学会報告 IJCNLP-AACL 2023 (第13回自然言語処理国際共同学会・第3回計算言語学会アジア太平洋支部大会)

野元　裕樹（nomoto@tufs.ac.jp）

Luncheon Linguistics, 2023/11/29

学会の概要

- The 13th International Joint Conference on Natural Language Processing (IJCNLP)
- The 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)
  - Cf. North American Chapter (NAACL), European Chapter (EACL)
- 日程：2023年11月1–4日
- 場所：インドネシアバリ島
  グランドハイアットホテル
- 実施形態：ハイブリッド（ポスターは会場のみ）
- 大会HP：http://www.ijcnlp-aacl2023.org

この報告で伝えたいこと

1. 学会の構成
2. 個々の発表の紹介
   言語記録活動（language documentation）や言語資源開発（language resource development）に関連するもの
3. 言語学の学会との違い
   自然言語処理・計算言語学の学会の国際学会に初めて参加してみて知ったこと

学会の構成

- 1日目：ワークショップ、チュートリアル
- 2–4日目：本大会
  - Main tracks (oral, poster, findings*)
  - Keynote speeches
  - Featured plenary talks
  - System demonstrations

Findings
- ここ最近ACLで始まったものらしい
- 口頭発表・ポスター発表での選択には至らなかった論文
- 予稿集には掲載される
- Findingsのセッションでは一発表2–3分で次々と論文の紹介をしていた

ワークショップ

- 応募して採択されたもの
- 採択されたワークショップはそれぞれ発表者の募集・選抜を行う

1. Student Research Workshop
2. 3rd Workshop on NLP for Medical Conversations
3. Workshop on Information Extraction from Scientific Publications
4. The Fourth Workshop on Evaluation & Comparison of NLP Systems
5. The 11th International Workshop on Natural Language Processing for Social Media
6. The First Workshop on South East Asian Language Processing
   - 「東南アジア」とはいうものの、インドネシアとフィリピンのみ。
   - 南インドの発表が２件…
7. The 6th Workshop on Financial Technology and Natural Language Processing
8. Second Workshop on Natural Language Interfaces
9. The ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI

チュートリアル

- 応募して採択されたもの
- 日本言語学会の「ワークショップ」に相当

1. Language and Robotics: Toward Building Robots Coexisting with Human Society Using Language Interface
2. Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity
3. Practical Tools from Domain Adaptation for Designing Inclusive, Equitable, and Robust Generative AI
4. Editing Large Language Models
5. Learning WHO Saying WHAT to WHOM in Multi-Party Conversations
6. Developing State-Of-The-Art Massively Multilingual Machine Translation Systems for Related Languages

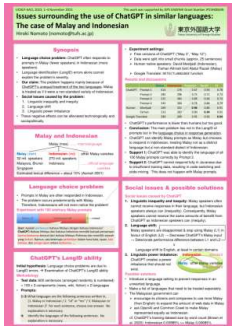
本大会プログラム

詳細はハンドブックを参照
https://github.com/IJCNLP-AACL23-Files/handbook/blob/main/IJCNLP_AACL_Handbook_2023_v5.pdf

1日目	Dialog Systems and Generation	Question Answering
	Demo	Regional Language Processing
	Demo	Findings Session
2日目	Resources and Evaluation	Data Mining, Information Extraction and Retrieval
	Poster Session	
3日目	Multilingual and Multimodal Analysis	Machine Learning and Model Interpretability
	Semantics	NLP Applications

