSPEECH
- SemDial

  千葉祐弥（NTTコミュニケーション科学基礎研究所）
- ICMI　岡田将吾（北陸先端科学技術大学院大学）
- EMNLP　佐藤翔悦（東京大学生産技術研究所）

国際会議参加報告
YRRSDS 2021
SIGDIAL 2021

田中翔平

奈良先端科学技術大学院大学

YRRSDS (7/27-28)

概要

https://sites.google.com/view/yrrsds-2021/

Young Researchers' Roundtable on Spoken Dialogue Systems

若手研究者が集まって自分の研究内容を紹介したり，最近の対話システム関連のトピックについて議論する座談会

今年度は Slack + Gathertown (+ Singapore) で開催

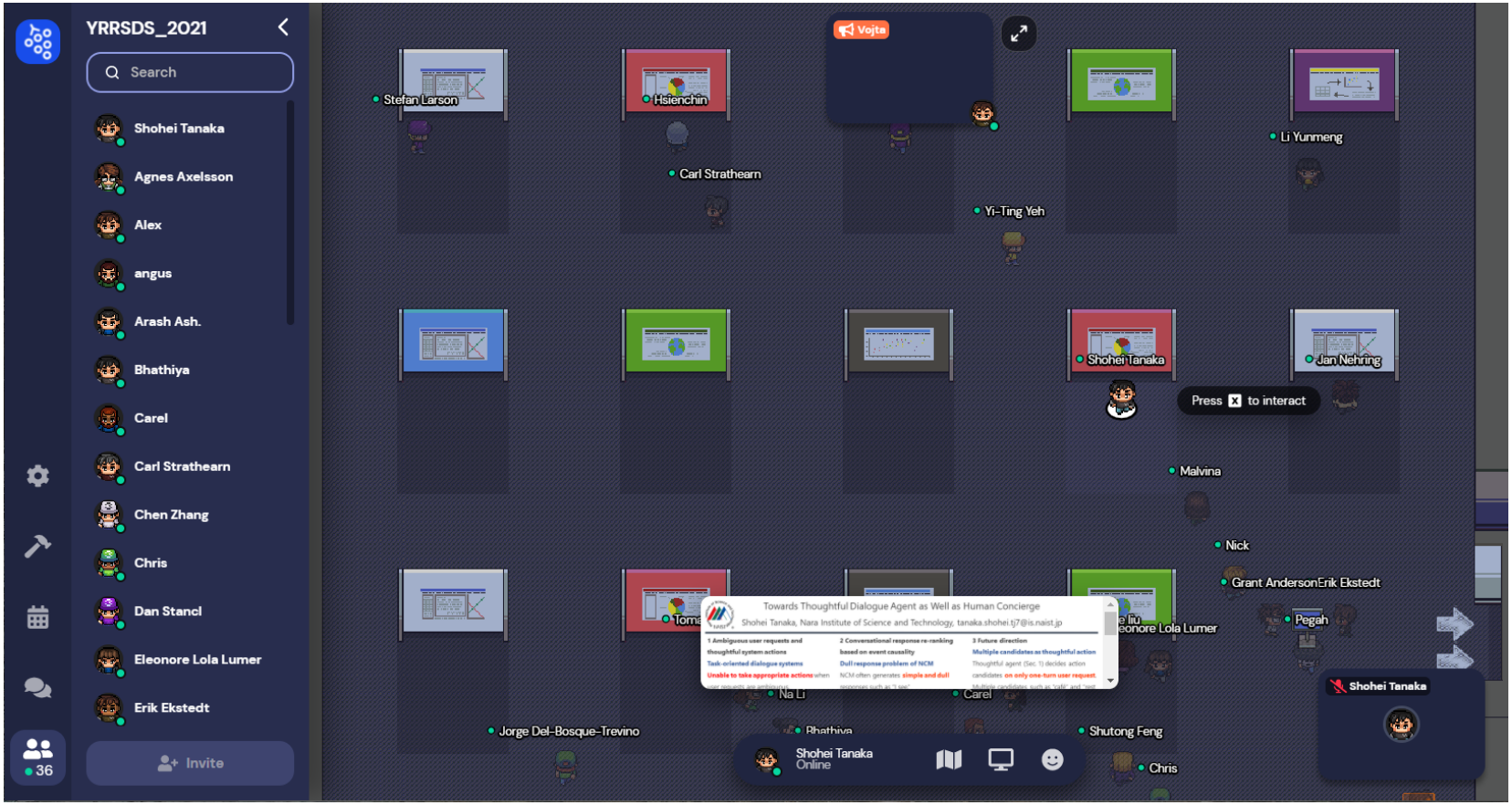
参加者は30-40人くらい

日本人の参加者が非常に少ないらしい

今年度も日本人は自分一人だけ

もっと日本人に参加してもらいたい

ポスターセッション



発表者の近くに行くとポスターが表示される

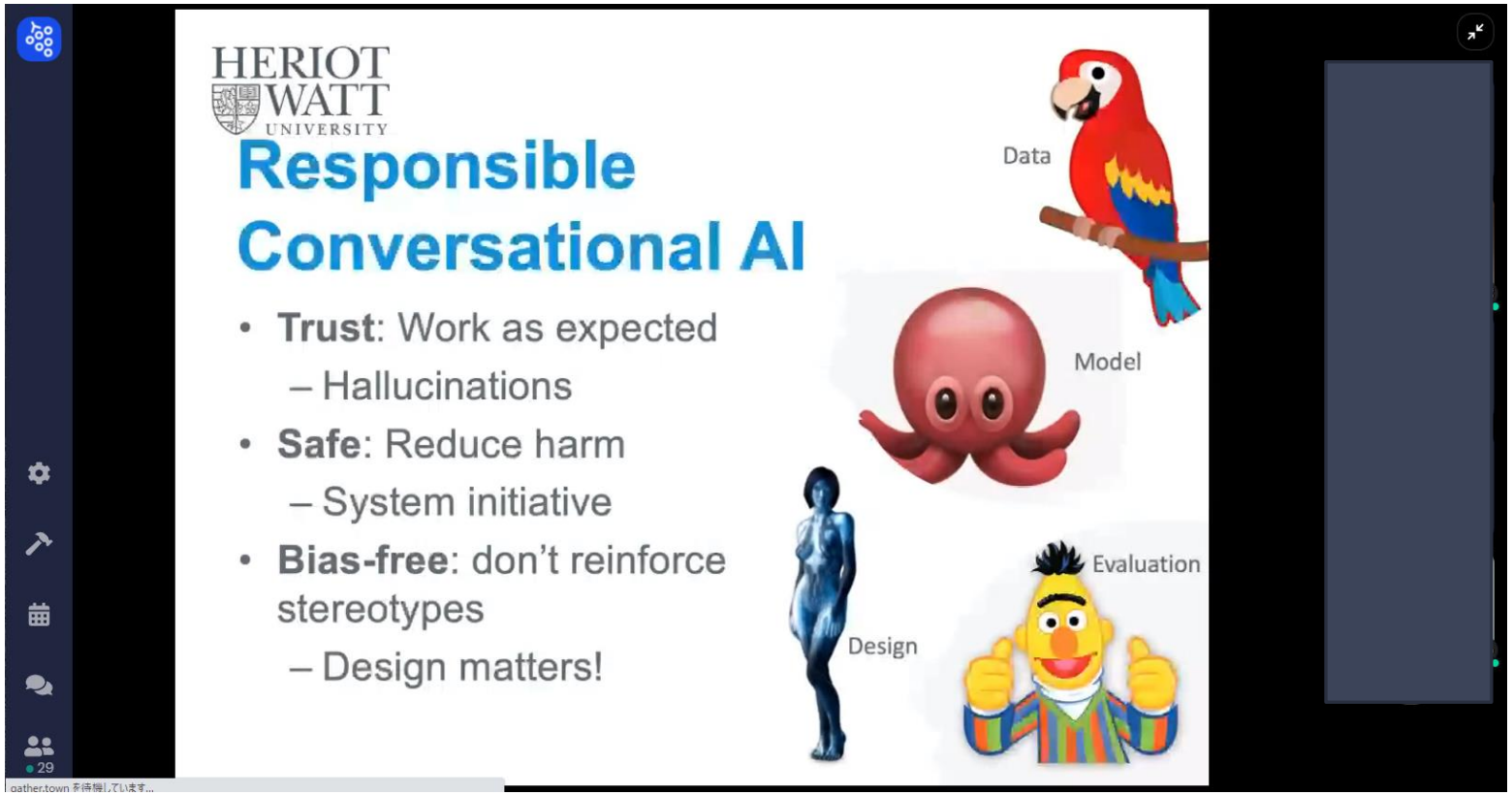
発表自体は zoom などと同じ感じで行える

ラウンドテーブル

円卓に集まって対話システム関連のトピック（e.g. システムの倫理）について議論する

普段会えない人たちの意見が聞けるので面白い

Verena Rieser 先生の講演



Responsible Conversational AI とは

期待した通りに動き，攻撃的な発言をせず，偏見を持たない対話システムで今後大きく発展するはず

SIGDIAL (7/29-31)

