ACL Is Not an AI Conference

Emily M. Bender
Bangkok, Thailand
August 14, 2024

ACL 2024 Presidential Address

https://bit.ly/EMB-ACL24

ACL Is Not an AI Conference

https://bit.ly/EMB-ACL24

ACL is Not an AI Conference

- I am not using “AI” as a synonym for “ML”.
- Machine learning (including deep learning) provides many techniques that are useful for language technology and computational linguistics
    - (Though both of those terms are problematic.)
- The problems of CL/NLP can also be illuminating for questions about ML
- The issues that I am concerned with arise when the focus shifts to “AI”

Compling/NLP asks questions such as

- How are languages similar/different?
- How is information represented in languages?
- How can we build technology that assists with: transcription, translation, summarization, information access … in different languages?
- How can we evaluate such technology?
- What kinds of intermediate representations are useful for such technology?
- How well do different ML techniques work for different tasks?
- How do language technologies interact with existing systems of power and oppression?

AI as a research & commercial field

Asks questions like:

- How do we build “thinking machines” that can do “human-like” reasoning?
- How do we build “thinking machines” that can “surpass” humans in cognitive work?
    - (and cure cancer, solve the climate crisis, make end-of-life decisions, etc)
- How do we automate the scientific method?
- How do we automate away such creative work as painting and writing?
    - Or: How do we steal artwork at scale and try to convince people this is “for the common good”?

AI as a research & commercial field

And makes assertions like:

- Humanity’s destiny is to merge with machines and become “transhuman”
- The singularity is coming: “AGI” is inevitable and will outstrip people in all ways that matter
- “AI” (really synthetic text extruding machines) is a suitable replacement for the services we owe each other (education, healthcare, legal representation)
- All of this is inevitable and refusal is futile

AI as a research & commercial field

Suffers from multiple scourges:

- Intense (though maybe waning?) interest from venture capital
- Intense (and not waning) interest from billionaires
- The racist history and present of the notion of “intelligence”/IQ
- Intense interest from proponents of TESCREAL ideologies (Gebru & Torres 2024)

Compling/NLP asks questions such as

- How are languages similar/different?
- How is information represented in languages?
- How can we build technology that assists with: transcription, translation, summarization, information access … in different languages?
- How can we evaluate such technology?
- What kinds of intermediate representations are useful for such technology?
- How well do different ML techniques work for different tasks?
- How do language technologies interact with existing systems of power and oppression?

Language processing is a prerequisite for “AI”, but that doesn’t mean that “AI” is the only goal of CL/NLP.

The “AI” questions lead to bad research practices

- Misappropriation of benchmarks (Raji et al 2021)
- Demands to evaluate against “SOTA” closed models (Rogers 2023)
- Unmanageably large data sets (Bender, Gebru et al 2021)
  - => Lack of held-out data
- Exploitative research & development practices (Luccioni et al 2024, Hao & Seetharaman 2023; see also Fort et al 2011, Strubell et al 2019)

If your question is “How do I prove my machine is intelligent?” this distorts research practices.

Contrast with best practices in CL/NLP research

We ask:

- How well does this technique work for this purpose?
- What can we learn about human language/linguistic behavior with this model?

Contrast with best practices in CL/NLP research

We answer with:

- Well-scoped evaluations (contrast “everything machines” per Gebru & Torres 2024)
- Intrinsic & extrinsic evaluations
  - Extrinsic ideally reflecting situated use cases
- Solid baselines
- Held-out test data
- Detailed error analysis
  - Ideally including consideration of impacts of different error types

Contrast with best practices in CL/NLP research

Grounded in understanding of our data:

- Knowledge of how languages work (i.e. linguistics)
  - Shameless plug of “100 things” books: Bender 2013, Bender & Lascarides 2019
- Dataset documentation (Bender & Friedman 2018)
  - Including naming the language(s) studied
  - See also Gebru et al 2018, 2021; McMillan-Major et al 2021; Bender et al 2021 inter alia

Contrast with best practices in CL/NLP research

Efforts towards replicability and reproducibility

- Not a new problem! (See Fokkens et al 2013, Fokkens 2017)
- But much, much worse with closed, commercial models
  - Claims of emergence are ascientific without access to training data (Rogers 2024, Rogers & Luccioni 2024)
  - Note that open-weights is not sufficient (and should not be called “open source”; Solaiman 2023, Lisenfield & Dingemanse 2024)
- We know that science is about building on previous research, not just climbing over each other to get to the top of the SOTA pile
  - Building on previous research requires open science

Contrast with best practices in CL/NLP research

With an eye towards societal impacts

- Ethics and NLP research goes back (at least) to Fort et al 2011
    - Journée ATALA in 2014
    - EACL workshop 2017
- Keeping the people in the frame
    - “The L in NLP is language, language means people” (Schnoebelen 2017)
    - Understanding the languages and their communities beyond the datasets (Bird & Yibarbuk 2024)
    - Who will the technology be used by/for/on, and who might be harmed, by being excluded -- or included? (Bender & Grissom II 2024)

“AI” focus leads to poor reviewing practices

Papers dismissed as uninteresting if they:

- Don’t use LLMs
- Don’t provide results from LLM with SOTA size
- Involve careful, detailed work on a specific language
- Give only carefully scoped claims

ACL 2024 papers that have nothing to do with “AI”

