RadicAI: A Radical, Though Not Entirely New, Approach to AI Paper Naming

Jim McCann
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
ix@tchow.com

Mike McCann
Department of Computational Mathematics, Science and Engineering
Michigan State University
East Lansing, MI 48824
michael.thompson.mccann@gmail.com

Abstract

It seems that nothing can stop the explosive, singularity-like growth of papers published about Machine Learning (AI). The floods of graduate students, funding, and computational resources show no signs of abating. But those publishing in AI will soon face their most daunting resource limitation: unique acronyms. In this paper, we quantify the scope of acronym scarcity, and profile one potential solution to the acronym scarcity problem by using visually-approximately-correct (VAC) digraph substitution.

1 Introduction

Now that AI has captured the pocketbooks of starry-eyed government funding agencies and thirsty venture capitalists, the research community has begun to produce machine learning papers at a prodigious rate. But this exponential growth in article production may soon be halted by acronym scarcity.

It is common knowledge that every good AI paper should be titled with a single English word – an *initial title word* (ITW) – followed by a colon, followed by a phrase that contains many letters in common with that single word. And, of course, a top-tier AI paper will include the digraph "AI" in this initial word.

But the English language, for all its *ruhmbedecktwortschatz*, is limited it its word count. Indeed, only 132,544 possible initialisms corresponding to reasonably common English words exist, and – of these – only 2,592 are top-tier.

Authors have already begun to scramble to avoid the acronym shortage by deploying a number of techniques, including using less-standard (or even entirely-made-up) words [NeRF:20] or names of Muppets [BERT:19, KERMIT:02]. Particularly brave and self-sacrificing authors have even gone so far as to avoid initialism entirely [Human-level15, Attention17], though it is unclear if such papers will ever have any impact.

In this paper, we propose a middle-ground solution: using the visual ambiguity of sans-serif fonts to develop paper titles which are both top-tier and visually, approximately correspond to real words.
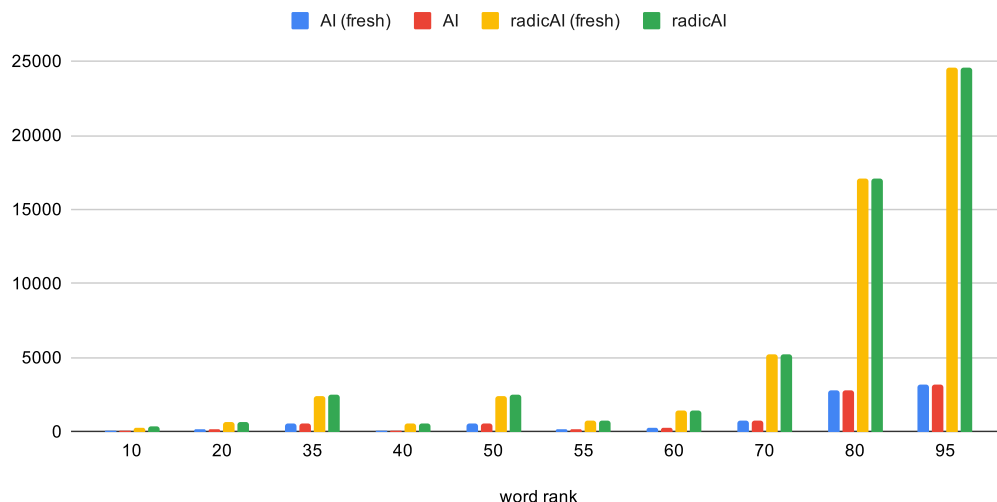
ITWs under RadicAI

Figure 1: With RadicAI, the number of fresh initial title words available for top-tier AI papers is increased nearly tenfold at all word complexity ranks. Note that "(fresh)" bars remove initial title words already in use.

2   Background

The technique reported in this paper was inspired by the Medium article "LocAI: AI Design for Local Contexts" [LocAI:21].

3   The Scope of the Solution

According to SCOWL [SCOWL:00], the English language contains somewhere between 4,068 and 465,999 words suitable for use as ITWs. The exact count depends on what portion of SCOWL one chooses to use. These portions correspond to word complexity/rarity ranks between 0 (common) and 100 (legendary), with words divided into 10 bins depending on their ranks[1].

Of these words, only 8,561 can possibly be used for top-tier papers, and only 2,592 of these lie at or below SCOWL rank 70. However, by taking the radical[2] step of using the fact that the digraph "AI" appears close to "AI" when typeset in sans-serif capital letters, 47,158 (11,398 at or below 70) new possible ITWs become available – Figure 1.

4   The Scope of the Problem

With our solution clearly in hand, we set out to discover the scope of the ITW-depletion crisis by examining paper titles in 1,854,689 papers from the most prestigious, peer-reviewed AI journal: ar14[3].

We accomplished this might feat of computering by using the tools provided by the arxiv-public-datasets project [On19].

For each paper, we extracted the title and an initial acronym by using the arcane might of regular expressions. In this complete dataset, 33,895 papers use initial title words and 750 of these papers

---

[1]These probably correspond to percentiles but I haven't actually read the README in SCOWL recently and have no intention of doing so now.
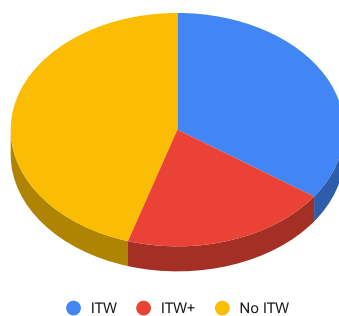
[2]Or, perhaps, "radicAI step"?