概要

https://www.colips.org/conferences/sigdial2021/wp/

The 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue

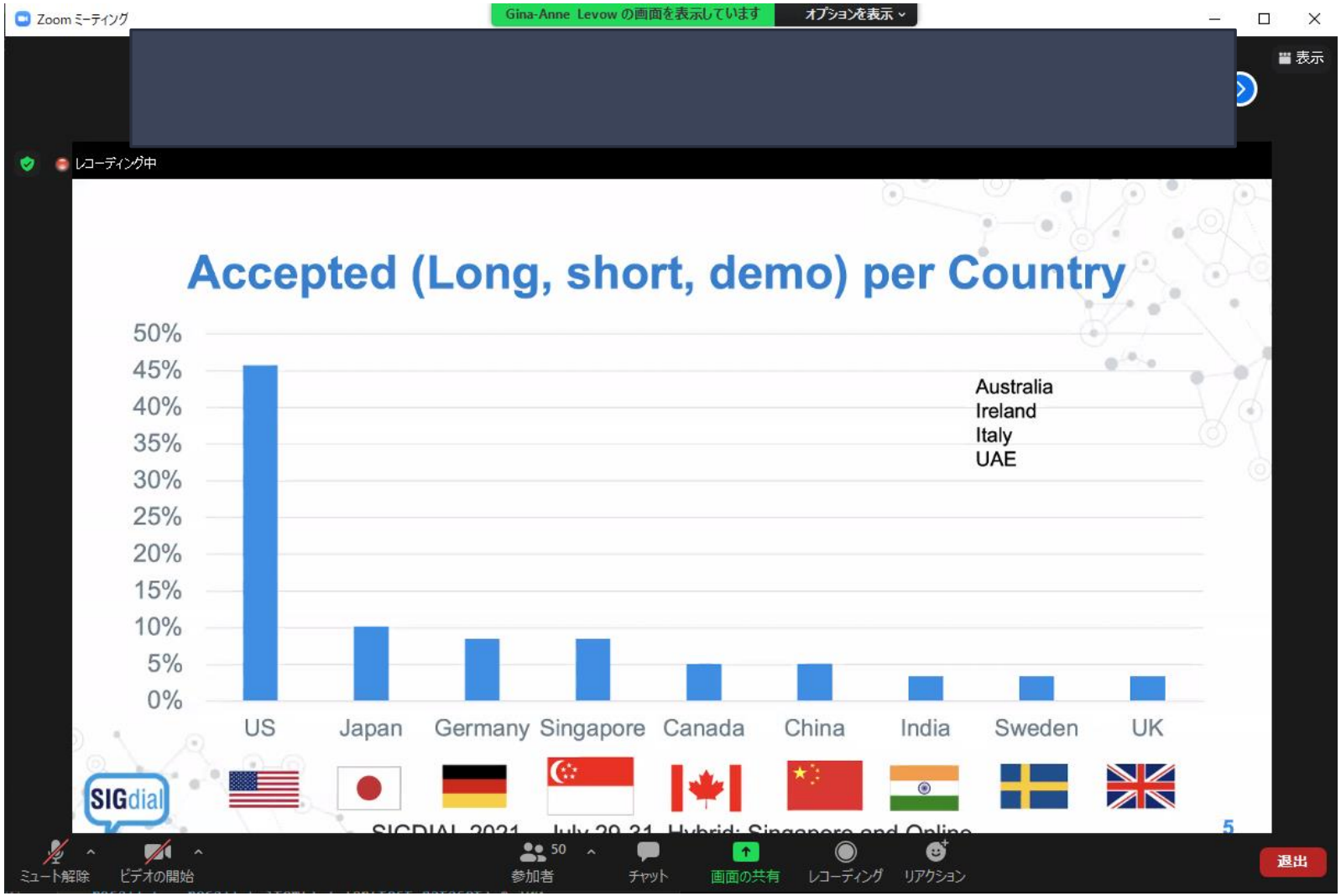
対話システムに特化した国際会議

今年度は Slack + Zoom (+ Singapore) で開催

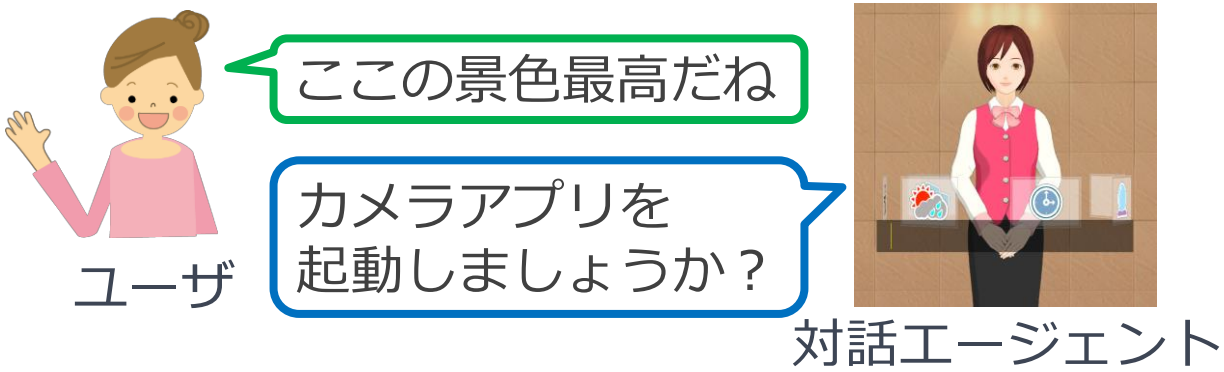
発表は事前に録画した2分のビデオ + 質疑応答

国ごとの投稿率

国ごとの採択率



ARTA: Collection and Classification of Ambiguous Requests and Thoughtful Actions



去年の SLUD で若手萌芽賞を頂いた研究の発展版

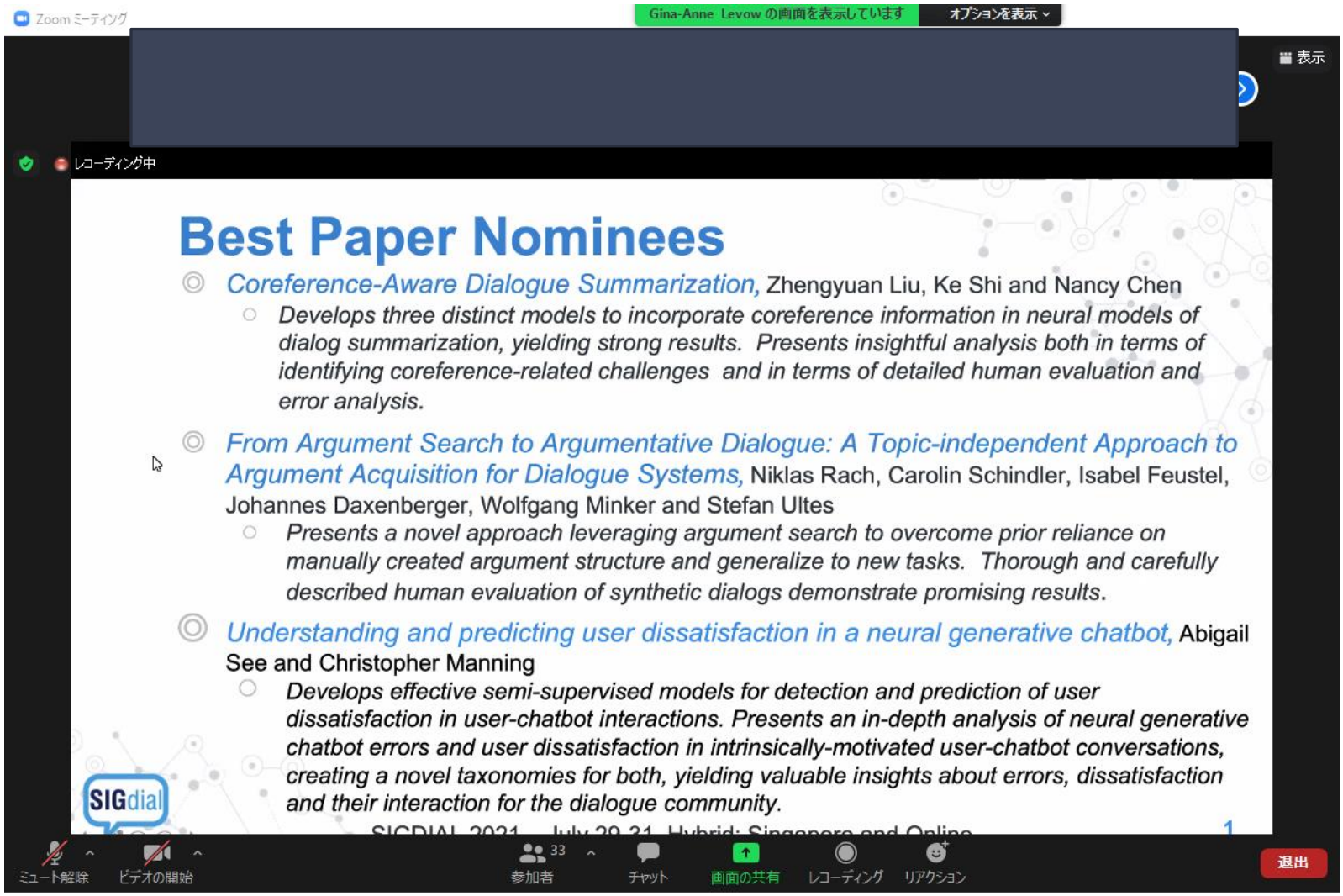
曖昧なユーザ要求に対して気の利いた行動をとることができる対話エージェントの構築を目指す

曖昧な要求と気の利いた行動をクラウドワークスで収集

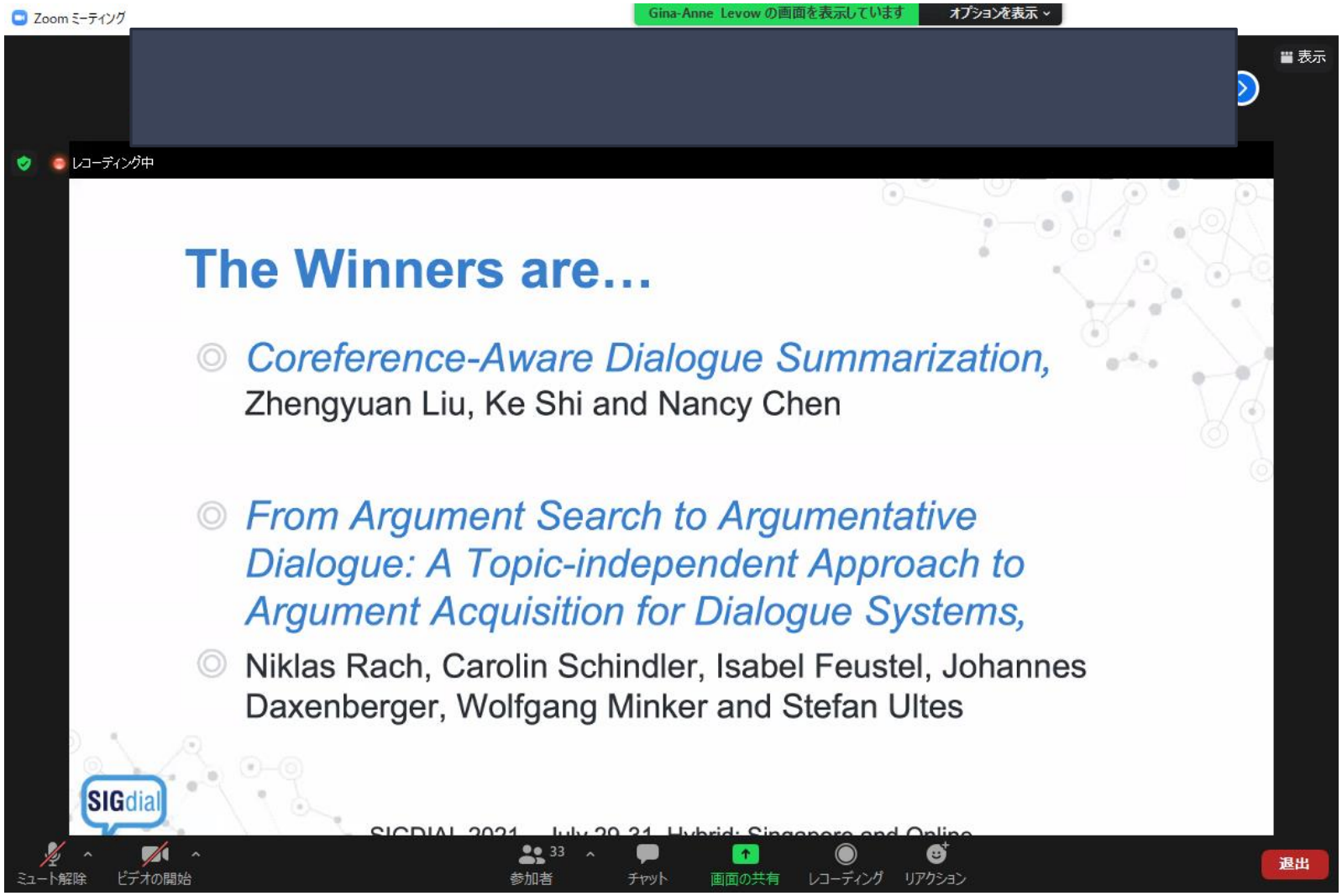
タスクに適した新しい損失関数でユーザ要求分類器を学習

提案した分類器が従来手法より高い精度で曖昧な要求に対して気の利いた行動をとれることを確認

Best Paper Nominees



Best Paper Award



第12回 対話システムシンポジウム

国際会議報告
ACL-IJCNLP 2021

塚越駿

名古屋大学 情報学研究科 修士1年 武田笹野研究室

2021/11/30

Photo by Milkovi on Unsplash

ACL-IJCNLP 2021 @ Bangkok, Thailand. 8/1 - 8/7

- Annual Meeting of the **A**ssociation for **C**omputational **L**inguistics (**ACL**)
- 自然言語処理分野における三大国際会議の一つ (他: EMNLP, NAACL)
  - 今年はIJCNLPとの同時開催 (なので名前が長い)
- オンラインでの開催

