The Long and Winding Road of NLP: Reflections on the Last 30 Years

Timothy Baldwin

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Talk Outline

1. My Connection to ANLP
2. The Long and Winding Road of NLP: Reflections on the Last 30 Years
   - What has Improved?
   - What has Stayed the Same?
   - What has Regressed?
   - Signs of the Times
3. Where does This Leave Us?

Personal NLP Beginnings

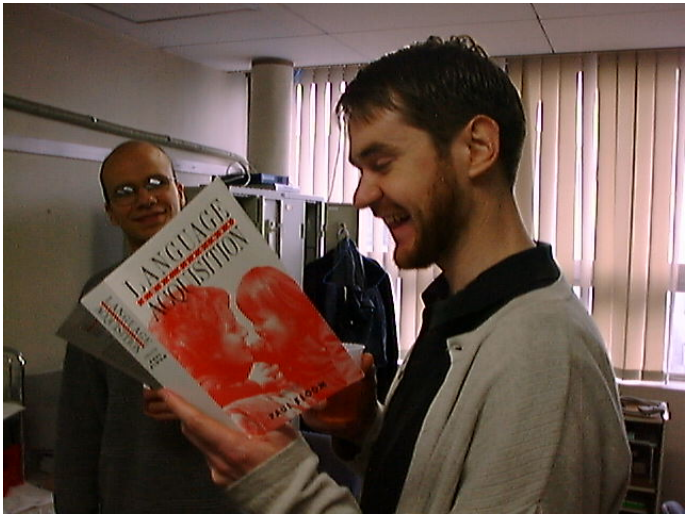
- My first humble beginnings in NLP were in building a reversible Japanese↔English machine translation system based on Montague semantics as an undergraduate at The University of Melbourne in 1994, with a vocabulary of ∼200 words
- I came to Japan as a research student in 1995 in the Tanaka–Tokunaga Lab of Tokyo Institute of Technology (TITech)

Source(s): https://www.unixuser.org/~euske/doc/memorial-prof-tanaka/pics1/tanaka.jpg

Personal ANLP Beginnings

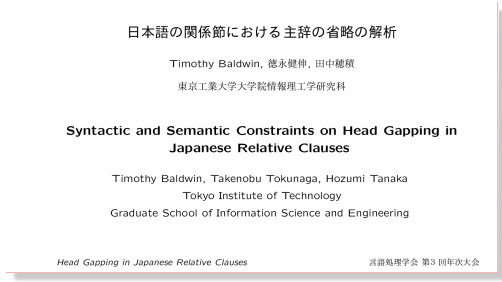
- I attended my first NLP conference on the TITech campus in 1996 (言語処理学会第2回年次大会)
  - 108 papers
  - ∼150 attendees
  - three parallel sessions
  - invited speaker = Key-Sun Choi

Personal ANLP Beginnings



Personal ANLP Beginnings

- My first JNLP conference publication was in 1997 in Kyoto (言語処理学会第3回年次大会):



日本語の関係節における主辞の省略の解析

Timothy Baldwin, 徳永健伸, 田中穂積
東京工業大学大学院情報理工学研究科

Syntactic and Semantic Constraints on Head Gapping in Japanese Relative Clauses

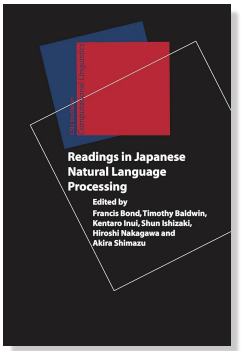
Timothy Baldwin, Takenobu Tokunaga, Hozumi Tanaka
Tokyo Institute of Technology
Graduate School of Information Science and Engineering

  - 144 papers
  - >200 attendees
  - four parallel sessions

Source(s): Baldwin et al. (1997)

Subsequent ANLP Engagement

- I published a total of 9 papers in JNLP conferences over the years (1997–1999, 2001–2002, 2004), including one best paper award in 2002
- I was also involved in setting up a (one-book) series on Japanese Computational Linguistics at CSLI Publications:



Talk Outline

1. My Connection to ANLP
2. The Long and Winding Road of NLP: Reflections on the Last 30 Years
   - What has Improved?
   - What has Stayed the Same?
   - What has Regressed?
   - Signs of the Times
3. Where does This Leave Us?

