Are machine learning corpora "fair dealing" under Canadian law?

Dan Brown
Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada
dan.brown@uwaterloo.ca

Lauren Byl
Library,
University of Waterloo
Waterloo, Ontario, Canada
lrbyl@uwaterloo.ca

Maura R. Grossman
Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada
maura.grossman@uwaterloo.ca

Abstract

We consider the use of large corpora for training computationally creative systems, particularly those that write new text based on the style of an existing author or genre. Under Canadian copyright law, a key concern for whether this is "fair dealing" is whether this usage will result in new creations that compete with those in the corpus. While recent law review articles in the United States suggest that training models on such corpora would be "fair use" in the United States, we argue that Canadian law may, in fact, forbid this use when the new products compete with works in the original corpus.

Introduction

The fair-dealing exception (section 29 – 29.2) in the Canadian Copyright Act (RSC 1985, c. C-42) allows the use of copyright-protected materials without permission or payment of royalties under certain circumstances. These circumstances include research, criticism, review, and private study, as long as what is done with the work is "fair." For the purpose of criticism and review, proper citation of the copyrighted material is also required.

"Fair" dealing is not the clearest of concepts: There is no checklist with a target score that ensures that use is certain to be judged fair. Instead, there are a collection of factors that are evaluated as part of the review process. This collection of factors is not part of the Copyright Act. Instead, it was provided in a Supreme Court of Canada ruling, *CCH Canadian Ltd.* v. *Law Society of Upper Canada* (2004 SCC 13). The factors include the purpose of the use, the amount of the copyrighted material being used, and the effect of the use on the original work (for example, does the use compete with the market for the original work?).

Natural language corpora, as used by machine learning systems, form an interesting test case for this "fair-dealing" regime. Current state-of-the-art text-generation systems use truly massive data sets of human-authored text and generate text that is more and more like what a human would write. In 2021, such systems are trained on billions of words of human-authored text (Brown et al. 2020).

A rapidly growing paradigm in text generation these days is to use GPT-2 fine-tuned with an author-specific or domain-specific corpus. (See, for example, Lee and Hsiang (2019). This approach yields texts with the surface features of an author/domain, while taking advantage of the high-grammatical fluency of transformers like GPT-2 (Radford et al. 2019). Therefore, this use can have two different components: modelling the corpus to inform the overall generation of the new text, and (possibly) inclusion of some short bits of the training corpus directly in generated works. (The current state-of-the-art system is actually GPT-3, which is fairly hard for most researchers and companies to access; in contrast, GPT-2 is readily available.) The resulting documents are often surprisingly hard to distinguish from human-authored text, although many still require a fair amount of human editing and correction. One example was an op-ed written by GPT-3, titled "A robot wrote this entire article are you scared yet, human?," published in *The Guardian* in September 2020, but the true story of a lot of these products involves a huge amount of human massaging and shaping (Uitdenbogerd 2020; Jordanous 2017).

Many corpora for these systems are created by Web crawlers. This is certainly the case for the base corpus on which GPT-2 is trained, and could be true for other creative systems; for example, a poetry generator might be trained with user-submitted poetry on a poetry forum. This raises the question of the copyright status of the source documents, which may be set by the organizers of these fora themselves, in their own policies.

All of this leads to the question: Is creating a machine learning corpus "fair dealing" if the material contained in it is itself copyrighted? Here, we investigate this question by looking at the factors defining "fair dealing" in Canadian case law, after briefly contrasting the situation in Canada

Proceedings of the 12th International Conference on Computational Creativity (ICCC '21)
ISBN: 978-989-54160-3-5

with the "fair use" model used in the United States. We focus our consideration on a factor concerned with creation of new works that compete with the copyrighted works being used for training, and explore a number of scenarios for this use. Ultimately, we argue that such use may, in fact, be unfair, and give a brief discussion of the consequences of this conclusion.