The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)

採択の難度

- main conferenceへの採録は過去5年で最も難しい年に
- 採択率は低下傾向

ACL Findingsの導入

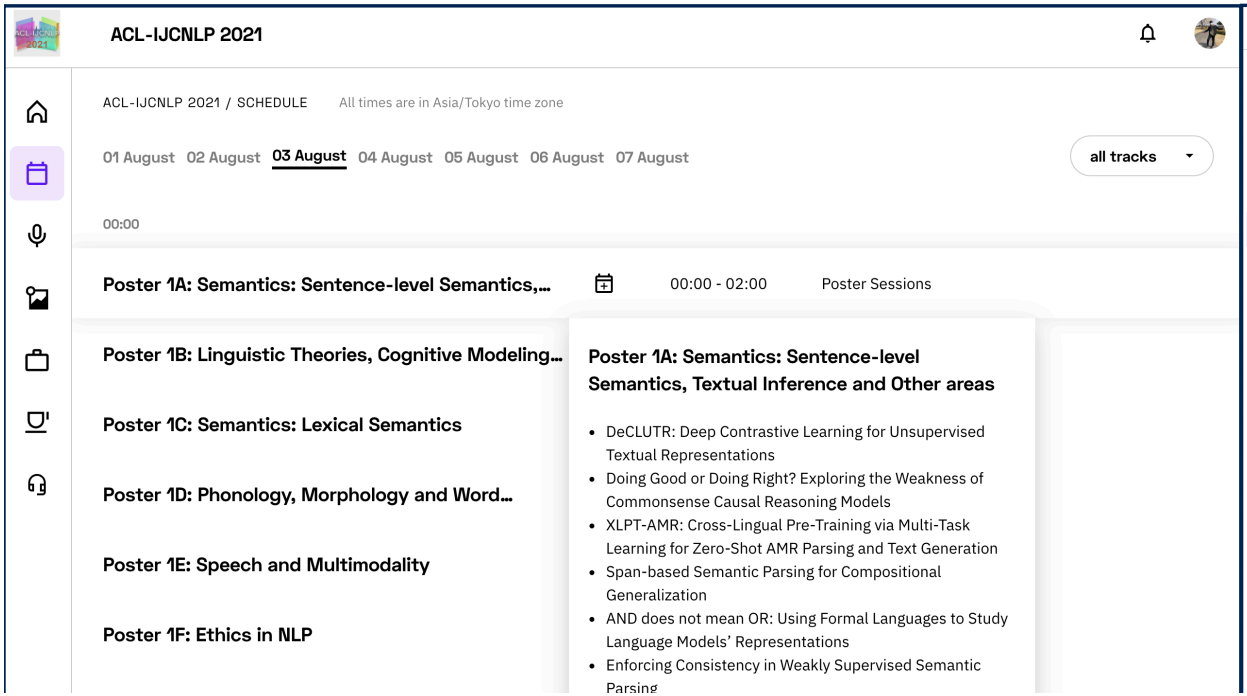
- main conferenceに準ずる質の高い論文を採録

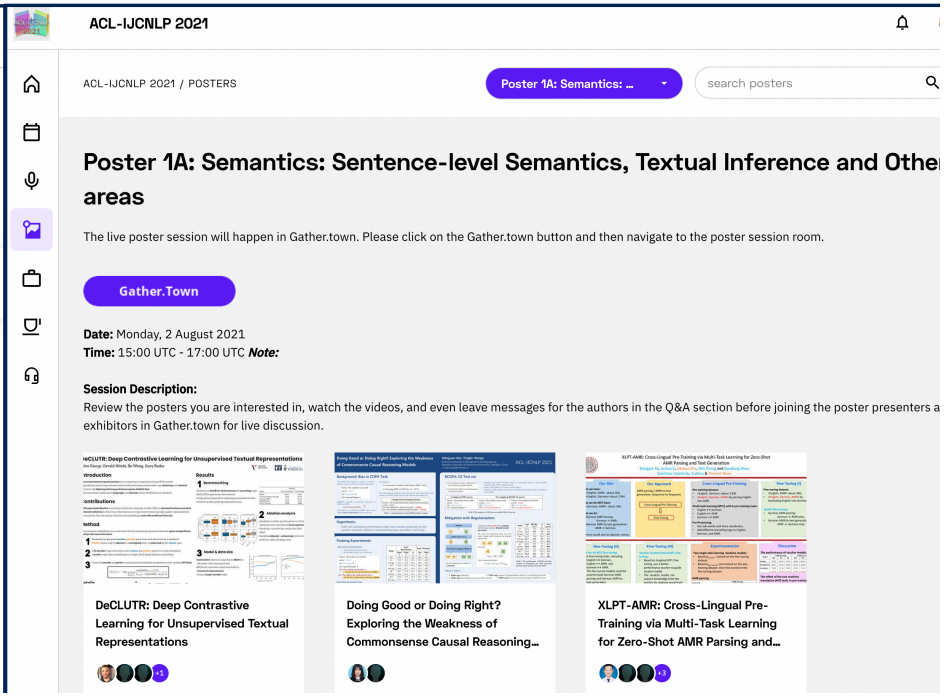
	Total submissions	Accepted	Acceptance rate (%)
ACL 2017	1297	302	23.3
ACL 2018	1544	384	24.9
ACL 2019	2680	660	24.6
ACL 2020	3429	779	22.7
ACL 2021	**3350**	**710 / 457**	**21.2 / 14.9**

main conference　　findings

オンライン開催について: プラットフォーム

- 動画ストリーミングとカンファレンスのためのプラットフォーム **underline.io** を利用
  - 論文についてのQ&Aや発表スケジュールなど種々の情報を集約
  - 発表時間を自動で日本時間にしてくれる等の便利機能も





オンライン開催について: 準備/発表

pre-recordingの撮影

- スライドを使って論文内容の紹介(口頭発表とほとんど同じ)
- SCREENCAST MATIC を用いた動画撮影・アップロード
- **long: 12分，short: 7分**

zoomでの口頭発表 or **gather.town でのポスター発表**

- 発表者が口頭発表かポスター発表かを選択
- 口頭発表の場合は
  - long: **5分**の発表 + **4分**の Q&Aセッション
  - short: **3分**の発表 + **3分**の Q&Aセッション
- ポスターの場合は2時間 Gather 内に設置された ブースで発表

Best paper awards論文の分析

Best paper

- Vocabulary Learning via Optimal Transport for Neural Machine Translation

Best theme paper

- Including Signed Languages in Natural Language Processing

Outstanding papers

- All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text
- Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning
- Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering
- Neural Machine Translation with Monolingual Translation Memory
- Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers
- UnNatural Language Inference

Best paper awards論文の分析: ジャンルごと色分け

Best paper

- **Vocabulary Learning via Optimal Transport for Neural Machine Translation**

Best theme paper

- **Including Signed Languages in Natural Language Processing**

- 手法提案
- 分析/評価
- その他

Outstanding papers

- **All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text**
- **Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning**
- **Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering**
- **Neural Machine Translation with Monolingual Translation Memory**
- **Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers**
- **UnNatural Language Inference**

Best paper awards論文の分析

Best paper

- Vocabulary Learning v…Machine Translation

Best theme paper

- Including Signed Languages in Natural Language Proce…

既存の評価手法や分析に疑問を投げかける/より詳細な分析を行う論文が複数

SOTAを追い求めるだけでなく，堅実に分野へ貢献する研究が一定の評価を得ている

Outstanding papers

- **All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text**
- **Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning**
- **Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering**
- Neural Machine Translation with Monolingual Translation Memory
- **Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers**
- **UnNatural Language Inference**

All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text

モデルの品質が向上しても，妥当な人手評価を実施できるか？

- GPT-3など大規模言語モデルの登場によって高品質な文生成が可能に，一方で評価が難化
- 人間の評価者に対して「GPT-2/GPT-3の生成したテキスト」と「人間が書いたテキスト」を見分ける課題を3つのテキストドメインで実施，判断基準の説明も収集
- さらに，評価者に練習をさせ，正解率が向上するか検証

実験結果，面白い知見

- **人間は「GPT-3の生成したテキスト」と「人間が書いたテキスト」を見分けることが出来ない** (正解率 約50%)
- 人間はテキストの**表面的な情報(句読点や詳細度)に頼る傾向がある**
- 人間は**モデルの性能を低く見積もる**傾向がある
  (例: "AIにしては生成文が自然すぎる")
- 練習内容(instruction)によって評価結果が有意に変化
  評価手法の詳細も論文に記述すべき

