Some Notes on Writing

Mark Schmidt

January 14, 2024

It is a bit hard to precisely define what "writing clearly" means. But in writing about machine learning research the goal should be that the answer is "yes" to the following question from the Journal of Machine Learning Research's guidelines on "clarity": "Is it written in a way such that an interested reader with a background in machine learning, but no special knowledge of the paper's subject, could understand and appreciate the paper's results?".

Below are some common issues seen in technical writing, and some specific suggestions for writing clearly:

1. The number of grammar errors should be relatively small. I understand that some typos are inevitable (you all know that I make them all the time), but there should not be a typo in every second sentence. If you are very uncomfortable writing in English, then you may want to find a friend to help you improve. Improving in this area will likely help you no matter what career path you choose, so it's a worthwhile investment.

2. Define all acronyms at their first occurrence. Don't just start using "SVM", "CRF", and so on assuming that the reader knows what these are. The first time you use the acronym write it as "support vector machines (SVMs)". Yes, you can make acronyms plural by adding an 's' at the end.

3. In general, stick with active voice "X verbed Y" sentences rather than passive voice "Y was verbed by X" or the worse "Y was what X verbed"; for example, "this final project applies different machine learning models to XY-prediction" is much clearer than "XY-prediction is what the machine learning models that are used in this course project are applied to". Reading several pages of the latter makes your brain hurt. Also, do not be afraid to use "we" or "I", even though it sounds informal. Many of the best writers use "we"; they often even use it for single-author reports, since "we" can refer to the author and the reader.

4. The sentence structures should be simple. For example, if you have a comma-separated list nested inside a sentence, 99% of the time you can write it more clearly/unambiguously as two sentences where the list ends one of the sentences. An example of a sentence with a nested list is: "in this project we applied kernel-SVMs, random forests, and deep neural networks, which are all non-parametric classifiers, to the task of XY-prediction". You could replace this with "this project applies several non-parametric classifiers to the task of XY-prediction. The non-parametric classifiers we considered are kernel-SVMs, random forests, and deep neural networks". Here are some notes giving examples of using simple sentences:
`http://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages#Simple_sentence_structure`
You will be surprised at how a few simple modifications make your reading much easier for others to understand. (These tricks also work when you are speaking and thinking, too.)

5. Related to the above, do not be afraid of short sentences. This sentence is short but concise. Conciseness, without over-simplifying, is a goal to strive for; avoid saying simple things in complicated ways just because you are afraid of short sentences, as long sentences are harder to parse and you will probably forget what the start of the sentence was about if the sentence starts getting too long. That last sentence should have been several short sentences.

6. Avoid excessive comma usage. In general, if you don't have a list in the sentence then I recommend trying to have at most two commas. If the sentence ends with a nested list, then I recommend having no commas before the list. You can fix this issue by splitting large/complicated sentences into shorter and simpler ones. However, I recommend using the Oxford comma. It doesn't hurt readability and may help resolve ambiguity. I've also never seen a valid reason to use ", therefore," (at least one of these commas shouldn't be there).

7. Regarding capitalization: Support Vector Machine is the wrong way to capitalize SVM, it should just be support vector machine as it is not a proper noun. On the other hand, you would use Gaussian radial basis function (GRBF) because Gauss was a person. For the plural, you can just add an "s" to the acronym: "we trained several SVMs". The "s" is not capitalized. Another weird capitalization issue in English is that we capitalize "Section 3" where there's a number but we don't capitalize "the next section".

8. Try to keep the vocabulary simple. There are no extra points in academic papers for using fancy words, and it might actually hurt you if non-native speakers have more trouble understanding your work.

9. Use hyphens to resolve disambiguities in adjective lists. For example, "the differentiable strongly convex function f" could be ambiguous: is f "differentiable-strongly" or is it "strongly-convex"? Add the hyphen to resolve this: "the differentiable strongly-convex function f".

10. Mathematical concepts should be described with a mixture of words and mathematical symbols. If you only have symbols, you assume the reader knows what everything means and your paper will be incomprehensible to people that use different notation. If you only have words, you rely on the reader filling in a lot information for you and your work will be much less reproduceable.

