Some Useful Things to Know When Combining IR and NLP: The Easy, the Hard and the Ugly

Omar Alonso
Amazon
Palo Alto, CA, USA
omralon@amazon.com

Kenneth Church
Northeastern University
San Jose, CA, USA
k.church@northeastern.edu

ABSTRACT

Deep nets such as GPT are at the core of the current advances in many systems and applications. Things are moving fast; techniques become obsolete quickly (within weeks). How can we take advantage of new discoveries and incorporate them into our existing work? Are new developments radical improvements, or incremental repetitions of established concepts, or combinations of both?

In this tutorial, we aim to bring interested researchers and practitioners up to speed on the recent and ongoing techniques around ML and Deep learning in the context of IR and NLP. Additionally, our goal is to clarify terminology, emphasize fundamentals, and outline problems and new research opportunities.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Information retrieval**;

KEYWORDS

Information Retrieval, Natural Language Processing, Fine-tuning, Prompt engineering, Evaluation and benchmarks

ACM Reference Format:
Omar Alonso and Kenneth Church. 2024. Some Useful Things to Know When Combining IR and NLP: The Easy, the Hard and the Ugly. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3616855.3636452

1 INTRODUCTION

Deep nets such as ChatGPT (Generative Pre-trained Transformers) have taken center stage and created rapid adoption of many applications. With the ever-increasing acceptance of machine learning and deep nets in many information seeking systems, the "tech stack" is undergoing constant change. We believe it is useful to take a step back and look at fundamentals. The goal is to identify useful things to know when working with infrastructure and applications that rely on IR and NLP technologies.

There is an incredible number of buzzwords and confusing terminology around the latest advancements. Part of our goal is to demystify all of these by providing a quick recap of the important concepts and distill their utility in the context of solving problems with IR and NLP components.

This interdisciplinary tutorial addresses an intermediate audience in terms of IR, NLP, ML, and evaluation methods expertise. The content should be of interest to researchers, practitioners, and graduate students.

2 CONTENT STRUCTURE

We organize the tutorial around degree of difficulty. Some things are easy to do, and some are hard. This organization allows us to focus on what is possible and realistic, given available resources.

2.1 Terminology Summary

Quick summary of terminology, main concepts and how they are related to IR, NLP, and related topics. This includes demystification of all the latest lingo and buzzwords that might be confusing.

2.2 Easy

The tutorial will start by introducing examples that are easy (and useful), such as using models on hubs (HuggingFace, AWS, Azure, PaddlePaddle) and APIs from OpenAI and elsewhere. It is relatively easy to use these models to do what they were designed to do, and not hard to modify them in creative ways to do much more, including things they were not designed to do. At the other end of the spectrum, it is extremely hard, unless you work in a well-funded industrial lab, to compete with industry on capital intensive tasks such as pre-training.

Ready-to-use. We will show how to run models on HuggingFace in simple ways for a number of applications in NLP, IR and beyond: classification, guess missing word (fill-mask; cloze task [1]), part of speech tagging [2], NER (named entity recognition), Question Answering (SQuAD [3]), Entailment (GLUE) [4, 5], Machine Translation, Speech (speech-to-text, text-to-speech), and Vision (classify objects, add captions).

Search. Full-text search features are available as part of relational databases (e.g., Postgres, Oracle, etc.) or stand alone libraries like Solr or Elastic Search. For IR, one can use BERT-like models [6–8], to map millions of documents in Semantic Scholar to vectors. And then, one can build a ranked retrieval system on top of approximate nearest neighbors (ANN) [9][1]. It does not require much compute power to apply these models to a small number of inputs. That can be done on a laptop. On the other hand, production systems that run on millions of documents is more challenging and requires more computing resources.

[1] https://pypi.org/project/annoy/1.0.3/

WSDM '24, March 4–8, 2024, Merida, Mexico

ACM ISBN 979-8-4007-0371-3/24/03…$15.00
https://doi.org/10.1145/3616855.3636452

Year	Deep Nets	Parameters (Billions)
2016	ResNet-50 [18]	0.02
2019	BERT [6]	0.34
2019	GPT-2 [19]	1.50
2020	GPT-3 [20, 21]	175.00
2022	PaLM [22]	540.00

Table 1: Deep nets are becoming larger (and more expensive).

Prompt Engineering. Prompt engineering has become super-popular with ChatGPT. The challenge is to find good applications, given ChatGPT's strengths (fluency) and weaknesses (trustworthiness). One suggestion is to help students write essays. Machines are more qualified than people for some sub-tasks and people are more qualified for other sub-tasks. To make the collaboration successful, we need to assign sub-tasks appropriately to the more qualified member of the collaboration. There is recent work that takes a programming view perspective to prompt [10]. At the same time, "jailbreaks" and other attacks against LLMs are possible [11].