8 個々の発表の紹介

1. 私の発表（ポスター）

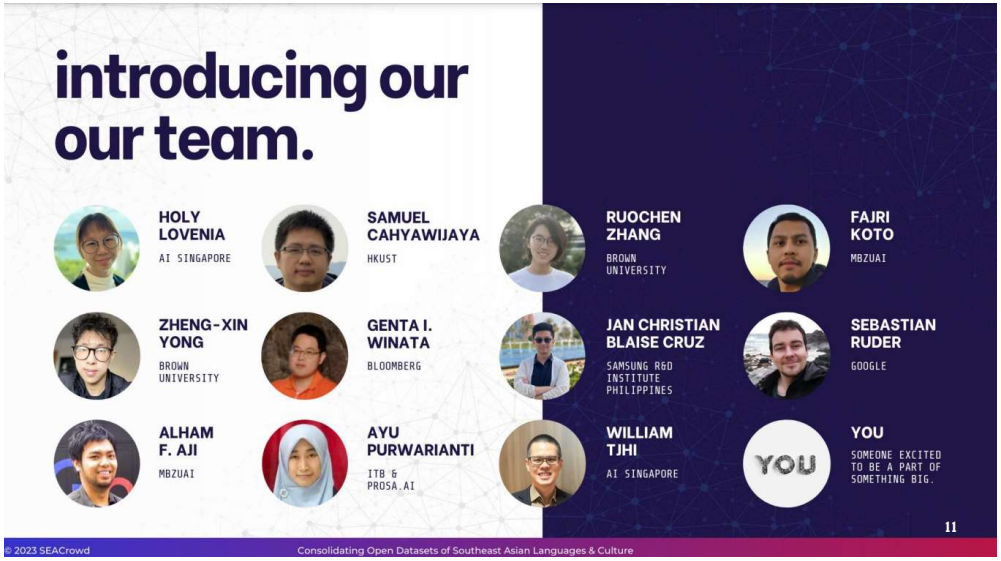


Issues Surrounding the Use of ChatGPT in Similar Languages: The Case of Malay and Indonesian

- ChatGPTにマレー語で聞くとインドネシア語で返ってくる問題
  - 実際の頻度
  - 問題が起きる理由
  - 問題がもたらす社会的・倫理的問題
  - 問題を解決する技術的・社会政治的方法
- ポスター実物は野元研究室（603）前に掲示中
  http://www.tufs.ac.jp/ts/personal/nomoto/IJCNLP-AACL2023-poster.pdf
- 予稿集論文
  http://www.afnlp.org/conferences/ijcnlp2023/proceedings/main-short/cdrom/pdf/2023.ijcnlp-short.9.pdf

2. SEA CROWD（チュートリアルから）

- 東南アジアの言語・文化の公開されたAI開発用データセットを一か所に集め、かつ標準化する試み
- チュートリアルスライド pp. 107-
  https://aacl2023-sea-nlp.github.io/assets/tutorial_slides.pdf



SEACROWDへの貢献方法

1. 既存の公開データセットの**メタデータ**を送る
2. #1のデータセットのためのデータローダーを作成する
3. （個人のPCや大学研究室のサーバーにあるとされる）**未公開のデータセットの情報を提供**する
4. 自らの未公開のデータセットを公開する（＋#3の情報提供）

3. NUSAWRITES

- 予稿集論文：http://www.afnlp.org/conferences/ijcnlp2023/proceedings/main-long/cdrom/pdf/2023.ijcnlp-long.60.pdf

NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages

Samuel Cahyawijaya[1,4*], Holy Lovenia[1,4*], Fajri Koto[3,4*], Dea Adhista[2†], Emmanuel Dave[4†], Sarah Oktavianti[2†], Salsabil Maulana Akbar[4†], Jhonson Lee[4†], Nuur Shadieq[4†], Tjeng Wawan Cenggoro[8,4†], Hanung Wahyuning Linuwih[2†], Bryan Wilie[1,4†], Galih Pradipta Muridan[2†], Genta Indra Winata[5,4†], David Moeljadi[7†], Alham Fikri Aji[3,4†], Ayu Purwarianti[6,4], Pascale Fung[1]
[1]HKUST [2]Prosa.ai [3]MBZUAI [4]IndoNLP [5]Bloomberg
[6]Institut Teknologi Bandung [7]Kanda University of International Studies [8]Bina Nusantara University
*Equal Contribution † Equal Contribution

3. NUSAWRITES

- 超低資源言語のコーパス作成は主にウェブスクレイピングや翻訳により行われてきた。
  - 利点：①効率的②低コスト
  - 欠点：①語彙的多様性がない②現地の文化と無関係な内容
- インドネシアの地方語を対象に母語話者による翻訳（NusaTranslation）とパラグラフライティング（NusaParagraph）のラベル付きコーパスを構築し、Wikipediaデータと比較して、欠点が解消されることを示した。
- 新たなベンチマーク（評価指標）NUSAWritesの構築・利用
  - 言語理解タスク（感情認識、感情分析、談話のモードの分類、トピックモデル）
  - 言語生成タスク（機械翻訳）
- データ・コード公開ページ：https://github.com/IndoNLP/nusa-writes

Language	ISO code	Status
Ambonese Malay	abs	wider communication
Mandailing	btm	threatened
Batak Toba	bbc	threatened
Betawi	bew	threatened
Bima	bhp	vigorous
Buginese	bug	wider communication
Javanese	jav	educational
Madurese	mad	developing
Makassarese	mak	threatened
Minangkabau	min	developing
Palembang / Musi	mui	wider communication
Rejang	rej	vigorous
Sundanese	sun	developing

4. MASAKHANEWS

- 予稿集論文： http://www.afnlp.org/conferences/ijcnlp2023/proceedings/main-long/cdrom/pdf/2023.ijcnlp-long.10.pdf