Training	Overall Acc.
None	0.50
Instructions	0.52
Examples	*0.55
Comparison	0.53

図表は当該論文より引用

Determinantal Beam Search

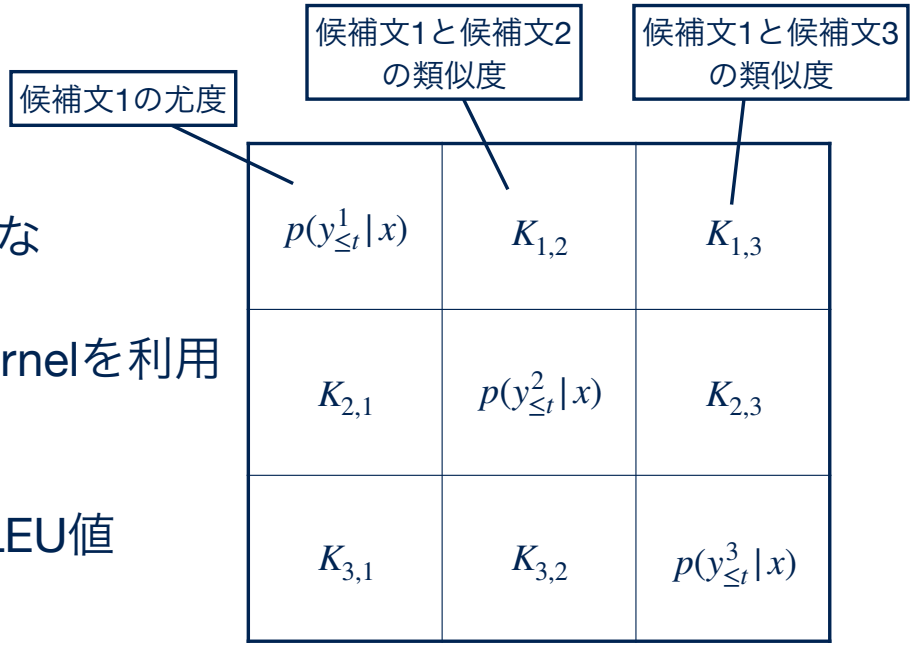
ビームサーチにおける候補文の多様性向上

- 文生成の基礎技術としてビームサーチは重要だが，候補文が多様性に欠けることがある
- 候補文同士の相互作用(interaction)を考慮して多様性(diversity)を高めることはできるか？

ビームサーチを行列式の最大化問題として再定式化

- 対角成分に候補文の尤度
- 非対角成分に候補文ペアの類似度

- 対角成分が大きく，非対角成分が小さくなるような候補文集合を選ぶ
- 論文では候補文の類似度にstring subsequence kernelを利用



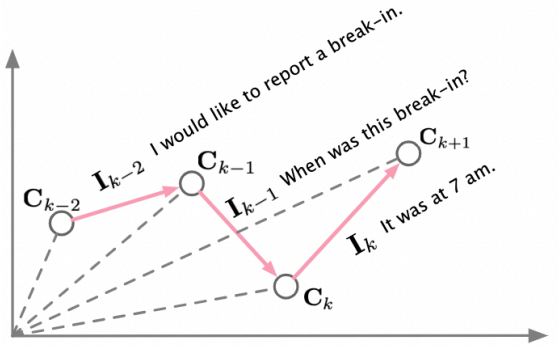
実験結果

- 機械翻訳タスクにおいて既存手法と比べて高いBLEU値を保ったまま多様な文を生成していることを確認

Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances

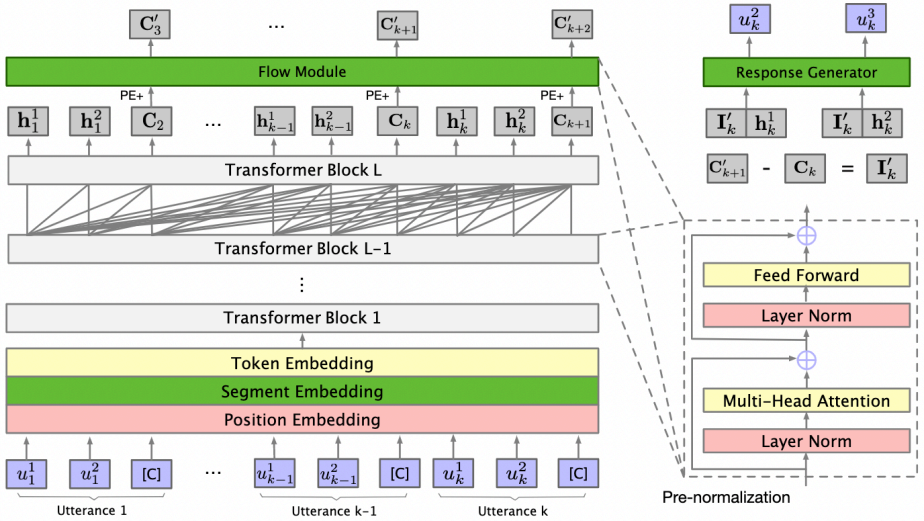
発話間の動的な情報の流れを考慮した対話モデルと事前学習

- 既存の対話モデルは過去の発話系列を全て等しく扱っている
- **発話ごとのコンテキストの変化を明示的にモデリング**することで応答の品質向上を狙う



提案手法: DialoFlow

- k番目までの発話を用いてコンテキストのベクトル表現 $C_k = \mathrm{Transformer}(u_{<k})$ を構成
- k+1番目でのコンテキストの表現 C'_{k+1} を予測
- 予測されるコンテキストの変化分 $I'_k = C'_{k+1} - C_k$ と隠れ状態を用いて応答生成



実験結果

- 対話生成タスクでDialoGPTなどを上回る性能

図表は当該論文より引用

A Cognitive Regularizer for Language Modeling

Uniform Information Density (UID) 仮説

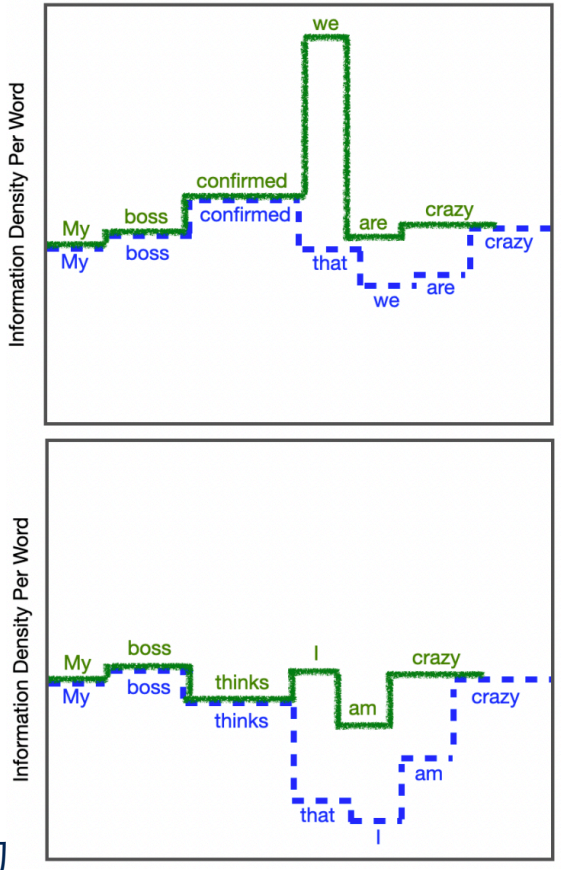
- 話者は文中の情報密度が一様になるように言語学的シグナルを分布させるという仮説
- 既存研究では言語学的現象の説明のためにUID仮説が用いられていた

UID仮説を言語モデルのinductive biasとして組み込む

- UID仮説に基づく正則化項は性能によい影響を与えるか？
- 二つのUID正則化 (**UID Regularizer**) を提案
  - 正則化項1: 文中の単語の情報量の分散
  - 正則化項2: 文中の隣り合う単語の情報量の差の二乗

実験結果

- UID正則化はともにパープレキシティを低減，特に低資源言語で有効



図表は当該論文より引用


ACL2022 @ Dublin, Ireland. 5/22 - 5/27
査読体制が大きく変更
• ACL Rolling Review (ARR) を中心とし，OpenReviewを用いて査読
• ARRでの査読→各conferenceへの投稿という流れに
• 投稿締切: 2021/11/15 AOE (終了済み😭)
Photos from www.2022.aclweb.org

ピックアップ論文リスト

- All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text
- Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances
- Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks
- UnNatural Language Inference
- Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers
- Contrastive Learning for Many-to-many Multilingual Neural Machine Translation
- Data Augmentation for Text Generation Without Any Augmented Data
- Lightweight Cross-Lingual Sentence Representation Learning
- When Do You Need Billions of Words of Pretraining Data?
- LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding
- Modeling Fine-Grained Entity Types with Box Embeddings
- A Cognitive Regularizer for Language Modeling

ピックアップ論文リスト

- Determinantal Beam Search
- Lower Perplexity is Not Always Human-Like
- Dynamic Contextualized Word Embeddings

NTT

INTERSPEECH 2021

千葉 祐弥 (NTT)

30 August – 3 September 2021
INTERSPEECH 2021
BRNO | CZECHIA
Speech everywhere!

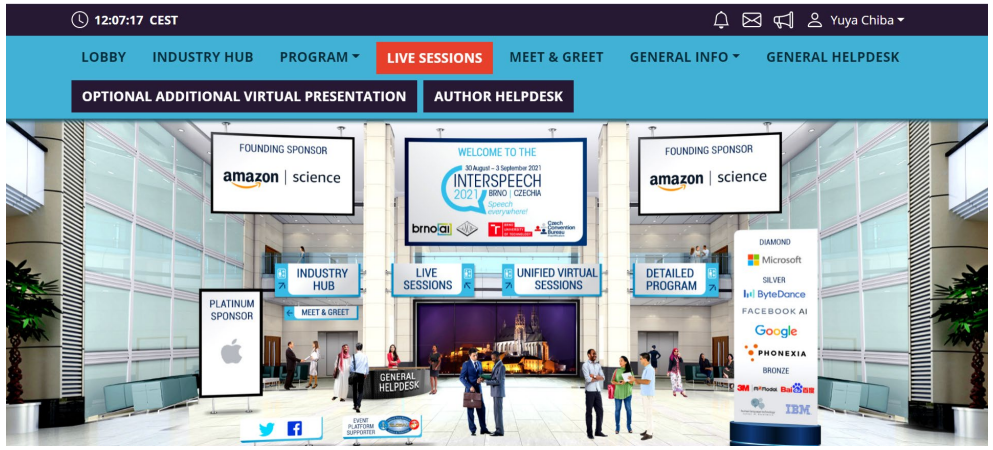
BRNO UNIVERSITY OF TECHNOLOGY

INTERSPEECHの概要

- International Speech Communication Association (ISCA)主催の会議
- 音声科学・音声工学のflagshipカンファレンス
  - 近年はICASSPに並ぶくらいの注目度

	Publication	h5-index	h5-median
1.	Conference of the International Speech Communication Association (INTERSPEECH)	89	150
1.	IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)	96	143

- 採択率 (963/2277 = 42.3%)
  - 例年50%くらい⇒ちょうどいい難易度か
  - ここ数年は徐々に難化？



https://www.interspeech2021.org/

Areas & Topics

- SCOPE (音声対話システム関連の抜粋)

  11.1 Spoken dialog systems
  11.2 Discourse and dialog structures
  11.3 Multimodal interaction and interfaces
  11.4 Conversation, communication and interaction
  11.5 Analysis of verbal, co-verbal and nonverbal behavior
  11.6 Language modeling for conversational speech (dialog, interaction)
  11.7 Interactive systems for speech/language training, therapy, communication aids
  11.8 Stochastic modeling for dialog
  11.9 Question-answering from speech
  11.10 Systems for spoken language understanding
  11.11 Other topics in Spoken dialog systems and conversational analysis