ChatGPT is good at producing thesis statements and outlines, but it does not capture the user's style, and it is worse on quotes. If you ask for quotes, it "hallucinates," a technical term for making stuff up. One concern is the reaction from less well-informed users, some of whom are expecting magic from ChatGPT. Some users are giving ChatGPT a list of URLs as prompts and are surprised when ChatGPT makes stuff up, rather than crawling them. It is not hard to find prompts that produce "hallucinations" [12]. The challenge, of course, is to find prompts that produce valuable outputs. One highly-cited suggestion is "Chain of Thought" prompting [13].

Recently, prompt engineering has become more popular than fine-tuning, perhaps because prompt engineering is easier, and more accessible to a broader audience. Prompt engineering offers *instant gratification*. Fine-tuning experiments often take hours. With prompt engineering, feedback is more immediate.

2.3 Hard

As illustrated in Table 1 [14], deep nets are becoming larger and larger (for better or for worse). Deep nets are getting bigger over time, where size can be measured in a variety of different ways:

(1) model size, m (number of parameters),
(2) number of dimensions, d (problem size),
(3) size of (annotated and unannotated) training data,
(4) staff (authors per paper),
(5) hardware (number of CPUs, GPUs and data centers)
(6) costs (including externalities such as carbon [15, 16]).

The consensus in industry, at least for practical applications, is that bigger nets are better. Most of these larger models are coming from industry; training large models has become too expensive for academia [17]. It is important to pick one's battles carefully. Industry is in a better position to make large investments, but after making large bets, they cannot afford to take chances. Academia can afford creativity, but not large bets.

There is considerable excitement recently in large language models (LLMs). But it is important to appreciate their strengths and weaknesses. Although there has been much progress, there are always opportunities for improvement. After seeing some amazing results with LLMs, some students wonder if it is too late to join in on the fun. If LLMs have already solved all the world's problems, is there any room left for the next generation of students to make improvements?

Empiricism and Rationalism. Obviously, there is plenty of room for students to contribute. Consider hallucinations, for example. It is natural to underestimate the magnitude of this problem, and suggest it will be fixed "in the next release." In fact, hallucinations are the tip of well-known (and very hard) problems that our teachers (Minsky and Chomsky) were beating their heads against. When we created EMNLP in 1990s, we were advocating a pivot away from those hard problems (rationalism) toward easier problems (empiricism) that their teachers worked on in 1950s [23].

- 1950s: Empiricism (Firth, Harris)
- 1970s: Rationalism (Minsky, Chomsky)
- 1990s: Empiricism (EMNLP)

To make progress on hallucinations, it may be necessary in the long-term to revisit some harder problems and revive rationalism, though in the short-term, it might be possible to use search to fact-check assertions.

2.4 Not So Hard

Fine-Tuning. Suppose one wanted to build a field-guide application, where the user could use their phone to take a picture of a flower, and the app would take the user to the appropriate web page. Fine-tuning is "unreasonably effective" in transferring knowledge from pre-trained base models to novel tasks. In this case, we start with the pre-trained model, ResNet [18], which was trained on 14 million images and 1000 class labels from ImageNet [24].

The fine-tuning task is to modify the base model to recognize 5 types of flowers instead of the 1000 ImageNet classes. The 5 flower classes are: *rose*, *tulip*, *daisy*, *sunflower* and *dandelion*.

For fine-tuning, we are given a training set and a validation set. Both sets consist of pictures of flowers, x, labeled with the 5 classes, y. There are 2915 flowers in the training set, and 383 flowers in the the validation set. The validation set is used to measure error. That is, after fine-tuning, the model is given a picture from the validation set, x, and asked to predict a label, $\hat{y}$. These predictions, $\hat{y}$, are compared with gold labels, y, to produce a score.

At inference/evaluation time, we are given a novel picture, x, and a set of possible class labels such as the 5 classes of flowers. The model predicts a label, $\hat{y}$, one of the class labels. Before fine-tuning, the model is performing at chance since the ImageNet uses different classes. After fine-tuning, the model is considerably better than chance, though far from state of the art (SOTA).

Fine-tuning is just one of many tools in the toolbox. If one wants to top the leaderboard, one needs an "unfair advantage," something better than what the competition is likely to do. Since fine-tuning is now well established within the literature, one should assume that the competition is likely to do that. One is unlikely to do much better than the competition (or much worse than the competition) if one uses obvious methods (such as fine-tuning) in obvious ways.

In Section 2 of [25], we describe three examples of fine-tuning in more detail (See GitHub for fine-tuning code):[2] Flowers, SQuAD

[2]https://github.com/huggingface/transformers/tree/main/examples/pytorch

[26] (question answering), GLUE [5, 27] (a collection of 9 NLP tasks that include both classification and regression).