- The Thai Discourse Treebank: Annotating and Classifying Thai Discourse Connectives (TACL)
- Feriji: A French-Zarma Parallel Corpus, Glossary & Translator (SRW)
- Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke (SRW)
- Wav2Gloss: Generating Interlinear Glossed Text from Speech (Main)
- DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages (Main)
- PyFoma: a Python finite-state compiler module (Demo)
- Z-coref: Thai Coreference and Zero Pronoun Resolution (SRW)
- Automatic Derivation of Semantic Representations for Thai Serial Verb Constructions: A Grammar-Based Approach (SRW)

ACL is historically and should remain:

- A venue for people who care about the *language* in language technology
- A community that fosters interdisciplinarity
- A research field that cares about language communities
- … and, as a result, a space where we can reason about societal impacts of our research and technology

https://bit.ly/EMB-ACL24

References

Bansal, V. (2024). Automatic derivation of semantic representations for Thai serial verb constructions: A grammar-based approach. In Fu, X. and Fleisig, E., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 422–437, Bangkok, Thailand. Association for Computational Linguistics.

Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool.

Bender, E. M., Freidman, B., and McMillan-Major, A. (2021a). A guide for writing data statements for natural language processing.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., and et al (2021b). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of FAccT 2021*.

Bender, E. M. and Grissom II, A. (2024). Power shift: Toward inclusive natural language processing. In *Inclusion in Linguistics*. Oxford University Press.

Bender, E. M. and Lascarides, A. (2019). *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Morgan & Claypool.

Bird, S. and Yibarbuk, D. (2024). Centering the speech community. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., and Anastasopoulos, A. (2024). DIALECTBENCH: A NLP benchmark for dialects, varieties, and closely-related languages. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Fokkens, A. (2017). Slowly growing offspring: Zigglebottom anno 2017 — guest post. Guest post on COLING 2018 PC blog, available at `https://coling2018.org/slowly-growing-offspring-zigglebottom-anno-2017-guest-post/`.

Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., and Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.

Fort, K., Adda, G., and Cohen, K. B. (2011). Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):8692.

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2018). Datasheets for datasets. arXiv:1803.09010v1.

Gebru, T. and Torres, E. P. (2024). The tescreal bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4).

Hao, K. and Seetharaman, D. (2023). Cleaning up ChatGPT takes heavy toll on human work. *The Wall Street Journal*, July 24, 2023.

He, T., Choi, K., Tjuatja, L., Robinson, N., Shi, J., Watanabe, S., Neubig, G., Mortensen, D., and Levin, L. (2024). Wav2Gloss: Generating interlinear glossed text from speech. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

Hulden, M., Ginn, M., Silfverberg, M., and Hammond, M. (2024). PyFoma: a python finite-state compiler module. In Cao, Y., Feng, Y., and Xiong, D., editors, *Proceedings of the 62nd Annual Meeting of* *the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 258–265, Bangkok, Thailand. Association for Computational Linguistics.

Keita, M., Ibrahim, E., Alfari, H., and Homan, C. (2024). Feriji: A French-Zarma parallel corpus, glossary & translator. In Fu, X. and Fleisig, E., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–9, Bangkok, Thailand. Association for Computational Linguistics.

Liesenfeld, A. and Dingemanse, M. (2024). Rethinking open source generative ai: open-washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 17741787, New York, NY, USA. Association for Computing Machinery.

Luccioni, S., Jernite, Y., and Strubell, E. (2024). Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 8599, New York, NY, USA. Association for Computing Machinery.

Mainzinger, J. and Levow, G.-A. (2024). Fine-tuning ASR models for very low-resource languages: A study on mvskoke. In Fu, X. and Fleisig, E., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 170–176, Bangkok, Thailand. Association for Computational Linguistics.

McMillan-Major, A., Osei, S., Rodriguez, J. D., Ammanamanchi, P. S., Gehrmann, S., and Jernite, Y. (2021). Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In Bosselut, A., Durmus, E., Gangal, V. P., Gehrmann, S., Jernite, Y., Perez-Beltrachini, L., Shaikh, S., and Xu, W., editors, *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

Prasertsom, P., Jaroonpol, A., and Rutherford, A. T. (2024). The Thai discourse treebank: Annotating and classifying Thai discourse connectives. *Transactions of the Association for Computational Linguistics*, 12:613–629.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). AI and the everything in the whole wide world benchmark.

Rogers, A. (2023). Closed ai models make bad baselines. Blog post, `https://hackingsemantics.xyz/2023/closed-baselines/`.

Rogers, A. (2024). A sanity check on 'emergent properties' in large language models. Blog post, `https://hackingsemantics.xyz/2024/emergence/`.

Rogers, A. and Luccioni, S. (2024). Position: Key claims in LLM research have a long tail of footnotes. In *Forty-first International Conference on Machine Learning*.

Schnoebelen, T. (2017). The carrots and sticks of ethical NLP. Blog post, `https://medium.com/@TSchnoebelen/ethics-and-nlp-some-further-thoughts-53bd7cc3ff69`.

Solaiman, I. Generative AI systems aren't just open or closed source. *WIRED*.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Suwannapichat, P., Tarnpradab, S., and Prom-on, S. (2024). Z-coref: Thai coreference and zero pronoun resolution. In Fu, X. and Fleisig, E., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 132–139, Bangkok, Thailand. Association for Computational Linguistics.