Contents

Automatic Evaluation

- Automatic evaluation was only just starting to emerge in 1995 (and largely for tasks such as POS tagging, constituency parsing, and IE template filling, over discrete representations using precision and recall); outside these limited areas (e.g. in MT) evaluation was:
  - manual
  - small-scale
  - largely over in-house datasets
- **The good:** Automatic evaluation metrics, combined with standardised evaluation datasets and automatic training methods, have revolutionised fields such as machine translation and summarisation, driving real progress

Automatic Evaluation

- **The bad:** BUT they have also come with downsides such as:
  - surprisingly incompatibilities in numbers based on low-level implementation details (Post, 2018; Deutsch and Roth, 2020)
  - over-reliance/interpretation of noisy metrics (Mathur et al., 2020; Koto et al., 2022)
  - "hill climbing" behaviours

Natural Language Generation (NLG)

- In 1995, generation was still in its infancy and based primarily on symbolic and rule/template/example-based methods (for text-to-text generation tasks) or planning (for more open-ended generation tasks)
- Methods based on *n*-gram language models soon followed (Brown et al., 1993; Langkilde and Knight, 1998), but generating text of any length was largely intractable
- For more than a decade, NLP work on language modelling (which drove NLG) focused on *n*-gram smoothing methods and data availability (Chen and Goodman, 1996)
- Neural language models emerged out of ML and Speech (Mikolov et al., 2010; Collobert et al., 2011), and revolutionised NLG

Natural Language Generation (NLG)

- **The Good:**
  - data and parameter efficiency
  - fluency
  - generalisability
  - flexibility
- **The Bad:**
  - lack of controllability(?)
  - lack of interpretability
  - less clear how/when to incorporate planning into NLG, although important work is emerging in this regard

Multilinguality

- In 1995, multilingual systems were either built pairwise for individual language pairs (and directions!), or assumed "interlingua" which were very hard to scale/generalise to typologically distinct language combinations
- The assumption in NLP was very much that for multilingual tasks, preprocessors such as language identification (Jauhiainen et al., 2019) needed to be used to identify individual languages, to "route" processing to language-specific components ... the idea of a multilingual language model was inconceivable (other than in the speech community: Ward et al. (1998); Wang et al. (2002))
- What little work did exist tended to focus on (bilingual) transfer from a high-resource language to a lower-resource language (Snyder et al., 2008; Lacroix et al., 2016)

Multilinguality

- **The Good:**
  - cross-lingual (task) transferability
  - data and parameter efficiency
  - flexibility
- **The Bad:**
  - entanglement of closely-related languages/dialects
  - controlling language switching effects
  - cultural bias from “large” languages

Open Source

- In 1995, open-source NLP (and ML) resources were scarce, meaning most things had to be built from scratch
- Japan was a pioneer in developing a culture of open-source tool development, including projects such as JUMAN (Kurohashi and Nagao, 1998), ChaSen (Matsumoto et al., 1999), and TinySVM, and pioneering systems elsewhere included the Brill tagger (Brill, 1995), Ratnaparkhi POS tagger (Ratnaparkhi, 1996), Charniak parser (Charniak, 2000), and Stanford parser (Klein and Manning, 2003)

Open Source

- **The Good:**
  - reproducibility
  - modularity
  - tractability
- **The Bad:**
  - black-box usage

Dataset Development & Sharing

- Along similar lines, there was very little culture of open sharing of datasets, and what datasets did exist tended to be very small in scale, making it very hard to perform like-for-like evaluation
- Notable exceptions to this were the Penn Treebank (Marcus et al., 1993) and MUC datasets (Grishman and Sundheim, 1996), which were major catalysts of innovation in POS tagging & parsing, and information extraction
- As part of the sharing of datasets, has been sharing of the resource-creation "recipe", promoting resource curation in new languages and domains
- Open access of datasets also allows for critical evaluation of their composition/validity (Gururangan et al., 2018; Gardner et al., 2021)

Dataset Development & Sharing

- **The Good:**
  - more training resources
  - more task variety
  - more comprehensive/varied evaluation
  - greater reproducibility of evaluation
  - ability to critically evaluate datasets
- **The Bad:**
  - fixation on (at times flawed) popular datasets
  - complacency re the art/complexity/cost of annotation
  - risk of data contamination