MasakhaNEWS: News Topic Classification for African languages

David Ifeoluwa Adelani[21], Marek Masiak[31], Israel Abebe Azime[2], Jesujoba Oluwadara Alabi[2], Atnafu Lambebo Tonja[10], Christine Mwase[4], Odunayo Ogundepo[7], Bonaventure F. P. Dossou[17,3,5], Akintunde Oladipo[7], Doreen Nixdorf, Chris Chinenye Emezue[5,33], Sana Sabah al-azzawi[23], Blessing K. Sibanda, Davis David[22], Lolwethu Ndolela, Jonathan Mukiibi[31], Tunde O. Ajayi[14], Tatiana Moteu Ngoli[27], Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka C. Obiefuna, Muhidin Mohamed[28], Shamsuddeen Hassan Muhammad[37], Teshome Mulugeta Ababu[15], Saheed S. Abdullahi[22], Mesay Gemeda Yigezu[3], Tajuddeen Gwadabe, Idris Abdulmumin[29], Mahlet Taye Bame, Oluwabusayo O. Awoyomi[25], Iyanuoluwa Shode[21], Tolulope Ana Adelani, Habiba Abdulganiy Kailani, Abdul-Hakeem Omotayo[38], Adetola Adeeko, Afolabi Abeeb, Anuoluwapo Aremu, Olanrewaju Samuel[24], Clemencia Siro[22], Wangari Kimotho[20], Onyekachi Raphael Ogbu, Chinedu E. Mbonu[39], Chiamaka I. Chukwuneke[39,35], Samuel Fanijo[39], Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge[40], Sakayo Toadoum Sari[41,35], Pamela Nyatsine, Freedmore Sidume[20], Oreen Yousuf, Mardiyyah Oduwole[42], Kanda P. Tshinu, Ussen Kimanuka[34], Thina Diko, Siyanda Nxakama, Sinodos G. Nigusie[26], Abdulmejid Tuni Johar, Shafie Abdi Mohamed[43], Fuad Mire Hassan[43], Moges Ahmed Mehamed[44], Evrard Ngabire[27], Jules Jules Ouguzoye, Ivan Ssenkungu, and Pontus Stenetorp[1]

[1]Masakhane NLP, Africa, [2]University College London, United Kingdom, [3]Stanford University, Germany, [4]Instituto Politécnico Nacional, Mexico, [5]Hugging Face, China, [6]University of Waterloo, Canada, [7]Lelapa AI, [8]McGill University, Canada, [9]Mila-Quebec AI Institute, Canada, [10]Lanfrica, [11]Technical University of Munich, Germany, [12]Luleå University of Technology, Sweden, [13]Kenyatta Data Lab, Tanzania, [14]Makerere University, Uganda, [15]Insight Centre for Data Analytics, Ireland, [16]Paderborn University, Germany, [17]Aston University, UK, [18]University of Porto, Portugal, [19]Ebo Howa University, Ethiopia, [20]Kabarak State University, Nigeria, [21]Ahmadu Bello University, Nigeria, [22]The College of Saint Rose, USA, [23]Montclair State University, USA, [24]University of California, Davis, [25]University of Rwanda, Rwanda, [26]University of Amsterdam, The Netherlands, [27]AIMS, Cameroon, [28]Nnamdi Azikiwe University, Nigeria, , [29]Lancaster University, United Kingdom, [30]Iowa State University, USA, [31]Haramaya University, Ethiopia, [32]AIMS, Senegal, [33]BIUST, Botswana, [34]MDEN, Nigeria, [35]RUSTI, Kenya, [36]Jamhuriya University, Somalia, [37]Somali National University, [38]Wuhan University of Technology, China, [39]Deutschzentrum an der Universität Burundi, Burundi

4. MASAKHANEWS

- アフリカの16言語のニューストピック分類用データセット
- データ・コード公開ページ：https://github.com/masakhane-io/masakhane-news