Existing Literature on the Question

The question of copyright and large-scale corpora is not novel; copyright issues for translation corpora, for example, have previously been discussed (Wilkinson 2006), and there is a short primer on how natural language corpora intersect with German copyright law (DFG Review Board n.d./2017). The World Intellectual Property Organization has studied whether or not such corpora should be permitted under international rules (2019), and some experts have highlighted the rulings in the United States in both the *Hathi Trust*[1] and the *Google Books*[2] cases as showing that the American "fair use" concept allows for the creation of corpora for research and other purposes.

In the Canadian context, Craig (2020) has explored whether computer-authored texts deserve copyright protection (her opinion is that they do not), and has expressed her concern that current law does not make clear how infringement concerns, if they were held to be valid, could be targeted at infringing systems that, in her view, lack autonomy as creators.

Our question is a bit different, as we focus on compu–tationally creative authoring systems. These have not been the focus of any Canadian legal scholarship that we can identify. We focus on Canadian law, because it governs the computational creativity research of the first author.

Fair Dealing, Not Fair Use

We specifically are asking about machine language corpora in light of the Canadian concept of "fair dealing," not the "fair-use" exemption available to users in the United States (Title 17 USC §107). U.S. law allows for researchers to make copies of copyrighted materials, and in both the *Google Books* and *Hathi Trust* cases, large-scale digital analysis of corpora (for example, to enable search) was seen as non-infringing, as long as users were not gaining access to chunks of the copyrighted materials in those corpora that were commercially relevant, such as whole pages of books, or definitions found in dictionaries. Some legal academics in the U.S. have argued for a much wider exemption, both for research in general, and for text and data mining in specific, claiming that the fair-use principles allow for it because of the positive benefits to society of scientific research (Carroll 2019). Sobel (2017) argues that Artificial Intellignece (AI) is in a potential crisis if learning from corpora and creating derived works is not held to be fair use, and argues that current U.S. law, which he deems to support this use, in fact gives U.S. researchers a competitive advantage.

Very recently, the U.S. Supreme Court has also ruled (in *Google* v. *Oracle*[3]) that Google's use of some Java application programming interface code was fair use, although the ruling did not answer the question of whether the code itself was copyrightable. This ruling is not directly applicable to the case of corpora and fair use, but nonetheless suggests a willingness to allow for technology innovation as fair use, consistent with Sobel's hopes.

More Detail: Factors for Fair Dealing

The full list of valid contexts for "fair dealing" are: research, private study, criticism, review, education, satire, parody, and news reporting. Creating a text generator is likely not education. If the generator is writing news articles, critiques or reviews, then the use of source materials about the events being reported on, or the work being reviewed or critiqued is fair dealing, but the base corpus being used to train language-model parameters is not itself being used in reporting, criticism or review.

Most of the time, a text generator is also not private study. Satire and parody may be the underlying goal for some text generators (e.g., mashups of the works of H.P. Lovecraft and the King James Bible, for example, Stross (2013)), but these are not the most common, and do not form the basis for the examples we discuss later in this paper.

The "Research" exception to the Canadian Copyright Act is broadly defined by case law; the ruling in *CCH v. LSUC* (2004) holds that "The fair dealing exception under s. 29 is open to those who can show that their dealings with a copyrighted work were for the purpose of research or private study. 'Research' must be given a large and liberal interpretation in order to ensure that users' rights are not unduly constrained. I agree with the Court of Appeal that research

[1] Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 2014 U.S. App. LEXIS 10803, Copy. L. Rep. (CCH) P30,617, 42 Media L. Rep. 1898, 111 U.S.P.Q.2D (BNA) 1001, 2014 WL 2576342 (United States Court of Appeals for the Second Circuit June 10, 2014, Decided). https://advance.lexis.com/api/document?collection=cases&id=urn:contentItem:5CD6-VH51-F04K-J015-00000-00&context=1516831

[2] Authors Guild v. Google, Inc., 804 F.3d 202, 2015 U.S. App. LEXIS 17988, Copy. L. Rep. (CCH) P30,832, Copy. L. Rep. (CCH) P30,832, 43 Media L. Rep. 2981, 116 U.S.P.Q.2D (BNA) 1423 (United States Court of Appeals for the Second Circuit October 16, 2015, Decided). https://advance.lexis.com/api/document?collection=cases&id=urn:contentItem:5H5B-G231-F04K-J02C-00000-00&context=1516831