General-purpose/End-to-end Tools

- In "early" NLP, tools were generally very limited in scope of use and input data modality, noisy, needed to be pipelined together, and/or were varyingly divorced from downstream tasks
- With the advent of LLMs and prompting, we suddenly find ourselves being able to perform a wide variety of tasks with a single model
- While they do not always work as well as they superficially appear to, the breadth of tasks they can potentially be applied to is unprecedented (e.g. identifying and mapping racial covenants in CA: `https://tinyurl.com/hai-convenant`)

General-purpose/End-to-end Tools

- **The Good:**
  - flexibility/generalisability
  - robustness
  - scalability
- **The Bad:**
  - model overconfidence
  - hallucinations
  - model refusals
  - lack of model consistency

Shared Tasks

- Another thing that evolved since 1995 is "shared tasks" (i.e. novel datasets that are put together as a proxy for a particular task, and where systems synchronously compete over held-out data), which evolved in a number of different forums: MUC (Sundheim, 1991), CoNLL (Tjong Kim Sang and Buchholz, 2000), SENSEVAL (Kilgarriff and Rosenzweig, 2000)
- These were inspired in part by TREC and its defining impact on information retrieval (Harman, 1992)
- Shared tasks helped establish new tasks (e.g. chunk parsing, NER, STS, ...) and descriptive/evaluation paradigms for them, and establish state-of-the-art tools/methodologies

Shared Tasks

- **The Good:**
  - establishing/catalysing research on new tasks
  - providing standard evaluation benchmarks for open research
  - lowering the bar to entry in engaging on new tasks
- **The Bad:**
  - in some instances, the "canonical" shared task version of a task is biased/limited in various ways (e.g. CoNLL NER)
  - equally, the framing/evaluation of a task can get stuck in a local minimum because of shared tasks (and it is hard — even for the organisers themselves — to break away from this)

AI Safety/Responsible AI

- One research area which was never discussed in NLP until very recently is AI safety/responsible AI, i.e. safeguarding NLP models to operate within pre-defined cultural/content/value-based norms
- Possible (cynical) takes on this are that: (a) NLP wasn't real-world relevant until recently for this to be an issue; or (b) it is only with the advent of hyper-parameterised, black-box LLMs that it has become an issue because of their unpredictability
- To me, this represents the realisation that if our models are to be used in the wild for all sorts of tasks beyond what we ourselves imagined, then we want to be confident that they can't cause harm, perpetuate/amplify biases, and are equitable in utility to all

AI Safety/Responsible AI

- My personal moment of "awakening" came in the context of `langid.py` (Lui and Baldwin, 2012) and analysis of Jurgens et al. (2017) into dialectal bias:

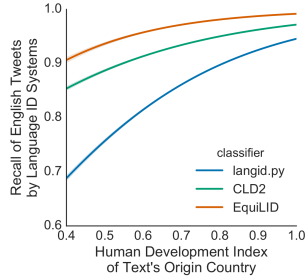


Figure 2: Estimated recall of tweets with health-related terms according to a logit regression on the Human Development Index of the tweet's origin country; bands show 95% confidence interval.

AI Safety/Responsible AI

- As LLMs and other generative AI models become mainstream, the power of the models for good and bad increases, as does the complexity of evaluating and imposing meaningful safeguards
- The onus of responsibility is even greater if we are to maintain a culture of releasing open-source/"open-weight" models, as once they are in the wild, they cannot be retracted/fixed
- To me, this encompasses the evaluation and safety alignment wrt different forms of potential "harm" (Wang et al., 2024a), localised to regional/cultural/political/religious/... sensitivities (Wang et al., 2024b), in a robust manner (to both false positives and false negatives); it also includes mitigation of a variety of model biases (Zhao et al., 2017; Han et al., 2023b,a)

AI Safety/Responsible AI

- **The Good:**
  - self-realisation of our "social/societal accord" = maturation of the field
  - (successful) work in this spaces requires a lot of detailed understanding of the models we are building
  - very hard research problems to be tackled, with high stakes
  - critical interfaces with people outside NLP (in psychology, sociology, law, philosophy, anthropology, ...)
- **The Bad:**
  - (successful) work in this spaces requires a lot of detailed understanding of the models we are building
  - very hard research problems to be tackled, with high stakes
  - critical interfaces with people outside NLP (in psychology, sociology, law, philosophy, anthropology, ...)