Language	Family/branch	Region	# speakers	News Source	# articles
Amharic (amh)	Afro-Asiatic / Ethio-Semitic	East Africa	57M	BBC	8,204
English (eng)	Indo-European / Germanic	Across Africa	1268M	BBC	5,073
French (fra)	Indo-European /Romance	Across Africa	277M	BBC	5,683
Hausa (hau)	Afro-Asiatic / Chadic	West Africa	77M	BBC	6,965
Igbo (ibo)	Niger-Congo / Volta-Niger	West Africa	31M	BBC	4,628
Lingala (lin)	Niger-Congo / Bantu	Central Africa	40M	VOA	2,022
Luganda (lug)	Niger-Congo / Bantu	Central Africa	11M	Gambuuze	2,621
Naija (pcm)	English Creole	West Africa	121M	BBC	7,783
Oromo (orm)	Afro-Asiatic / Cushitic	East Africa	37M	BBC	7,782
Rundi (run)	Niger-Congo / Bantu	East Africa	11M	BBC	2,995
chiShona (sna)	Niger-Congo / Bantu	Southern Africa	11M	VOA & Kwayedza	11,146
Somali (som)	Afro-Asiatic / Cushitic	East Africa	22M	BBC	2,915
Kiswahili (swa)	Niger-Congo / Bantu	East & Central Africa	71M-106M	BBC	6,431
Tigrinya (tig)	Afro-Asiatic / Ethio-Semitic	East Africa	9M	BBC	4,372
isiXhosa (xho)	Niger-Congo / Bantu	Southern Africa	19M	Isolezwe	24,658
Yorùbá (yor)	Niger-Congo / Volta-Niger	West Africa	46M	BBC	6,974

17

言語学の学会との違い

言語学の学会との違い①

- 参加費が高い！
- 若い人たちは指導教員や勤務先が支払うのが一般的（？）

Timeline	Early Bird (by 30 Sep 2023)	Standard (by 21 Oct 2023)	Onsite (After 21 Oct 2023)	Remark
Regular (Normal)	750	850	950	Normal rate for non-student
Student (Normal)	450	550	650	Normal rate for fulltime student
Regular (Special)	700	800	900	Not to cover a paper. Special rate is for non-student from developing countries/regions.
Student (Special)	400	500	600	Not to cover a paper. Special rate for fulltime student from developing countries/regions.
Regular (Online)	300	400	500	Online participantion, not to cover a paper
Student (Online)	200	225	250	Online participantion, not to cover a paper

18

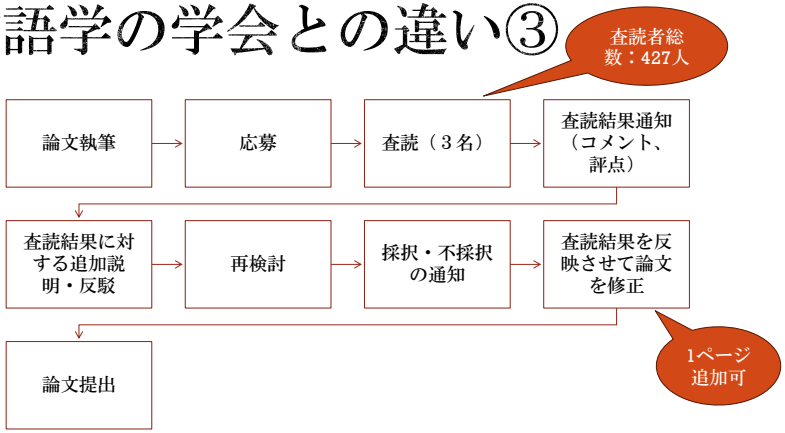
言語学の学会との違い②

- 応募先：直接 or ARR (ACL Rolling Review) という学会共通応募システム https://aclrollingreview.org
- 種目：Long Paper (8ページ＋参考文献・付録) or Short Paper (4ページ＋参考文献・付録)
- 発表形態（口頭orポスター）は内容に応じて主催者が決定

	Long			Short			計
	直接	ARR	Finding	直接	ARR	Finding	
応募数	229	19	–	109	6	–	363
採択数	72		23	22		14	131
採択率	29.0%		(38.3%)	19.1%		(31.3%)	36.1%

19

言語学の学会との違い③



20

言語学の学会との違い④

- Limitations、Ethics Statementの節が論文中で必須（ページ制限外）
  Ethics Statementの例（NusaWritesの論文より抜粋、強調は発表者による）
  - Within this work, the annotators are properly rewarded above the national average minimum wage in Indonesia.
  - We have obtained informed consent from all annotators and adhered to data protection and privacy regulations for releasing the corpus and benchmark.
  - Throughout our research process, we have made conscious efforts to engage with the language communities, involve local experts, and respect their linguistic and cultural nuances.
- Responsible NLP Checklistを論文と一緒に提出
  https://aclrollingreview.org/responsibleNLPresearch/

言語学の学会との違い⑤

- 参加者が若い
  Are you a Ph.D. student?
- 中国、インド、中東からの参加者が多い（逆に、日本からの参加者が少ない）
- 企業からの参加者もいる
  Do you work in industry or academia?
- スポンサー企業がいる
- 賞がたくさんある
- 口頭発表でNo showが結構ある