2.5 Ugly

Risks. Much of the literature on Responsible AI focuses on Risks 1.0 (fairness and bias), but in [28], we argued that there should also be more work on Risks 2.0 (addictive, dangerous, deadly and insanely profitable). The use of ML to maximize engagement in social media has created a Frankenstein Monster that is exploiting human weaknesses with persuasive technology, the illusory truth effect, Pavlovian conditioning, and Skinner's intermittent variable reinforcement. In [29], we introduced a third risk: it is bad to treat people badly (Risk 1.0), but worse to kill them (Risk 2.0) and even worse to do so with malicious intent (Risk 3.0):

(1) Risks 1.0: Biased and unfair [30].
(2) Risks 2.0: Addictive, dangerous, deadly and insanely profitable [31, 32].
(3) Risks 3.0: Proliferation of military-grade spyware to 50,000 or more cell phones, targeting journalists and their friends and families and many others [33, 34].

Evaluation and Benchmarks. More and more benchmarks, datasets, and evaluation tasks are becoming available online at rapid speed. This is extremely useful for the community and allows researchers and practitioners to test and evaluate new techniques. However, the construction, evaluation, and maintenance of such benchmarks is opaque which creates problems with respect to stability and true representations.

It is easy to produce impressive numbers but it is very difficult to produce numbers that we can trust [35]. We describe scenarios where numbers look too good to be true. We will also discuss some aspects of benchmarks that make it difficult to test validity and reliability. Tasks and use cases are in many cases very artificial and don't reflect the real world making the entire evaluation process detached from reality.

Ground Truth and Labeling. Most of the AI-based solutions require ground truth data that is created in a labeling step. Large-scale labeling (aka annotations) is very hard [36–38] and even LLM-based solutions still require a layer of human computation. There is very recent work on using LLMs for labeling as a mechanism to scale up the process and reduce the human cost. That said, task design, task complexity, and aggregation techniques are still required.

2.6 Research Opportunities

For each of the above items we outline research opportunities and problems that need to be solved.

2.7 Supporting Material

Slides and related material will be posted on GitHub. This material will make it easy for participants to do many easy things (inference and prompt engineering), and some not so hard things (fine-tuning).

3 FORMAT, SCHEDULE, AND PRESENTERS

The tutorials will be 3 hours, on-site, and the outline is as follows.

3.1 Schedule

The schedule below allows 30 minutes for breaks, Q&A and slippage.

(1) Introduction (10 minutes)
(2) Easy (60 minutes): Inference, Prompting, Pattern Matching; examples will be selected from NLP and IR
(3) Hard (20 minutes): Pre-training, fixing hallucinations, understanding
(4) History (10 minutes): ChatGPT's strengths (fluency) and weaknesses (trustworthiness)
(5) Not so Hard (30 minutes): Fine-tuning; knowledge graphs
(6) Ugly/Responsible AI (20 minutes): Risks, labeling, and evaluation.
(7) Conclusions (10 minutes)

3.2 Presenters

O. Alonso is a senior applied science manager at Amazon working on the intersection of information retrieval, knowledge graphs, and human computation. In the past, he was a principal applied scientist lead at Microsoft, where he worked on the Bing search engine. He is the co-organizer of DESIRES, a new information retrieval conference with a focus on system implementation and experimental design.

K. Church[3] was an early advocate in 1990s of the revival of empirical methods and corpus-based lexicography. He introduced what is now known as PMI (point-wise mutual information) [39]. He was a founder of EMNLP, a major conference in Computational Linguistics. Church was president of the ACL in 2012. He was an AT&T Fellow in 2001, ACL Fellow in 2015 and Baidu Fellow in 2018. Church led a team at a 6-week summer workshop, JSALT-2023.[4] The team used deep nets and spectral clustering in an Academic Search application.[5]

Both Alonso and Church have considerable experience in industry and academia (Northeastern, Columbia, Johns Hopkins, University of Pennsylvania). They taught a joint class at Northeastern (Spring 2023). They have given many tutorials at ACL[6] and elsewhere. They will be giving a beta version of this tutorial at CIKM-2023.[7] Some of the material for this tutorial will be borrowed from a class Church is teaching this term.[8]

Alonso has co-presented a number of tutorials on IR topics at WSDM, WWW, and SIGIR. Church has written a number of tutorials on Deep Nets [25, 40, 41]. "Deep nets for poets" [40] is based on "Unix for poets" [42], a tutorial Church gave at a linguistics summer school nearly 30 years ago. Church was a TA for Weizenbaum in 1978 and has first-hand knowledge of how horrified he was when people took ELIZA seriously. The comparison of ChatGPT and Weizenbaum's ELIZA was called out in a recent article in the Guardian.[9]

[3] https://kwchurch.github.io/
[4] https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html
[5] https://www.semanticscholar.org/api-gallery/better-together
[6] https://github.com/kwchurch/ACL2022_deepnets_tutorial
[7] https://uobevents.eventsair.com/cikm2023/tutorials
[8] https://kwchurch.github.io/teaching/2023-fall/CS6120/syllabus.html
[9] https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai

REFERENCES

[1] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," *Journalism quarterly*, vol. 30, no. 4, pp. 415–433, 1953.