Cultural NLP

- Along similar lines, there was no dialogue relating to cultural wrt the "old" models we were building ... possibly for similar reasons
- As our models are being used by more diverse populations for more diverse purposes, it's critical that that are embedded with cultural sensitisation/understanding that is relevant to those users (Hershcovich et al., 2022; Liu et al., 2024)
- This is still a very young area of NLP, but critically important if we are serious about the "reach" of the field and true multilinguality

Cultural NLP

- **The Good:**
  - Very young field: lots to be done!
  - A critical area in terms of striving towards techno-social equity in NLP
- **The Bad:**
  - What is culture?
  - Does it make sense to consider culture in the isolated context of text?

Branching out beyond Generic "Language"

- 30 years ago, there was very little talk of specialist domains or interdisciplinary work with other research fields, whereas now areas like BioNLP, clinical NLP, LegalNLP, computational social science, and computational psycholinguistics are mainstays of NLP
- Equally, there is very exciting work building off "language" models to multimodal models (graphs, images, video, speech, time series numeric data, geospatial data, ...), biological foundation models, foundation models for robotics, agentic language models, ...

Branching out beyond Generic "Language"

- **The Good:**
  - we are at the starting line for what "language" models can really achieve
- **The Bad:**
  - the stakes get considerably higher, and safety margins smaller as we integrate language models with real-world applications

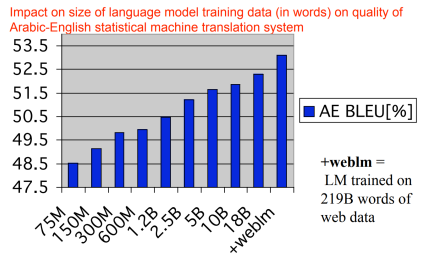
Contents

Data and Compute Scalability

- While there is a lot of discussion about the "data scalability" of current-generation LMs and the need for more and more high-quality data, this is not the first time we have met with this "data scaling crisis"
- The first iteration of it was in the context of "googleology" (Kilgarriff, 2007) in the mid-2000s, i.e. the use of page counts from commercial search engines to estimate token frequencies in a range of disambiguation tasks which could be lexically fingerprinted (Keller and Lapata, 2003; Nakov and Hearst, 2005)

Data and Compute Scalability

- Around the same time, at the peak of statistical machine translation, there was a race to source as much data as possible to train effective *n*-gram LMs (Och, 2005; Saphra et al., 2024), a race where the commercial groups had an unfair advantage:



Data and Compute Scalability

- This has artfully been described as the "Bitter Lesson" (Sutton, 2019):

  We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

Data and Compute Scalability

- What was the outcome of that story? data scaling didn't work for many language pairs in SMT, and advances in computation and learning algorithms in the form of NMT ultimately meant that web-scale SMT was a local minimum
- Saphra et al. (2024) is a wonderful read in terms of the lessons that can be learned from this era, with the following recommendations: (1) scale is supreme; (2) evaluation is a bottleneck; (3) there is no gold standard; (4) progress is not continuous; and (5) do research

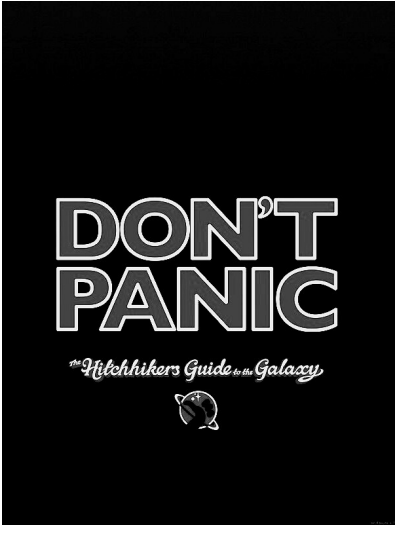
The Academic Intractability of Commercial-grade NLP

- A related lament that is not new is that academic NLP research cannot compete with commercial-grade NLP research
- This was (perhaps even moreso) the case in the era of rule-based MT systems, where lexical resources were the major differentiator (the "linear accelerators" of NLP at the time), and it was impossible to compete with the budgets of commercial players