[3] Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 209 L. Ed. 2d 311, 2021 U.S. LEXIS 1864, 2021 U.S.P.Q.2D (BNA) 391, 28 Fla. L. Weekly Fed. S 727, 2021 WL 1240906 (Supreme Court of the United States April 5, 2021, Decided). https://advance.lexis.com/api/document?collection=cases&id=urn:contentItem:62CD-04Y1-F8KH-X250-00000-00&context=1516831

is not limited to non-commercial or private contexts." (at ¶ 51). As such (and see below as well), "research" probably does include commercial research, although the intent of the researchers is clearly relevant.

Once the use is assigned to the general category of "research," we must look at the various factors typically assessed by Canadian case law. The following purposes were laid out in the *CCH* v. *LSUC* ruling: "[T]he purpose of the dealing, the character of the dealing, the amount of the dealing, the nature of the work, available alternatives to the dealing and the effect of the dealing on the work are all factors that could help determine whether or not a dealing is fair. These factors may be more or less relevant to assessing the fairness of a dealing depending on the factual context of the allegedly infringing dealing. In some contexts, there may be factors other than those listed here that may help a court decide whether the dealing was fair." (2004, at ¶ 60).

1) The purpose of the dealing: What is the user's motive? This factor looks at the user's "purpose or motive in using the . . . work" (¶ 54).

This is complicated. Most academic research is not directly commercial, but a lot of the use in the context of non-profit work is itself building data sets that will be used commercially. This may create situations where the motive changes over time. The boundary between non-profit and commercial research is fuzzy these days, and particularly in AI.

2) The character of the dealing: What was done with the work? Was it isolated or an ongoing use? How widely was the work distributed?

This is also complicated. The usage is often a single-use event (e.g., to fine-tune the language model), but the resultant parametrized model is repeatedly used, possibly with different prompts. The actual activity is usually to model sentence structure and how sentences flow from one sentence to another, but in practice, since the parametrization of a deep learning model is impossible to easily describe, it is entirely possible that chunks of the work will be directly "copied" into the parameters of the model, and may appear in the resultant generated texts. These occurrences of small chunks of copying would likely be an "insubstantial use,"[4] and therefore, not require copyright permission, but it is unpredictable whether or not they would happen, or how often. The copyrighted work in the corpus typically would not be distributed, but the parameterization, in the form of the structure and weights of a neural network model, typically would be.

3) The amount of the dealing: How much of the work was used? How important was the content that was used?

Typically, the entire work is used; the goal is to have as large of a corpus of domain-specific writing as possible.

4) Alternatives to the dealing: Could a different work have been used?

This, again, is complicated. The goal of these generative models is to use as rich of a corpus as possible: Using such a corpus allows for more of the natural flow in an author's writing to be modeled. So, if the goal is "as refined an approximation as possible," then, no, a different work could not have been used.

5) The nature of the work being used: Does dissemination aid the public interest?

No. The work is, in fact, not being disseminated via this kind of modeling; to the extent that it is found within the parametrization, recipients of the model could not reconstruct the original work.

6) The effect of the dealing on the original work: Does the use compete with the market of the original work?

This is possibly the most complicated question. In the next section, we look at this question in more detail, exploring a variety of computational creativity scenarios.

Competitive Use and Computational Creativity

To the question of "does the use compete with the market of the original work?," the answer is especially complicated. In the *CCH v. LSUC* (2004) ruling, it was made clear that this is not merely an abstract question: Copyright owners have to supply evidence of harm to their market because of the use in question. The onus of proving that dealing is fair is on the user, but such users typically lack the ability to see sales figures. "If there had been evidence that the publishers' markets had been negatively affected by the Law Society's custom photocopying service, it would have been in the publishers' interest to tender it at trial." (¶ 72). In a follow-up ruling, *Alberta (Ed)* v. *CCLA* (2012), the mere fact that sales have declined was not taken as sufficient proof that the use of copyrighted materials was a material factor in proving unfair use: "[O]ther than the bald fact of a decline in sales over 20 years, there is no evidence from Access Copyright demonstrating any link between photocopying short excerpts and the decline in textbook sales." (¶ 35).