[3]Also known as "arXiv" if you still use Roman numerals.

Prevalence of Initial Title Words (all ar14)

Prevalence*of Initial Title Words (cs.AI)

ITW   ITW+   No ITW

ITW   ITW+   No ITW
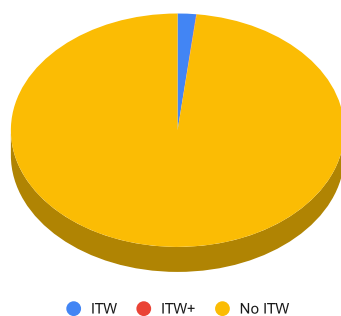
*log scale to show texture

Figure 2: Although, as far as we know, every paper on ar14 is an AI paper, the papers in the "cs.AI" category appear, in general, to use far more initial title words (ITW) as well as initial title words containing "ai" (ITW+).

contain the digraph "ai". Interestingly, in the cs.AI subcategory only 2,748 papers use initial title words and only 96 of these ITWs contain "ai" (Figure 2). I suppose one must, therefore, conclude that the majority of top-tier AI papers are not even published as AI papers.

Further, and perhaps distressingly, many common English words have already been used as initial title words (Figure 3). This depleted stock has already resulted in a fair number of collisions [W-net:20c, W-Net:19b, W-Net:17, W-Net:19c, w-Net:20b, W-Net:20a, W-Net:19a].

5   Conclusions

By taking the radicai step of confusing the digraphs ai and al, the space of top-tier initial words for paper titles is greatly expanded.

"Big data. I get it."

Acknowledgments and Disclosure of Funding

Well, um, that's awkward. Were we supposed to have funding?

Initial Title Words as Percentage of Available English Words

Initial Title Words vs Available English Words

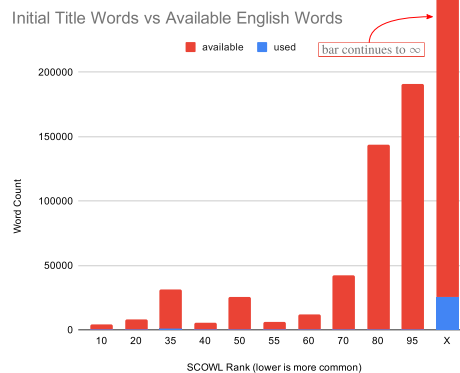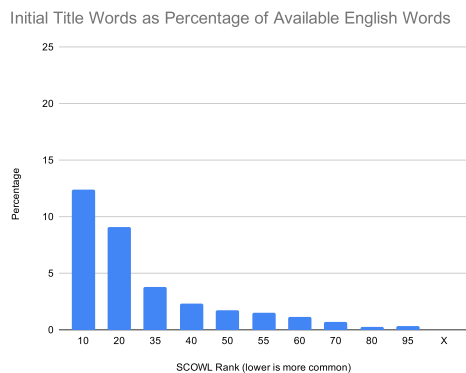available   used

bar continues to ∞

Figure 3: The stock of low-complexity initial title words is already significantly decreased, though most ITWs are actually non-words (rightmost bar in the graphs).

References

[Attention17]   Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[BERT:19]   Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[Human-level15]   Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[KERMIT:02]   Pramuditha Suraweera and Antonija Mitrovic. KERMIT: A constraint-based tutor for database modeling. In *Intelligent Tutoring Systems*, pages 377–387, Berlin, Heidelberg, 2002.

[LocAI:21]   People+ AI Research. Locai: Ai design for local contexts. `https://medium.com/people-ai-research/locai-ai-design-for-local-contexts-9ecfde4aeac8`, February 2021.

[NeRF:20]   Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, 2020.

[On19]   Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of arxiv as a dataset, 2019.

[SCOWL:00]   Kevin Atkinson. Scowl: Spell checker oriented word lists. `http://wordlist.aspell.net/`, 2000.

[W-Net:17]   Xide Xia and Brian Kulis. W-Net: A deep model for fully unsupervised image segmentation, 2017.

[W-Net:19a]   Changhun Jung, Mohammed Abuhamad, Jumabek Alikhanov, Aziz Mohaisen, Kyungja Han, and DaeHun Nyang. W-Net: A cnn-based architecture for white blood cells image classification, 2019.

[W-Net:19b]   Kwang-Hyun Uhm, Seung-Wook Kim, Seo-Won Ji, Sung-Jin Cho, Jun-Pyo Hong, and Sung-Jea Ko. W-Net: Two-stage u-net with misaligned data for raw-to-rgb mapping. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Oct 2019.

[W-Net:19c]   Varun Kannadi Valloli and Kinal Mehta. W-Net: Reinforced u-net for density map estimation, 2019.

[W-Net:20a]   Gautam Rajendrakumar Gare, Jiayuan Li, Rohan Joshi, Mrunal Prashant Vaze, Rishikesh Magar, Michael Yousefpour, Ricardo Luis Rodriguez, and John Micheal Galeotti. W-Net: Dense semantic segmentation of subcutaneous tissue in ultrasound images by expanding u-net to incorporate ultrasound rf waveform data, 2020.

[w-Net:20b]   Bo Wang, Lei Wang, Junyang Chen, Zhenghua Xu, Thomas Lukasiewicz, and Zhigang Fu. w-Net: Dual supervised medical image segmentation model with multi-dimensional attention and cascade multi-scale convolution, 2020.

[W-net:20c]   Hongwei Zhao, Chengtao Peng, Lei Liu, and Bin Li. W-net: Simultaneous segmentation of multi-anatomical retinal structures using a multi-task deep neural network, 2020.