Abstract

This literature review assesses the contributions of 'smart' and 'dumb' devices to the space of text extraction from PDF files and correction of encoding errors. Given the universality of PDFs and the vast amounts data contained in them, this exploration could be useful in advancing the quality of machine learning datasets. I cover four themes: (1) A high-level overview of my understanding of the meaning of "smart" and "dumb" as applied to machines, (2) The current application of "smart" devices in the space of text extraction and error correction, (3) the challenges encountered when training language models with data from PDFs, and (4) an overview of prior attempts to address these challenges. This review will serve as a guide to the development of an application aimed at making text extraction and error correction from PDFs more efficient and accessible.

Theme 1: Big Brain, Small Brain

The use of "smart" and "dumb" as applied to devices has much in common with their use as applied by humans, to humans. We ascribe intelligence or its lack to people, and we extend this analogy to our machine devices. I divide this concept into three key parameters: adaptiveness, incisiveness, and responsiveness.

The adaptiveness of a device corresponds to the versatility of human intelligence. A "smart" device, like an intellectually adaptable person, can handle a diverse range of tasks. A good natural language model can speak in various languages, about a wide breadth of domains. A "dumb" device is limited, able to perform only a narrow set of functions – a bank chatbot, for example, is only able[1] to address the concerns of a banking customer.

The incisiveness of a device corresponds to the depth and precision of human thought. A "smart" device doesn't just function in a variety of environments; it does so meaningfully, and precisely. Its outputs appear insightful and show a clear "grasp"[2] of the context in which the device operates. In contrast, a "dumb" device may provide superficial or erroneous responses, much like shallow or ill-conceived[3] human thought processes.

Finally, the responsiveness of a device is akin to the speed and immediacy of a human's cognitive processes – how "quick" they are. In a human being, this is coming up with the perfect joke to diffuse a social faux pas right after it happens, rather than on the way home[4]. A technological example is an iPhone's swift lowering or increasing of screen brightness in response to changes in local light conditions.

Cutting edge Artificial Intelligence (AI) approaches stand as the ultimate "smart" devices, defined by their adaptiveness, incisiveness, and responsiveness relative to other technologies.

[1] Arguably

[2] These scare quotes are as far into theory of mind as this review is going to get

[3] This includes the vast sum of my own thoughts

[4] Or five years later, as one is settling into bed

Theme 2: Smart Devices and Their Role in Text Extraction and Error Correction

Smart devices have been applied to a wide span of domains, from the deconstruction of songs into their constituent parts[5] and the prediction of highly complex protein structures[6], to the indiscriminate persecution of farmland weeds[7] and Apple Engineers[8,9].

Of particular interest for this review is their role in data processing and digitization. Here, smart devices[10,11] have played a crucial role in extracting and processing text from various sources, digital and otherwise. One critical application is data extraction from PDFs.

Optical Character Recognition (OCR) technology is one of the backbones of this process. OCR uses Machine Learning (ML), Artificial Intelligence (AI), and Computer Vision (CV) to convert different types of data – from hand-written characters to full manuscripts – into machine-encoded text.[12]

Large Language Models (LLMs) are another technology that have increased our ability to interact with PDF documents; these technologies allow individuals to converse with and summarise PDF documents, making it easier to extract understanding from them.[13]

These technologies allow human beings to harness the ungodly[14] power of machines to easily access, edit, and search text locked in physical or digital formats, and thus further extending the domain of humanity's unabated pursuit of power and control.[15,16]

[5] Facebook Research, "Demucs," GitHub, last modified March 21, 2020. https://github.com/facebookresearch/demucs/tree/v2

[6] DeepMind , "AlphaFold", DeepMind, last modified July 28, 2020. https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe

[7] Freethink, "*Laser 'death ray' kills weeds 80x faster than humans*", June 30, 2023. https://www.freethink.com/series/hard-reset/laser-weeding

[8] Catherine Thorbecke, "*Tesla on autopilot had steered driver towards same barrier before fatal crash, NTSB says*", ABC News, Feb 12, 2020. https://abcnews.go.com/Business/tesla-autopilot-steered-driver-barrier-fatal-crash-ntsb/story?id=68936725

[9] This juxtaposition might read as odd, but try explaining that to the Google Maps car writing this for me

[10] "reCAPTCHA: Easy on Humans, Hard on Bots," Google, last modified August 2, 2023, https://www.google.com/recaptcha/intro/?zbcode=inc5000.

[11] Including the human smart devices used to train the silicon-substrate smart devices

[12] "OCR: Text Recognition for your Business," Google Cloud, last modified August 1, 2023, https://cloud.google.com/use-cases/ocr.

[13] John Biggs, "5 AI tools for summarizing a research paper," Cointelegraph, August 2, 2023, https://cointelegraph.com/news/5-ai-tools-for-summarizing-a-research-paper

[14] Jeremy Naydler, "The Archetype of the Binarius and the Prehistory of the Computer," (pamphlet, Guild of Pastoral Psychology)

[15] *Ibid*

[16] "Industrial society," Wikipedia, last modified July 31, 2023, https://en.wikipedia.org/wiki/Industrial_society

Theme 3: Obstacles In Extracting Training Data From PDFs

The quest for more advanced AI systems is increasingly accepted as an inevitable societal goal.[17,18] This is largely driven by those in power,[19] whose unrelenting greed[20] and bloodlust[21,22] often lead to significant resources[23,24] being diverted towards these ends.[25]

A common obstacle in the process of training more advanced AI systems is the quality of data available.[26,27] In light of this, PDFs represent an untapped treasure trove. The format's prevalence across numerous sectors – academia, business, government, etc – makes it a rich source of high-quality data and expert opinion, just waiting to be harnessed.