With this context in mind, we can consider a variety of ways works might be used in corpora, and examine whether these uses would compete in the market with the original:

1) Automated news-writing systems. The system that is built certainly competes with the journalists whose work is being used, but likely is not competing with the specific articles of the corpus, since the news-writing system will create stories about events occurring after the articles it was trained on. The text humans write about individual news events may be copyrighted, but the facts themselves are not protected by copyright. The system may compete with the owners of the copyright (the newspapers in which the work appeared), as well: If a start-up using a computational news writer builds a corpus of articles from *The Globe and Mail*,

[4] Access Copyright - Tariff for Provincial and Territorial Governments, 2005-2014. [2015] Copyright Board of Canada. Retrieved from the Copyright Board website. https://decisions.cb-cda.gc.ca/cb-cda/decisions/en/item/366791/index.do

that start-up's product may compete with the *Globe*. Systems that write news editorials are closer to the creative writing systems discussed next.

2) Automated creative writing systems. The system that is built will create works in the same genre as the original work; in fact, they may be even designed to mimic a single author's style. As of 2021, such systems are mostly curiosities: They require quite a bit of editing to make song lyrics (Ackerman and Loker 2016) or newspaper op-eds, and again, the training data are about moments in the past, but they are progressing quite speedily. It is entirely possible that in certain genres or domains, consumers will (knowingly or not) purchase computer-generated works, trained on corpora of human-generated work in the same field, in preference to new or existing work by humans.

To that end, yes, the systems that result from corpora can compete with the market of the original work. We note that this overall space is huge: systems that generate text, music, visual art, dance patterns, and more. This suggests that if the training data are copyrightable, the same overall questions as we are raising in this paper probably apply more broadly to other domains.

3) Automated patent generation. The goal of these systems is to create a patent "autocomplete" system which, trained on a corpus of patents, can start with the preamble of a patent and generate text that belongs in the patent (Lee and Hsiang 2019). This is obviously a remarkable goal, but here, again, the extent to which the new inventions might compete with the existing ones is likely to become more of an issue in the future. More generally, automatic legal authoring systems are definitely starting to come into their own, drafting motions, for example (Hudgins 2020), and their work absolutely competes with the work of human lawyers.

4) Automated scientific paper generation. Again, this is a blue-sky idea. We would be delighted if we did not have to chase down the citations this short article requires, and instead, a computer did it for us. But in theory, one could write an abstract and the paper could write itself around the abstract. Previously, this has mostly been used for satire,[5] not for real research. Regardless, it is entirely possible that the existence of computationally generated papers would make the peer-review process collapse due to the required labour to assess all of those new papers, or that such papers might well compete with those in the corpus, particularly if the corpus was made of papers that "stand the test of time," or the like.

In this manuscript, we do not consider chatbots. In this case, the system probably does not use copyrighted materials much, and will be trained using transcripts of natural dialogues, or successful customer-service interactions, or similar sources. Chat bots are also not consistent with the rest of the frame of this article, as the model will involve a lot more discourse analysis and (often) more of a model of what customer-service interaction a human participant is seeking to resolve.

[5] For example, SCIgen (http://pdos.csail.mit.edu/scigen), an automatic CS paper generator, was used to generate a submission that was accepted to a predatory CS conference in 2005 (Ball 2005).

Conclusion

Overall, then, the question of competition between the result of a corpus' use and the corpus itself is complicated, but particularly thorny in the context of creative-writing systems. Ultimately, if computational creative-writing systems become successful, then such systems will be competing with exactly the producers (and quite possibly even the creative artifacts themselves) through which the systems were trained. Rather than buying a greeting card for your spouse's birthday, you might just send an automatically generated message, with a cheery computer-generated video based on a corpus of existing greeting cards. Rather than using existing pop songs as the soundtrack for a promotional video, you might use a "new" song, whose lyrics are produced by an engine trained on a corpus of existing pop songs. Rather than buying a copy of a Newbury winning book for your grandson's birthday, you might buy a book written by a computer, using a corpus of existing children's books. The possibilities are endless.