- 対話におけるメジャーな研究トピック
  - 音声処理を導入した対話制御 (意図推定，認識仮説のリランキング，認識誤りへの対応)
  - 音声言語理解

Keynotes

- “Forty years of speech and language processing: from Bays decision rule to deep learning” (Prof. Hermann Ney)

- “Ethical and Technological Challenges of Conversational AI” (Prof. Pascale Fung)

- “Adaptive listening to everyday soundscape” (Prof. Mounya Elhilali)

- “Language Modeling and Artificial Intelligence” (Prof. Tomas Mikolov)

https://www.interspeech2021.org/keynotes

ピックアップ論文

- 対話＋音声認識
  - Leveraging ASR N-Best in Deep Entity Retrieval
  - User-initiated Repetition-Based Recovery in Multi-Utterance Dialogue Systems
- 対話＋音声合成
  - Controllable Context-Aware Conversational Speech Synthesis
- マルチモーダル対話
  - Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information

Leveraging ASR N-best in Deep Entity Retrieval

Haoyu Wang[1], *John Chen*[2†], *Majid Laali*[3], *Kevin Durda*[3], *Jeff King*[3], *William Campbell*[1], *Yang Liu*[1]

- Amazon Alexa, USA
- エンティティ検索のタスク
  - ユーザ中の発話のエンティティ候補に対する最もふさわしいカタログ内のエンティティを検索
  - 対話システムにおけるドメイン固有の音声認識はいまだに難しい（出現が稀，Out-of-vocabulary, …）
- N-best結果から得られたエンティティリストを用いて検索精度を向上
- Dual encoderベースの手法，N-best結果の統合方法を比較
- 提案手法であるself-attentionでは11.07%のRelative Error Reductionを達成

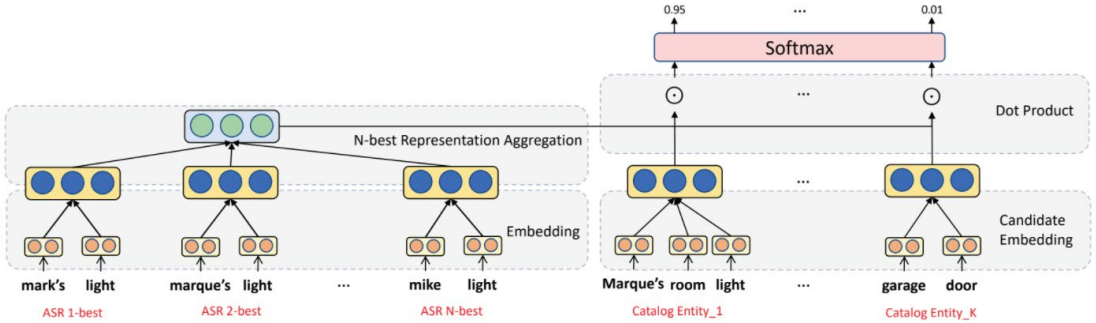


Figure 1: *Dual encoder architecture for ASR N-best ER model. All ASR N-best mentions will be aggregated into a single representation. Each candidate entity will be encoded individually and the score will be calculated by a dot product with ASR N-best representation.*

Table 1: *Experiment results for the in-domain data. The Relative Error Reduction for a model m is calculated by comparing the relative difference between* $(100\% - Accuracy_m)$ *and* $(100\% - Accuracy_{baseline})$.

	Relative Error reduction(%)
Baseline w/o ASR N-best	0.00
Mean Pooling	6.02
Learnable Weights	8.61
Global Attention	8.57
Concatenation	8.87
Self-Attention	**11.07**

https://www.isca-speech.org/archive/interspeech_2021/wang21b_interspeech.html

User-Initiated Repetition-Based Recovery in Multi-Utterance Dialogue Systems

Hoang Long Nguyen, Vincent Renkens, Joris Pelemans, Srividya Pranavi Potharaju, Anil Kumar Nalamalapu, Murat Akbacak

- Apple, USA
- 音声認識誤りを含むユーザ発話とユーザの繰り返し発話を用いて最初の発話を正しく書き換えるタスク
  - 繰り返し発話は認識誤りに起因するおかしなシステムの挙動に対するユーザの行動として一般的 [Swertze et al., 2000]
- Query Rewriteの音声対話システムへの適用
  - 書記素レベルの埋め込み + 2-Step pointer network
- 元発話と生成発話の編集距離で書き換えるかどうか判定
- 提案手法は不要な書き換えを抑えつつ通常のpointer networkより高いWERRを達成

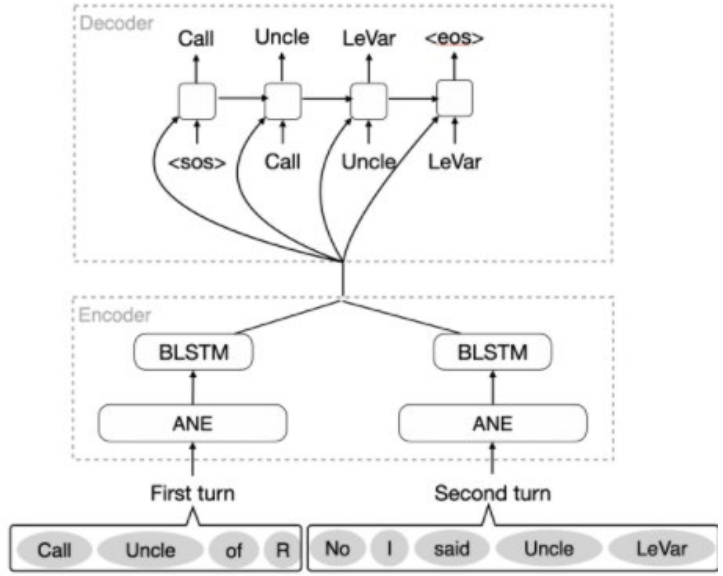


Table 2: *We evaluate the model performance on the test set. We find the best possible WERR and show its corresponding FAR*

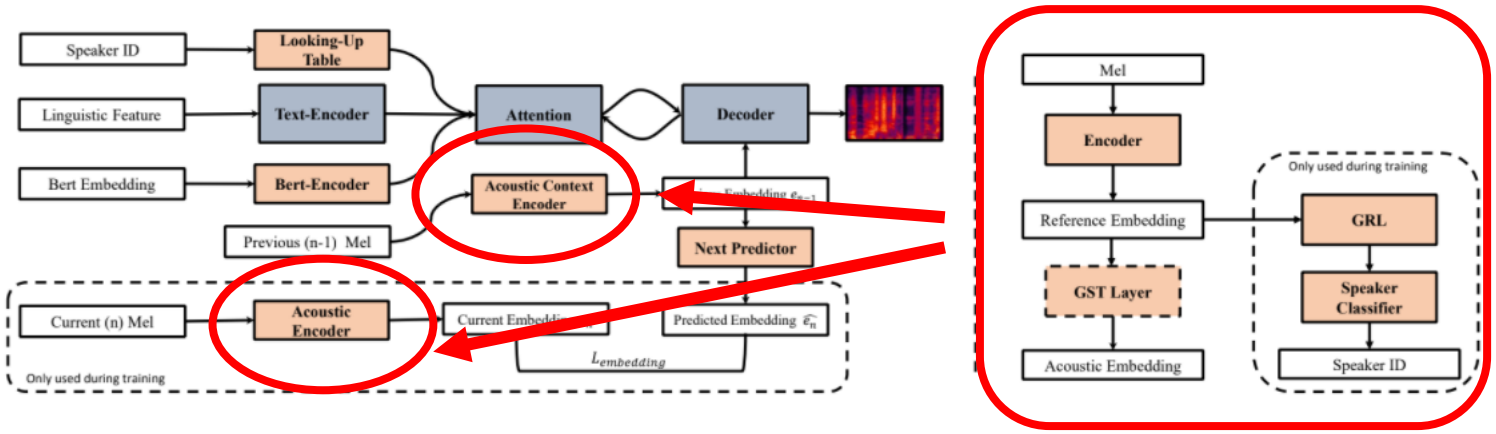
Model	Max WERR	FAR
rule-based baseline	26.60%	48.88%
pointer-network	19.03%	**2.15%**
2 step attention pointer-network	25.80%	6.77%

https://www.isca-speech.org/archive/interspeech_2021/nguyen21_interspeech.html

Controllable Context-aware Conversational Speech Synthesis

Jian Cong[1†], *Shan Yang*[2], *Na Hu*[2], *Guangzhi Li*[2], *Lei Xie*[1*], *Dan Su*[2]

- 対話用音声合成 (個人的な注目トピック)
- Prolongationとfilled pauseに対応するトークンを導入してTacotron2ベースのモデルを学習
- 同調を考慮するためAcoustic context encoderを導入
- 推論時のトークンの位置はBERTベースの方法で推定 (= controllable)



- 先行発話の音響情報をEncodeし，Tactron2に導入
- 学習時には先行発話から後続発話のAcoustic Embeddingを推定し，実際のEmbeddingと近くなるように学習
- 話者性をなくすように学習

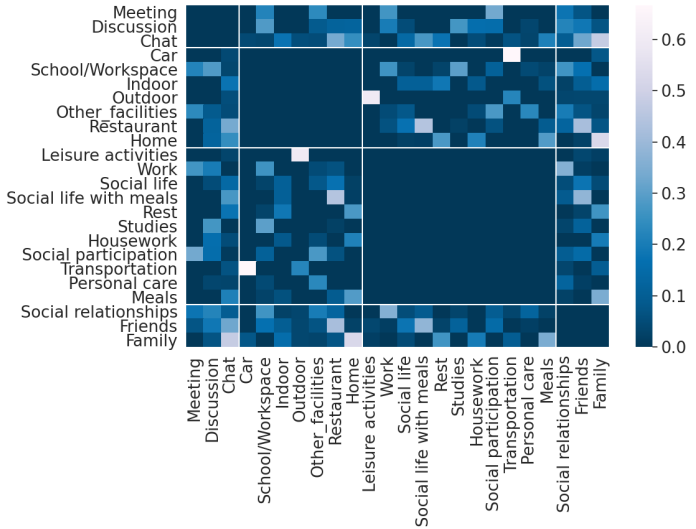
Cf.) Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, “Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems,” in Proc. INTERSPEECH, 2021