The Academic Intractability of Commercial-grade NLP

- Similar to the SMT vs. NMT story, history shows that rule-based MT was a local minimum (given sufficient training data), but there are deeper similarities with what is happening now with LLMs:
  - as the number of rules grew, systems became more impenetrable/black-box and unmaintainable
  - given their complexity, in order to run the systems in reasonable time, specialist compute was required
  - clever engineering by scrappy start-ups resulted in the release of MT systems that were a fraction of the cost and run on commodity hardware ... killing off much of the high-end rule-based MT industry (and EBMT and SMT finished the job)

The Academic Intractability of Commercial-grade NLP

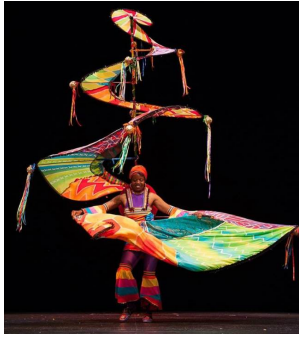


Generalisability and Reproducibility

- Two constant concerns in NLP research have been:
  - generalisability (across unseen data, novel domains, etc), e.g. in the form of domain adaptation (Daumé III, 2007; Ramponi and Plank, 2020)
  - reproducibility of research findings (under the same data conditions) (Fokkens et al., 2013; Belz et al., 2021)
- As we have shifted as a field in computational paradigms (e.g. symbolic $\rightarrow$ statistical $\rightarrow$ pre-trained neural $\rightarrow$ generative neural), the particular sensitivities wrt these issues has changed, as have standard paradigms for dealing with/evaluating them, but they have (rightly) never gone away

Spiralling under Control?

[We] can see old ideas reappearing in new guises ... But the new costumes are better made, of better materials, as well as more becoming: so research is not so much going round in circles as ascending a spiral, if only a rather flat one.



Source(s): Sparck Jones (1994); image attribution = `https://tinyurl.com/spiral-costume`

Contents

The Ability to Comprehend Data Nuances

- While I am in no way suggesting a return to the "good olde times" of rule-based or symbolic NLP and ad hoc evaluation datasets, one thing it did teach was "deep data engagement" via the necessarily hands-on nature of the research methodology
- This fed into the development of a highly-refined skillset for understanding methodological limitations and being able to diagnose/categorise errors
- As part of an increasing shift towards end-to-end trained models and auto-eval, we are losing the ability/motivation to look carefully at our datasets/system outputs to diagnose errors, and an unhealthy level of disengagement from the core formulation of research tasks

The Ability to Comprehend Data Nuances

- **What to do about it?** Encourage a greater culture of data engagement, in terms of:
  - thinking critically about the task that is being performed, and whether a given dataset is aligned with it or not
  - thinking critically about dataset labelling, how it has been done, and what it captures
  - thinking critically about exactly what different evaluation metrics are measuring, and whether this is an accurate reflection of the true task
  - developing deeper skills in error analysis of system outputs, and the ability to reason about *why* a system is behaving in the way it is

Lack of Modularity

- As powerful as the end-to-end training and zero-/few-shot prompting are, a natural consequence of current research methodology is monolithic black-box models
- While I am not evangelising for a return to traditional NLP pipelines (with all of the complexities and issues with error propagation that entailed), there is potential value in more modular architectures in terms of being to plug-and-play sub-models with specific functionality in terms of:
  - parameter localisation/efficiency
  - open-research collaboration/scalability, esp. for academic research
  - modular interpretability
- Modern neural architectures make this more methodologically tractable (in terms of model integration/combined parameter updates) than was traditionally the case

The Science of Annotation/Dataset Construction

- There is insatiable appetite for ever-larger datasets to train/robustly evaluate ever-larger models on, which are inevitably more expensive to construct
  $\Rightarrow$ *the barrier to dataset construction is getting higher and higher*
- Paired with this, as more and more large-scale datasets are made available and accessible via (amazing) sites like HuggingFace, the more they are taken for granted
  $\Rightarrow$ *the value placed on dataset construction is getting lower*
- Combined, this means that dataset construction is increasingly happening via "large science" projects which can be hard to break into the inner circle of, at the same time as "dataset complacency", with the corollary that less individuals are involved in the core design and curation of major datasets