11. Mathematical symbols and displayed equations should be readable as part of a sentence. This includes adding punctuation to your equations. As an example consider the finite sum problem,

$$\operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

    When you read the paper, the symbols above are actually the second part of the sentence. We can similarly have equations like

$$f(x) = \|Ax - b\|^2,$$

    that are placed in the middle of a sentence (in this case use a comma). You should be able to *read* all symbols as a natural part of the sentence. If you can't read the symbols as you go along, then the math in your work is probably too dense and you should consider making it more reader-friendly.

12. x should not appear at the start of the sentence. Displayed math should never start a sentence.

13. New concepts, that would not be familiar to a general machine learning audience, should be defined where they first occur. We should also mention the relevance of new concepts as they are introduced.

14. I recommend using an "author-year" citation style. So when citing the Wolpert "no free lunch paper", it would appear in the text as "Wolpert [1996]" or as "[Wolpert, 1996]". The LaTeX package "natbib" allows you to use bibliography styles like this, and supports the commands \citep and \citet which generate "[Wolpert, 1996]" and "Wolpert [1996]" respectively. You use \citep when the citation is not part of the sentence and \citet when it is part of the sentence, don't confuse these.

15. Related to the above, a citation like "[Wolpert, 1966]" or "[45]" is not a noun and should not be part of the sentence. Your document should not say "in [Wolpert, 1996]" or "by [45]", but instead should say "in Wolpert [1996]" or "by Wolpert [45]". The correct usage of the \citep citation style would be a sentence like "no model obtains the smallest generalization error for every problem [Wolpert, 1996]" or with numbered citations you would use "no model obtains the smallest generalization error for every problem [45]".

16. I can't believe how often I see references cited without a space before the citation. Wrong: "stochastic gradient method[12]". Right: "stochastic gradient method [12]". There should be a space, do not forget the space, where have you seen people leaving out the space that you think there should not be a space?

17. Some people don't think we should use contractions in scientific writing, like replacing "do not" with don't. The argument is that it makes the writing more informal, and some contractions like "y'all" defintely sound informal. But I don't believe that using standard contractions hurts the clarity of the writing.

18. Any time you want to write "e.g.", replace it with "for example". Any time you want to use "i.e.", replace it with "in other words". You aren't saving blackboard space, and are putting more cognitive load on the reader.

19. Paragraphs have at least three sentences. If your paragraph has one or two sentences, then either combine it with the next paragraph or split those 1-2 sentences into more sentences.

20. You'll need to learn to use "determiners" in the right way. These are the nastiest part of speech in English, and unfortunately I don't think there is an easy way to pickup all the rules/exceptions regarding these nasty things without a ton of practice. Here is an attempt at giving some rules: `https://grammar.yourdictionary.com/parts-of-speech/nouns/what/what-is-a-determiner.html`

21. Don't do "this" with your quotation marks.

22. Footnotes go after punctuation.[1]

23. Check capitalization in the bibliography. In your bibtex file you can capitalize bibliography letters using braces. For example, use {SVM} so that it appears as SVM and not svm.

24. I often see papers jump into describing technical details, without first discussing *why* what they are describing is important/relevant. If you want the reader to understand your writing, you don't want the reader to be confused about why you are discussing a particular topic.

25. A guide to "that" vs. "which" (and "who"): `http://www.kentlaw.edu/academics/lrw/grinker/LwtaThat_Versus_Which.htm`

26. The expression "$f(x)$" is not a function: f is the function and $f(x)$ is the value of the function evaluated at x.

27. Do not say "it is easy...". Readers can come from wildly different backgrounds, and what might be easy for one person may be easy for others. A replacement for something like "it is easy to show" would be something like "it can be shown" (and you should probably give hints about how is shown).

The goal is that an informed reader can understand the most important points of the paper, by reading it once from start to finish. You will limit the impact of your work if the reader has to read your paper twice to understand what the main points are, as they might just read someone else's paper instead of reading your paper a second time.

Related Notes

Tamara Munzner has a set of related suggestions here:

`https://www.cs.ubc.ca/~tmm/writing.html`

Maryam Kamgarpour has a set of guidelines related to writing style here:

`https://ece-kamgarpour-2019.sites.olt.ubc.ca/files/2020/08/ScientificCommunication.pdf`

[1] Like this.