This trove is, however, quite unwieldy for lay-practitioners due to the challenges associated with parsing PDFs. Encoding errors, usually caused by complicated layouts, low-quality documents, or unusual fonts, can mess up extracted text.[28,29,30,31] And while OCR remedies this a little bit, it's not without its problems. The largest being that one must generally pay for access to these tools[32,33], and most people don't enjoy paying for things[34]. Another being that these solutions aren't designed for use at speed at the scale of ML training.[35] Yet another being that it's hard to control its process.

[17] Billy Perrigo, "U.K. Chancellor Rishi Sunak Wants to Regulate AI. It Could Set a Global Trend," Time, August 2, 2023, https://time.com/6287253/uk-rishi-sunak-ai-regulation/.
[18] Rob Price, "Sundar Pichai says Google's AI is more profound than fire, electricity, or the internet," Business Insider, April 23, 2023, https://www.businessinsider.com/sundar-pichai-google-ai-bard-profound-tech-human-history-2023-4.
[19] Robert Hackett, "Bill Gates, Jeff Bezos, and Jack Ma are betting on this AI unicorn to revolutionize mining," Fortune, June 20, 2023, https://fortune.com/2023/06/20/bill-gates-jeff-bezos-jack-ma-ai-unicorn-kobold-mining/.
[20] Daniel Fisher, "Wall Street's Need For Trading Speed," Forbes, September 27, 2010, https://www.forbes.com/forbes/2010/0927/outfront-netscape-jim-barksdale-daniel-spivey-wall-street-speed-war.html.
[21] Daniel Gros, "Why the EU must now tackle the risks posed by military AI," Centre for European Policy Studies, August 3, 2023, https://www.ceps.eu/why-the-eu-must-now-tackle-the-risks-posed-by-military-ai/#:~:text=Large%20investments%20in%20military%20AI,love%20affair%20with%20the%20technology.
[22] Valerie Insinna, "Artificial intelligence flies XQ-58A Valkyrie drone," Defense News, August 3, 2023, https://www.defensenews.com/unmanned/2023/08/03/artificial-intelligence-flies-xq-58a-valkyrie-drone/.
[23] Alex Hern, "How the internet was invented," The Guardian, July 15, 2016, https://www.theguardian.com/technology/2016/jul/15/how-the-internet-was-invented-1976-arpa-kahn-cerf.
[24] Cade Metz, "Salaries for Artificial Intelligence Researchers Are Skyrocketing. That's a Problem," The New York Times, April 19, 2018, https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html.
[25] "Dark triad," Wikipedia, last modified August 2, 2023, https://en.wikipedia.org/wiki/Dark_triad.
[26] Forbes Tech Council, "Data Quality: The Real Bottleneck In AI Adoption," Forbes, February 1, 2023, https://www.forbes.com/sites/forbestechcouncil/2023/02/01/data-quality-the-real-bottleneck-in-ai-adoption/.
[27] Andrew Griffin, "AI training is being fed by 'the worst of the internet'," The Independent, August 3, 2023, https://www.independent.co.uk/tech/ai-training-data-internet-junk-b2360570.html.
[28] "Encoding problems when extracting text with PyPDF2," Stack Overflow, February 16, 2021, https://stackoverflow.com/questions/66225205/encoding-problems-when-extracting-text-with-pypdf2.
[29] "Issue #37," GitHub, opened on March 3, 2013, https://github.com/py-pdf/pypdf/issues/37.
[30] "Issue #235," GitHub, opened on July 31, 2023, https://github.com/py-pdf/pypdf/issues/235.
[31] Examples of this kind of issue go on forever. I've faced it myself
[32] "Adobe Acrobat DC plans & pricing," Adobe, last modified August 3, 2023, https://www.adobe.com/uk/acrobat/pricing.html.
[33] "Pricing," ABBYY FineReader PDF, last modified August 3, 2023, https://pdf.abbyy.com/pricing/.
[34] "Homo economicus," Wikipedia, https://en.wikipedia.org/wiki/Homo_economicus.
[35] You want evidence? Try making a 100-book training set for a language model using Adobe Acrobat.

Theme 4: People Who've Already Done Some Or All Of What I'm Planning On Doing

The market already offers a range of applications designed for text extraction. One such application is Adobe Acrobat DC. Another example is ABBYY FineReader. And so on. Here's what's important: *none of these softwares have an effective way of manually accounting for encoding errors*.

They provide basic tools to address these issues, like find-and-replace operations or manual editing, but these methods are, in my experience, often not enough to handle encoding errors. This particularly frustrating when one is dealing with large documents, where manual correcting such problems can become extremely tedious.

Dumb, smart – it doesn't matter. There are no tools, yet, with the either the adaptability, incisiveness, or responsiveness to solve this problem all the way. What exists right now only takes us part of the way.

It seems to me that the current landscape of text extraction software highlights a significant need for tools capable of helping users accurately and efficiently correct encoding errors. It goes without saying that I am just the undergraduate to take on such a task.

Conclusion

I like training ML models. I and many others have found that it's difficult to create high quality datasets with PDFs due to the bad encodings created during text extraction. Seeing as I have been unable to find a pre-existing, suitably priced[36], and lightweight solution to this problem, I intend to create one that is able to find and replace bad encodings in a way that is both controllable, and context sensitive

[36] Free