The Science of Annotation/Dataset Construction

- Through this effect, the proportion of the field that is engaged in dataset construction is falling, and the skills of designing and managing such datasets are being lost
- At the same time, amazing work is being done on:
  - how to label/train models in a more "human-authentic" way (Plank et al., 2014; Baan et al., 2023; Wang et al., 2022)
  - standardising definitions of evaluation and documentation of annotation methodology (Howcroft et al., 2020)
- Related to the point about data nuances, important to ensure that more people are engaged in dataset construction, and that more value is ascribed to these activities

Regression Testing

- For those engaged in symbolic/rule-based NLP, there was a strong culture of "regression testing" (= checking that nothing has been unintentionally "broken" in making a particular set of changes, wrt standard evaluation datasets)
- As part of this, a common practice was to pairwise compare the outputs of different versions of the same system over a standard dataset, to measure (potentially very fine-grained) progress over time (Lehmann et al., 1996)
- The modern-day equivalent with large models is to benchmark against a range of different datasets, and use LLM "arenas" to perform aggregate-level evaluation, given the lack of instance-level control

Regression Testing

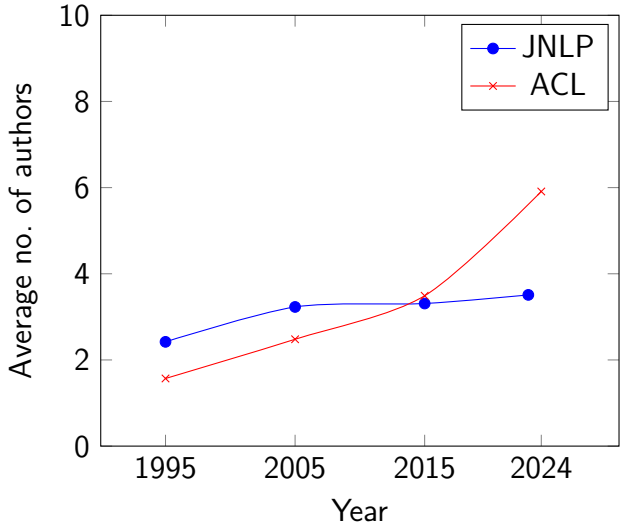
- What can be done by model developers with the fine-grained feedback that comes from regression testing is a somewhat open question, but there was great value in the richness of feedback that came from this style of testing, that we should be looking back at

Contents

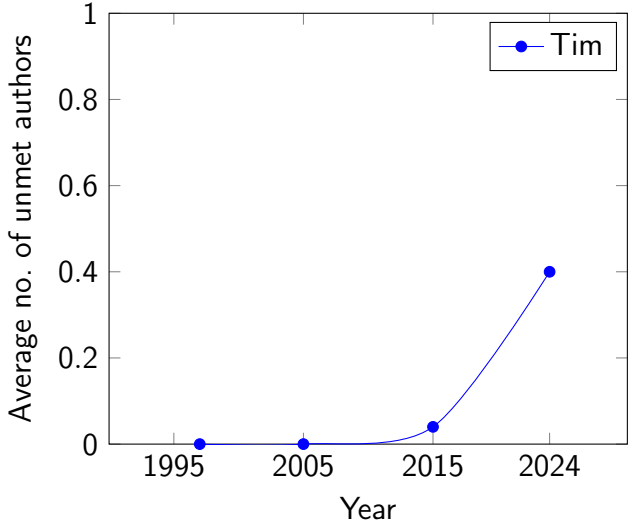
Signs of the Times

- Somewhat as an aside, I spent a bit of time playing around looking at some potentially interesting data points in the proceedings of JNLP vs. ACL (short and long papers) over the years, in terms of:
  - the average number of authors per paper over time
  - the proportion of co-authors who I had not met at the time of publication, for papers in a given year

Average Number of Authors per Paper



Average Number of Unmet-at-time-of-submission Authors



What does This Tell Us?

- Big science/industrial research is increasing in NLP, coming off a lower base in ACL than JNLP (there was stronger representation from industry in the early days of JNLP than in ACL at the same time)
- Some combination of: (a) Tim is getting old/more senior/more distracted; and (b) our field is moving towards more distributed collaboration models

Talk Outline

1. My Connection to ANLP
2. The Long and Winding Road of NLP: Reflections on the Last 30 Years
   - What has Improved?
   - What has Stayed the Same?
   - What has Regressed?
   - Signs of the Times
3. Where does This Leave Us?

Where does This Leave Us?