[2] K. Church, "A stochastic parts program and noun phrase parser for unrestricted text," in *Proceedings of the second conference on Applied natural language processing*. ACL, 1988.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[4] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://www.aclweb.org/anthology/W18-5446

[5] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[7] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[8] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2270–2282. [Online]. Available: https://aclanthology.org/2020.acl-main.207

[9] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.

[10] L. Beurer-Kellner, M. Fischer, and M. T. Vechev, "Prompting is programming: A query language for large language models," *Proc. ACM Program. Lang.*, vol. 7, no. PLDI, pp. 1946–1969, 2023.

[11] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023.

[12] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in chatgpt: implications in scientific writing," *Cureus*, vol. 15, no. 2, 2023.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *arXiv preprint arXiv:2201.11903*, 2022.

[14] K. W. Church, "Emerging trends: Deep nets thrive on scale," *Natural Language Engineering*, vol. 28, pp. 673 – 682, 2022.

[15] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *ArXiv*, vol. abs/1906.02243, 2019.

[16] R. Schwartz, J. Dodge, N. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, pp. 54 – 63, 2019.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Kohd, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," 2021.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019.

[20] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *NeurIPS*, 2020.

[21] R. Dale, "GPT-3: What's it good for?" *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, 2021.

[22] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

[23] K. Church, "A pendulum swung too far," *Linguistic Issues in Language Technology*, vol. 6, no. 5, pp. 1–27, 2011.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[25] K. W. Church, Z. Chen, and Y. Ma, "Emerging trends: A gentle introduction to fine-tuning," *Natural Language Engineering*, vol. 27, no. 6, p. 763–778, 2021.

[26] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://www.aclweb.org/anthology/D16-1264

[27] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, 2018.

[28] K. Church, A. Schoene, J. E. Ortega, R. Chandrasekar, and V. Kordoni, "Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable," *Natural Language Engineering*, p. 1–26, 2022.

[29] K. W. Church and R. Chandrasekar, "Emerging trends: Risks 3.0 and proliferation of spyware to 50,000 cell phones," *Natural Language Engineering*, vol. 29, no. 3, pp. 824–841, 2023.

[30] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[31] M. Fisher, *THE CHAOS MACHINE: The Inside Story of How Social Media Rewired Our Minds and Our World*. Little, Brown & Company, 2022.

[32] M. Bergen, *Like, Comment, Subscribe: Inside YouTube's Chaotic Rise to World Domination*. Viking, 2022.

[33] L. Richard and S. Rigaud, *Pegasus: How a Spy in Your Pocket Threatens the End of Privacy, Dignity, and Democracy*. Henry Holt and Company, 2023.

[34] N. Perlroth, *This is How They Tell Me the World Ends: The Cyberweapons Arms Race*. Bloomsbury Publishing, 2021.

[35] D. T. Ron Kohavi and Y. Xu, *Trustworthy Online Controlled Experiments*, 2020.

[36] O. Alonso, *The Practice of Crowdsourcing*, ser. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2019.

[37] A. Braylan, O. Alonso, and M. Lease, "Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks," in *WWW*, 2022, pp. 1720–1730.

[38] A. Braylan, M. Marabella, O. Alonso, and M. Lease, "A general model for aggregating annotations across simple, complex, and multi-object annotation tasks," *Journal of Artificial Intelligence Research*, 2023.

[39] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990. [Online]. Available: https://www.aclweb.org/anthology/J90-1003

[40] K. W. Church, X. Yuan, S. Guo, Z. Wu, Y. Yang, and Z. Chen, "Emerging trends: Deep nets for poets," *Natural Language Engineering*, vol. 27, no. 5, p. 631–645, 2021.

[41] K. W. Church, X. Cai, Y. Ying, Z. Chen, G. Xun, and Y. Bian, "Emerging trends: General fine-tuning (gft)," *Natural Language Engineering*, pp. 1–17, 2022.

[42] K. W. Church, "Unix™ for poets," *Notes of a course from the European Summer School on Language and Speech Communication, Corpus Based Methods*, 1994.