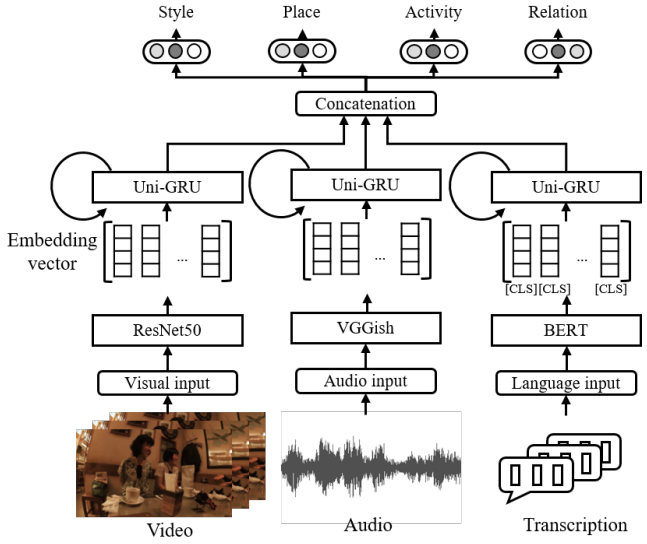
https://www.isca-speech.org/archive/interspeech_2021/cong21b_interspeech.html

Dialogue Situation Recognition for Everyday Conversation Using Multimodal Information

Yuya Chiba[1] *and Ryuichiro Higashinaka*[2]

- 周囲の状況に合わせた応答や行動，機能の調整が必要
  - 包括的な対話状況の推定手法を検討
- CEJCを利用し，状況間の関連を調査
- マルチモーダル情報を用いたマルチタスク学習により対話の状況を認識





https://www.isca-speech.org/archive/interspeech_2021/chiba21_interspeech.html

INTERSPEECH 2022

- Paper submission due: 2022年3月21日
- Paper update due: 2022年3月28日
- Notification: 2022年6月13日

https://interspeech2022.org/

SemDial 2021

千葉 祐弥 (NTT)

SemDialの概要

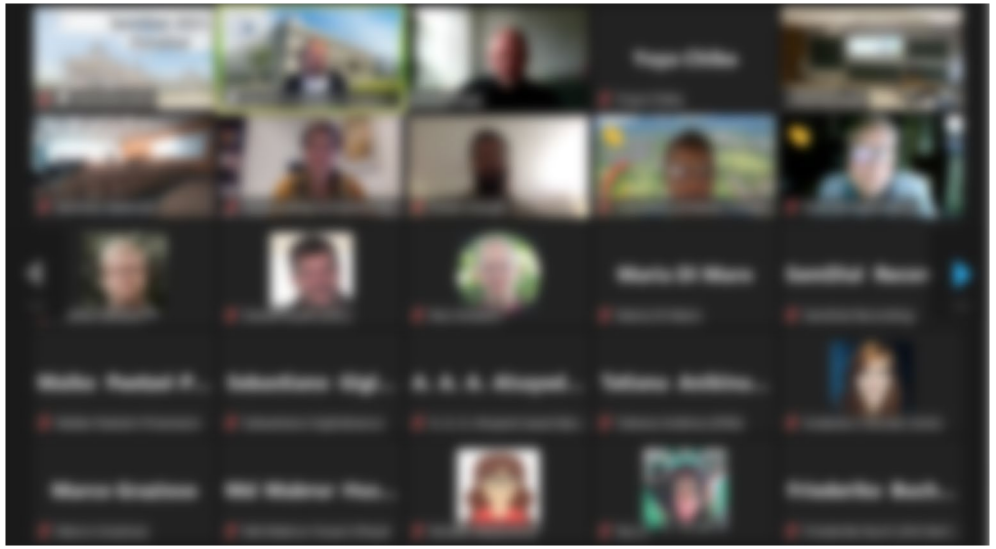
- 独立系の対話の会議
  - 対話の理論，計算言語学，人工知能，哲学，心理学，神経科学などにフォーカス
- 会議の歴史は長い
  - The 25th Workshop on the Semantics and Pragmatics of Dialogue
  - SIGDIALとフォーカスが似ているがやや理論的な側面が強い？
- 採択率 (Full paper: 15/28 = 53.6%)
- 今年は日本からの発表は1件
  - 8ページの原稿が結構大変だが、途中段階の研究も許容

Areas & Topics

- The dynamics of agents' information states in dialogue
- Common ground/mutual belief
- Turn-taking and interaction control
- Semantic/pragmatic interpretation in dialogue
- Categorization of dialogue phenomena in corpora
- The psycholinguistics of dialogue
- Multimodal and multi-party dialogue
- Conversation analysis applied to human/agent interaction
- Dialogue management
- Designing and evaluating dialogue systems

SIGDIALに近いSCOPE（semantic/pragmaticがやや高い位置にある）

会議の特徴



- 発表件数（ロングペーパー）
  - 大体17件前後
  - ここ10年では10～21件
- シングルトラックなので研究を知ってもらえる
- 対話研究者が集まっているので深い議論ができる

Keynotes

- Human-aware conversational agents (Prof. Stefan Kopp)

- How to Escape the Encodingism Stranglehold: Dynamic Syntax, Process and Interaction (Prof. Ruth Kempson)

- Leveraging the wisdom of the crowd to realize a character-like chatbot (Prof. Ryuichiro Higashinaka)

https://semdial2021.ling.uni-potsdam.de/programme/

ピックアップ論文

- Confusionの推定
  - Detecting Interlocutor Confusion in Situated Human-Avatar Dialogue: A Pilot Study
- Rapportのモデル化
  - By the way, do you like Spider Man?” — Towards A Social Planning　Model for Rapport
- 親密さの推定
  - Speaker intimacy in chat-talks: Analysis and recognition based on verbal and non-verbal information

Detecting Interlocutor Confusion in Situated Human-Avatar Dialogue: A Pilot Study

Na Li, John D. Kelleher, Robert Ross

- タスク指向対話を対象としたマルチモーダルシステムにおけるConfusionの検出
- 最近似たタスクをちらほら見る(Keynote, ICMI, HRI, …) (個人的に注目のタスク)
- WOZシステムとの対話でデータを収集
  - 質問の複雑さ，情報の過不足，エージェントの動作などで話者のConfusionを誘発
- Confusionを誘発した場合とそうでない場合の画像特徴量を比較
  - 表情: Confusion時によりNegative
  - 顔向き: Confusion時に変動が小さい
  - 視線: Confusion時に変動が多い

http://semdial.org/anthology/papers/Z/Z21/Z21-3013/

“By the way, do you like Spider Man?” — Towards A Social Planning Model for Rapport

Alafate Abulimiti[1], Justine Cassell[1], Jonathan Ginzburg[2]

- タスク指向の対話においてOff-taskの対話を挿入するべきタイミングを検討
  - 特にRapportベースの方法を検証
- 仮説
  - Off-task対話はRapportが低い状態が続いたときに起こる
  - Off-task対話はタスク達成率を高める
- Tutoring対話を収集・分析
  - Off-task対話の長さと学習効果の相関は0.3558
  - Off-task対話はRapportが対話全体の平均よりも低いときに起こる傾向
- 対話の進行に伴いRapportが蓄積するモデルを構築し，検証

$$r_a(t) = rp(t_0) + \gamma * rp(t_1) \ldots + \gamma^{(\Delta t-1)} * rp(t_0 - \Delta t)) = \sum_{t=t_0-\Delta t}^{t=t_0} \gamma^{t-t_0} * rp(t)$$

http://semdial.org/anthology/papers/Z/Z21/Z21-3004/

Speaker Intimacy in Chat-Talks: Analysis and Recognition based on Verbal and Non-Verbal Information

Yuya Chiba[1], Yoshihiro Yamazaki[2], Akinori Ito[2]

- 対話相手との関係に基づく対話制御モデル構築のための親密さ推定
- 親密さの段階による話者の言語・非言語行動の違いを分析
  - 単語の利用や対話行為，音声の同調，表情の同調，視線の変動に差
- Context BLSTMとMulti-stream BLSTMを用いた三段階の親しみ推定
  - 全てのモダリティを考慮した場合に最も高い性能
  - 3発話応答対用いた場合に性能が最大

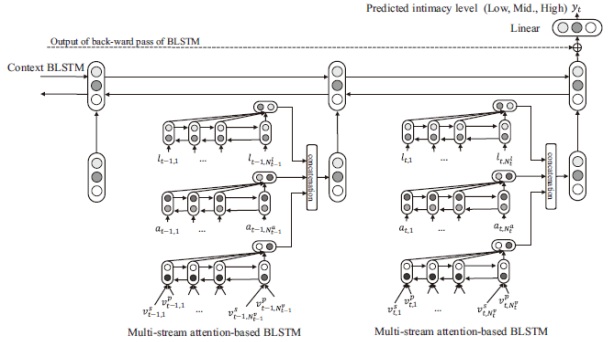


Figure 3: Network architecture for intimacy recognition: $l_{t,n}$ and $a_{t,n}$ are linguistic and acoustic features at frame n of t-th utterance. $v^i_{t,n}$ is the visual features of participant $i \in (s, p)$. s and p represent the speaker and the partner, respectively. N^l_t, N^a_t, and N^v_t are sequence length of linguistic, acoustic, and visual features. y_t is prediction result. $\oplus$ shows the summation.

Table 6: Intimacy Recognition Results: A, V, and L denote acoustic, visual, and linguistic features, respectively. Rec., Pre., and F1. represent the recall, precision, and F1-score. Bold fonts are the best performance between modalities. Chance shows results when all test samples are classified as high-level intimacy, which is the most frequent label.

	Low			Middle			High			Macro Average		
Modality	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
A	0.313	0.739	0.439	0.399	0.354	0.375	0.565	0.257	0.353	0.425	0.450	0.389
V	0.262	0.554	0.356	0.165	0.032	0.053	0.568	0.495	0.529	0.332	0.360	0.313
L	**0.857**	0.770	**0.811**	0.272	0.196	0.228	0.652	0.762	0.703	**0.594**	0.576	0.581
A+V	0.275	0.709	0.397	0.499	0.338	0.403	0.587	0.266	0.367	0.454	0.438	0.389
A+L	0.627	**0.835**	0.716	0.469	0.314	0.376	0.671	0.672	0.672	0.589	0.607	0.588
V+L	0.759	0.791	0.775	0.258	0.115	0.159	0.652	**0.801**	**0.719**	0.557	0.569	0.551
A+V+L	0.567	0.811	0.667	**0.506**	**0.432**	**0.466**	**0.693**	0.608	0.648	0.589	**0.617**	**0.594**
Chance	–	–	–	–	–	–	–	–	–	0.177	0.333	0.231

http://semdial.org/anthology/papers/Z/Z21/Z21-3007/

SemDial 2022

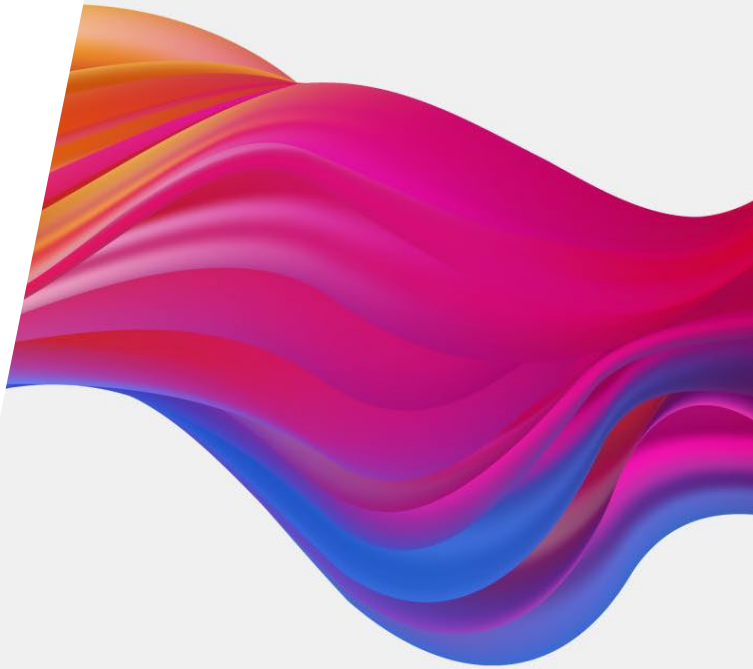
- Submission due (参考): 2021年3月31日 ⇒ 2021年6月7日
- 3月くらいには投稿準備をしたい

https://twitter.com/semdialmeeting/status/1440670477903544327

第12回対話システムシンポジウム
国際会議報告

ACM International Conference on Multimodal Interaction 2021 (ICMI 2021)
October 18-22, 2021