As such, producing corpora for training creative-writing systems will, over time, diminish the market for the copyrighted works in those corpora. Thus, building corpora of copyrighted materials for the purpose of training machine language models that compete in the same market as the training materials is unlikely to be fair dealing under Canadian law, particularly when it is practiced in commercial research.

This state of affairs places Canadian researchers at a disadvantage compared to researchers from other countries, most notably the U.S. If, as Sobel (2019) argues, U.S. law enables training of machine learning models from large collected corpora, and Canadian law does not, then researchers, such as the first author, must either gain permission to train models from copyright holders, use only materials with open licenses or in the public domain, or risk infringement lawsuits. We terminated a recent project, in part because gaining a good corpus of non-infringing materials was not easy. Either a change is needed in the Canadian copyright regime, or certain research may be chilled.

Acknowledgements

The work of authors DB and MRG is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Ackerman, M., and Loker, D. 2016. Algorithmic Songwriting with ALYSIA. arXiv:1612.01058

Alberta (Education) v. *Canadian Copyright Licensing Agency (Access Copyright)*, 2 SCR 345 (Supreme Court of Canada 2012). https://canlii.ca/t/fs0v5

Ball, P. 2005. Computer conference welcomes gobbledegook paper. Nature 434:946.

Brown, T.B., et al. 2020. Language Models are Few-Shot Models. arXiv:2005.14165.

Carroll, M.W. 2019. Copyright and the Progress of Science: Why Text and Data Mining Is Lawful. *UC Davis Law Review* 59:893-964.

CCH Canadian Ltd. v. *Law Society of Upper Canada*, 1 SCR 339 (Supreme Court of Canada 2004). https://canlii.ca/t/1glp0

Copyright Act, RSC 1985, c. C-42. https://laws-lois.justice.gc.ca/PDF/C-42.pdf

Copyright Act of 1976, 17 U.S.C. §107. https://www.copyright.gov/title17/

Craig, C. 2021. AI and Copyright. in F. Martin-Bariteau and T. Scassa, eds., *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021)

GPT-3. 2020, September 8. A robot wrote this entire article. Are you scared yet, human? *The Guardian*. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

DFG Review Board. 2017. *Guidelines for building language corpora under German law*. (E. Ketzan, J. Wilgans and J. Weitzmann, Trans.). https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdten/guidelines_review_board_linguistics_corpora.pdf

Hudgins, V. 2020. Casetext Launches New Brief-Writing Automation Platform Compose. https://www.law.com/legaltechnews/2020/02/25/casetext-launches-new-brief-writing-automation-platform-compose/

Jordanous, A. 2017. Has computational creativity successfully moved "Beyond the Fence" in musical theatre? *Connection Science* 29(4):350-386.

Lee, J.-S., and Hsiang, J. 2019. Patent Claim Generation by Fine-Tuning OpenAI GPT-2. arXiv:1907.02052

Radford, A., *et al*. 2019. Language Models are Unsupervised Multitask Learners. https://openai.com/blog/better-language-models/

Sobel, B. 2018. Artificial Intelligence's Fair Use Crisis. *Columbia Journal of Law and the Arts* 41(1):45-97.

Stross, C. 2013. Lovebible.pl. http://www.antipope.org/charlie/blog-static/2013/12/lovebiblepl.html

Uitdenbogerd, A.L. 2020, September 17. Can robots write? Machine learning produces dazzling results, but some assembly is still required. *The Conversation*. http://theconversation.com/can-robots-write-machine-learning-produces-dazzling-results-but-some-assembly-is-still-required-146090

Wilkinson, M. 2006. Legal aspects of compiling corpora to be used as translation resources. *Translation Journal,* 10(2). http://www.bokorlang.com/journal/36corpus.htm

World Intellectual Property Organization. 2019. The WIPO conversation on intellectual property and artificial intelligence. https://www.wipo.int/about-ip/en/artificial_intelligence/conversation.html