- NLP is (mostly) in rude health, and has made great progress in the last 30 years across many dimensions
- There are always lessons to be learned from the past, esp. in terms of how the field has evolved (or not) and the issues it has faced
- Regional associations such as 言語処理学会 have played a major role in this progress, and are critical to the further development of the field
- Congratulations once again on the first 30 years (and thanks for all the fish)!

References

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Timothy Baldwin, Takenobu Tokunaga, and Hozumi Tanaka. 1997. Syntactic and semantic constraints on head gapping in Japanese relative clauses. In *Proceedings of the Third Annual Meeting of the Japanese Association for Natural Language Processing*, pages 277–280, Kyoto, Japan.

Anja Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393.

References

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Eugene Charniak. 2000. A maximum entropy-based parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2000)*, Seattle, USA.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, Santa Cruz, USA.

References

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic.

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online. Association for Computational Linguistics.

References

Antske Fokkens, Marieke Van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah A Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference – 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pages 466–471, Copenhagen, Denmark.

References

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2023a. Everybody needs good neighbours: An unsupervised locality-based method for bias mitigation. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2023b. Fair enough: Standardizing evaluation and model selection for fairness research in NLP. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–312, Dubrovnik, Croatia. Association for Computational Linguistics.

References

D.K. Harman, editor. 1992. *NIST Special Publication 500-207: The First Text REtrieval Conference (TREC-1)*. Department of Commerce, National Institute of Standards and Technology.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

References

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Volume 2: Short Papers*, pages 51–57.

References

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Adam Kilgarriff. 2007. Last words: Googleology is bad science. *Computational linguistics*, 33(1):147–151.

Adam Kilgarriff and Joseph Rosenzweig. 2000. English Senseval: Report and results. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1239–1244, Athens, Greece.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 423–430, Sapporo, Japan.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. FFCI: A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, 73:1553–1607.

References

Sadao Kurohashi and Makoto Nagao. 1998. *Nihongo keitai-kaiseki sisutemu JUMAN* [Japanese morphological analysis system JUMAN] version 3.5. Technical report, Kyoto University. (in Japanese).

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1058–1063.

Irene Langkilde and Kevin Knight. 1998. The practical value of N-grams in generation. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 248–255.

References

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Lorna Balkan, and Doug Arnold. 1996. TSNLP — test suites for natural language processing. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

References

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4984–4997.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, and Yoshitaka Hirano. 1999. *Japanese Morphological Analysis System ChaSen Version 2.0 Manual*. Technical Report NAIST-IS-TR99009, NAIST.

References

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048, Makuhari, Japan.

Preslav Nakov and Marti Hearst. 2005. Search engine statistics beyond the *n*-gram: Application to noun compound bracketting. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 17–24, Ann Arbor, USA.

Franz Josef Och. 2005. Statistical machine translation: Foundations and recent advances. In *Proceedings of Machine Translation Summit X: Tutorial notes*, Phuket, Thailand.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

References

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP — a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96)*, Philadelphia, USA.

Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. 2024. First tragedy, then parse: History repeats itself in the new era of large language models. *arXiv preprint arXiv:2311.05020*.

References

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050.

Karen Sparck Jones. 1994. Natural language processing: A historical review. In Antonio Zampolli, Nicoletta Calzolari, and Martha Palmer, editors, *Current issues in computational linguistics: in honour of Don Walker*. Linguistica Computazionale.

Beth M. Sundheim. 1991. Overview of the third Message Understanding Evaluation and Conference. In *Third Message Understanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991*.

Rich Sutton. 2019. The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`.

References

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of the 4th Conference on Computational Natural Language Learning (CoNLL-2000)*, Lisbon, Portugal.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024a. Do-Not-Answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

References

Yuxia Wang, Zenan Zhai, Haonan Li, Xudong Han, Shom Lin, Zhenxuan Zhang, Angela Zhao, Preslav Nakov, and Timothy Baldwin. 2024b. A Chinese dataset for evaluating the safeguards in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3106–3119, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhirong Wang, U. Topkara, T. Schultz, and A. Waibel. 2002. Towards universal speech recognition. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 247–252.

Todd Ward, Salim Roukos, Chalapathy Neti, Jerome Gros, Mark Epstein, and Satya Dharanipragada. 1998. Towards speech understanding across multiple languages. In *Fifth International Conference on Spoken Language Processing*.

References

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2979–2989.