岡田将吾(北陸先端大)



ICMIのスコープ

- **マルチモーダルデータ処理・インタラクションに関するACM国際会議**
  - premier international forum for multidisciplinary research on multimodal human-human and human-computer interaction, interfaces, and system development.
- **Scientific と Technical Novelty の双方を重視**
- **ICMI2021年のメインテーマ**
  - Behavioral Health and Virtual Connectivity

ICMIが関連する会議

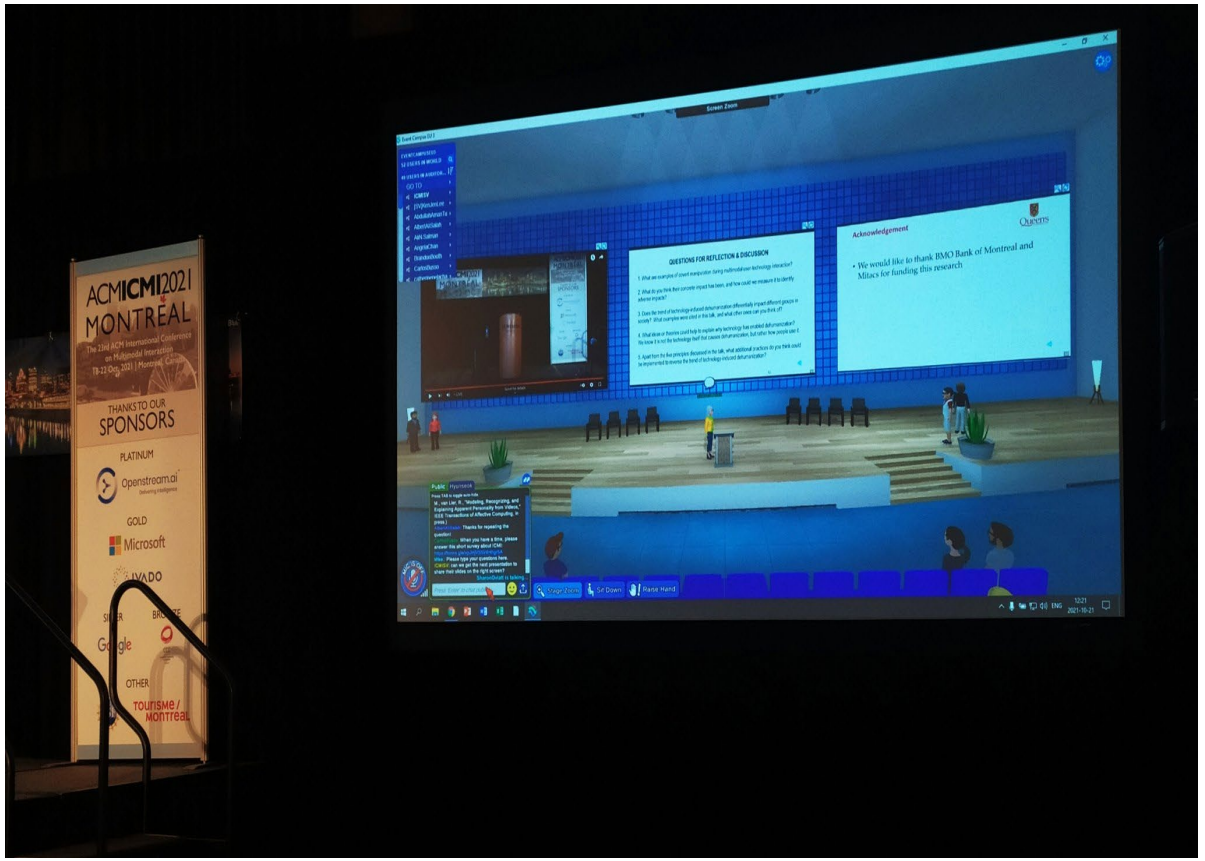
- Affective Computing and Intelligent Interaction: ACII (AAAC，IEEE系)
  - Affective Computing (感情を扱う研究全般)
- Intelligent Virtual Agent: IVA (ACM系)
  - バーチャルエージェント，会話エージェント（ECA）
- 他：
  - ACM Multimedia
  - ACM Intelligent User Interface (IUI)
  - ACM Computer Supported Cooperative Work (CSCW)
  - ACM CHI
  - ACM Human Robot Interaction (HRI)
  - ACM Human Agent Interaction (HAI)

ICMI2021 stats:

- Regular paper (Long (8p+ref.) and Short (4p+ref.))
  - 投稿数247,　採択論文数93（38%: オーラル (14%)）
- Late Breaking Papers track :7 papers (28%)
- Doctoral Consortium track: 9 papers (69%)
- Demonstrations track 7 papers (54%)
- Blue Sky Papers track (新しい研究の方向性を提言する論文) 3papers (27%)
- Special tutorials
  - Tutorial on Ethics in AI and human-machine interaction (5件のトーク)
- Registration: **315人（59人がオフライン参加**）

開催方法：ハイブリッド

- 現地（カナダ：モントリオール）
  ＋オンライン



11 workshops:

ヘルスケア

- Automated Assessment of Pain (AAP)
- Social Affective Multimodal Interaction for Health (SAMIH) (NAIST **田中先生，中村先生**)
- Socially-Informed AI for Healthcare – Understanding and Generating Multimodal Nonverbal Cues

グループダイナミクス/ 社会的スキル

- Insights on Group & Team Dynamics
- Corpora and Tools for Social Skills Annotation (CATS2021)

感情

- Modelling socio-emotional and cognitive processes from multimodal data in the wild
- Affective Social Multimedia Computing (ASMMC 2021)
- Multimodal Affect and Aesthetic Experience

ロボットとエージェント

- Empowering Interactive Robots by Learning Through Multimodal Feedback Channels
- Generation and Evaluation of Non-verbal Behavior for Embodied Agents

社会科学と幼児の行動分析

- Bridging Social Sciences and AI for Understanding Child Behavior

3 keynotes:

From Differentiable Reasoning to Self-supervised Embodied Active Learning
Ruslan Salakhutdinov (CMU)

Incorporating haptics into the theatre of multimodal experience design; and the ecosystem this requires
Karon MacLean (UBC)

Theory Driven Approaches to the Design of Multimodal Assessments of Learning, Emotion, and Self-Regulation in Medicine
Susanne P. Lajoie (McGill Univ.)

Presentations (Best paper nominees):

A Systematic Cross-Corpus Analysis of Human Reactions to Robot Conversational Failures:

- ユーザ応答のマルチモーダル情報よりインタラクションの失敗を自動検出するための分析

ViCA: Combining Visual, Social, and Task-oriented Conversational AI in a Healthcare Setting:

- タスクベース対話と，視覚情報に基づいた対話（ユーザの服装に関する話など．．）を統合した病院の受付ロボットの提案・評価

What's Fair is Fair: Detecting and Mitigating Encoded Bias in Multimodal Models of Museum Visitor Attention:

- 博物館における来館者のエンゲージメント推定モデルの学習におけるバイアスの検出と緩和方法の提案

Presentations (Best paper nominees):

Exploiting the Interplay between Social and Task Dimensions of Cohesion to Predict its Dynamics Leveraging Social Sciences 🏆 :

- グループディスカッションにおける社会的関係性と課題解決スコアの間の相互作用分析と集団の結束力予測

Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis:

- 2つのモダリティ間の融合と相違のモデル化に基づいたマルチモーダルセンチメントモデリング（SOTA更新）

A Multimodal Dataset and Evaluation for Feature Estimators of Temporal Phases of Anxiety:

- 不安反応の動的変化分析のための生体信号を含むマルチモーダルデータの提案と不安状態の推定

Impact of the Size of Modules on Target Acquisition and Pursuit for Future Modular Shape-changing Physical User Interfaces :

- ユーザーの入力やシステムの出力に対して複数のインタラクションモダリティをサポートするために、物理的な形状を変化させるUI

Presentations

- Recognizing Social Signals with Weakly Supervised Multitask Learning for Multimodal dialogue Systems (Oral)

一致度の低いラベルなどのノイズラベルを除去する
マルチモーダル弱教師付き学習手法の提案

- Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems (Poster)

ユーザの対話システムとの対話の満足度をマルチモーダル情報から予測する方法の提案・評価

- ICMI2022はインドのBangaloreで開催予定

EMNLP 2021 参加報告

東京大学 生産技術研究所
佐藤 翔悦

研究紹介: トピック別 (1/5)

Knowledge-grounded:

- 低頻度エンティティの積極的な利用を促すため、知識のグラフ構造を活用するモデルを提案
  - *EARL: Informative Knowledge-Grounded Conversation Generation with Entity-Agnostic Representation Learning*
- 大規模応答生成モデルの知識を検索ベースのモデルに蒸留し、高速化を図る研究
  - *Distilling the Knowledge of Large-scale Generative Models into Retrieval Models for Efficient Open-domain Conversation (findings)*
- 複数文書を根拠とした会話タスク・データの提案 (4796対話、平均14ターン、488文書)
  - *MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents*

大規模モデルの重さ・分析の困難さから、外部知識を陽に扱う手法への関心が高まっている印象

研究紹介: トピック別 (2/5)

Multi-modal (visual):

- BlenderBotをマルチモーダル化、Redditデータで実験
  - *Multi-Modal Open-Domain Dialogue*
- CVAEライクな、分布を陽に仮定したアーキテクチャを提案しどの物体に注目すべきかをより正確に判断
  - *Learning to Ground Visual Objects for Visual Dialog (findings)*
- 衣類・家具店を模したシミュレータ上で、店員と客のやり取りを収集したデータ (約11.2k対話、117k発話)
  - *SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations*

機械翻訳などと同様に、近年発表件数が増加している分野。新規データセット作成の動きも活発

研究紹介: トピック別 (3/5)

Dialogue Summarization:

- Code-switchingと呼ばれる会話の途中で使用する言語が部分的に変わる問題への対処
  - *GupShup: Summarizing Open-Domain Code-Switched Conversations*
- 多人数会話要約のデータセット提案。既存のSAMSumと比較して、新規ドメインに転移した際高性能との結果も
  - *ForumSum: A Multi-Speaker Conversation Summarization Dataset*
- カスタマーサービスドメインの対話要約データ
  - *TWEETSUMM - A Dialog Summarization Dataset for Customer Service*

発表件数がとても多かった印象。需要が増え、かつ解けていない課題としての要約側からの注目も？

研究紹介: トピック別 (4/5)

Dialogue State Tracking (DST):

- GPT-2ベースのState trackerに知識グラフをエンコードするモジュールを加えて学習、MultiWOZ 2.0で実験
  - *Knowledge-Aware Graph-Enhanced GPT-2 for Dialogue State Tracking*
- ドメイン適応を繰り返しモデルを成長させる手法における致命的忘却を解決、蒸留ベースの手法を提案
  - *Domain-Lifelong Learning for Dialogue State Tracking via Knowledge Preservation Networks*
- 多言語BERTをタスク・ドメインと関連する他の言語のデータに対して追加で事前学習, DSTでの性能向上を確認
  - *Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking*

事前学習済み言語モデルを前提とするものが多数
ドメイン知識の利用にフォーカスした研究も

研究紹介: トピック別 (5/5)

Personalization:

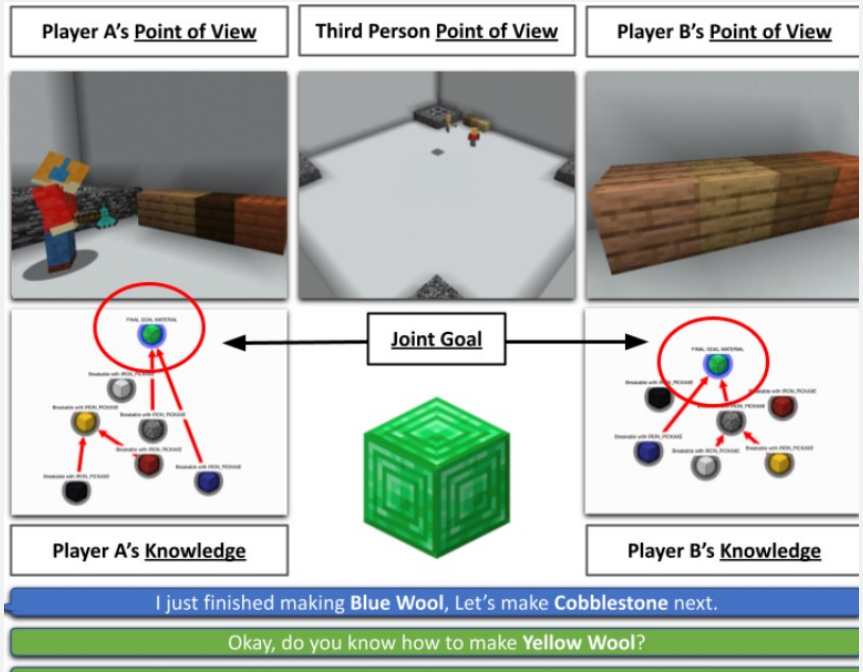
- 与えられたペルソナを元に応答を書き換えるタスクを提案
  - *Transferable Persona-Grounded Dialogues via Grounded Minimal Edits*
- 話者ペルソナを既存の会話から推定するためのデータ、モデルを提案。Persona-Chatに対応
  - *Detecting Speaker Personas from Conversational Texts*

Evaluation:

- 対話モデルの応答が外部知識と矛盾しないかどうかを、質問作成・応答モデルの会話への適用によって評価
  - *Q2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering*
- 既存の23種類の評価手法についての議論・検証論文
  - *A Comprehensive Assessment of Dialog Evaluation Metrics (workshop)*

MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks (Outstanding paper)

- MineCraft上でのGoal-oriented dialogueの実験
  - ２人のクラウドワーカが対話しながら協力して材料を組み合わせゴールとなるブロックを作成
  - それぞれのプレイヤーは何と何を組み合わせたら何になるか、という知識が部分的に与えられる



MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks (Outstanding paper)

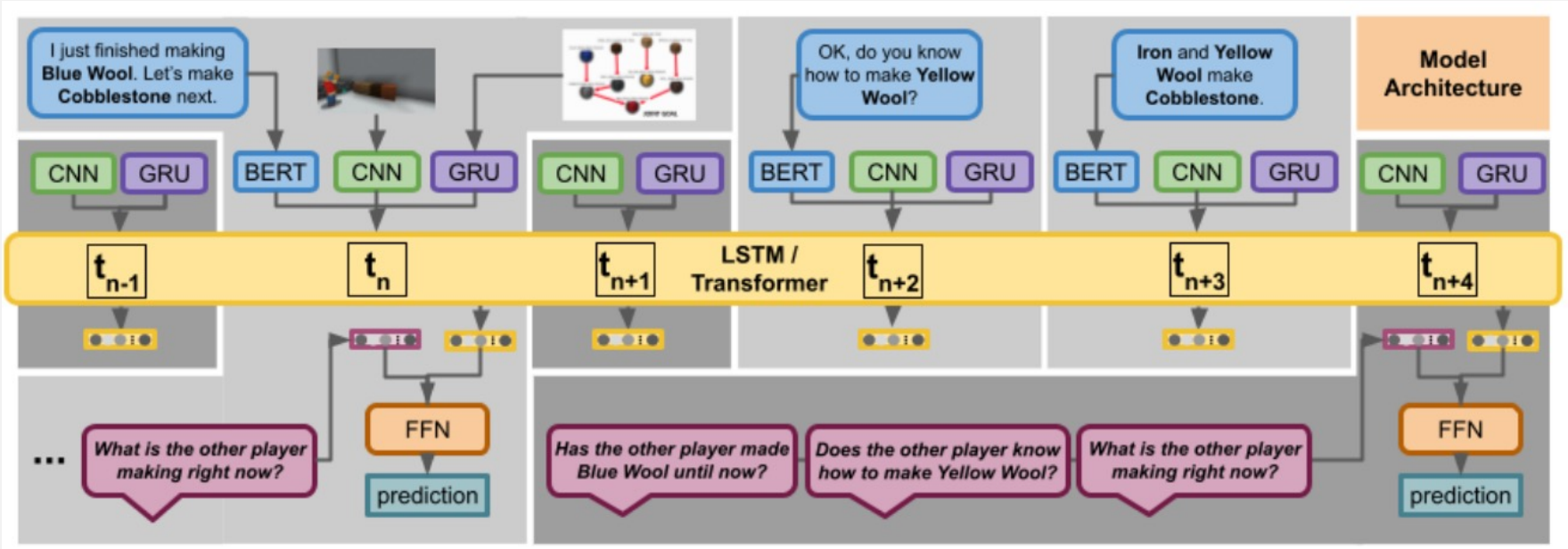
- パートナーの状態について自分がどう理解しているか (belief state) の質問が定期的に行われる
  - パートナーは○○を作り終わった？ (Yes/No/Maybe)
  - パートナーは○○の作り方を分かっているか？ (Yes/No/Maybe)
  - パートナーは今何を作っている？ (ブロック名)
- これらの質問に正しく答えられたかと、時間や行われた対話のターン数との関連性を分析
- データは100ゲーム分、平均20.5ターンほど

こうした話者の考え・精神状態を陽にモデリングしている事(mind modeling) が新規性だと主張

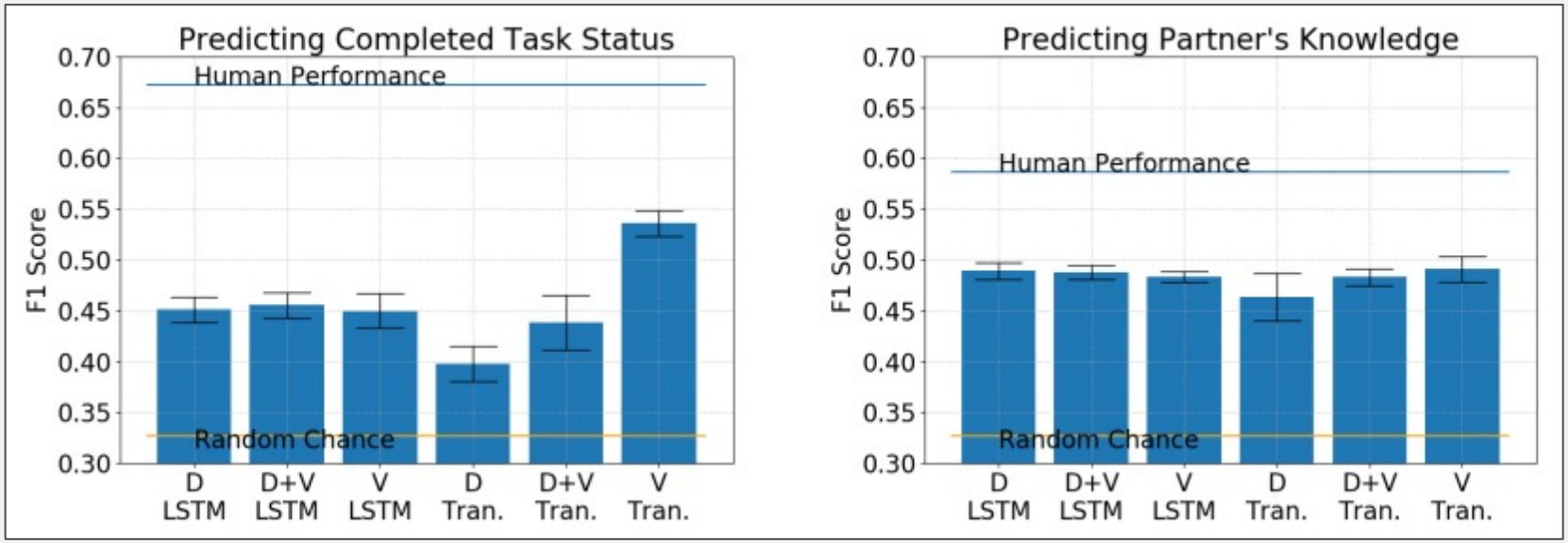
MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks (Outstanding paper)

- 構築したデータをtrain/dev/test = 6:2:2 に分割し、belief stateを推定するモデルを訓練
  - LSTM/Transformerどちらが良いか？
  - 対話・画像のどちらが効いているか？

  などについての分析を行っていた



MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks (Outstanding paper)



D: Dialogue, V: Visual

- ランダムよりは解けているが, Yes/No/Maybeの3択でMaybeをあまり選ばないだろう事を考えると…？
  - データ数 100 * 20 ターンはやはり学習が厳しい？
  - Human Performanceが結構低い