Some Guidelines for Writing Scientific Papers

Aris Anagnostopoulos

1 Introduction

I have written these guidelines because I have found that many M.Sc. and new Ph.D. students tend to do some common mistakes when writing papers. There exist several books that one can consult, but I wanted to create a short list with the most common errors.

These are either rules for correct English, or accepted style for scientific publications. Writing is, to a large extent, a matter of style, so different editors or writers may provide slightly different recommendations, and some may be different from the ones written here. Nevertheless, the advice is to follow the following guidelines and deviate only if you have a good reason to; for instance, because you read it in some book and it suits better your writing style. Feel free to send me any errors, questions, comments, or objections.

I begin with some comments for English syntax and grammar and then I write about LaTeX. In Section 4, you can find some books that I recommend for improving your writing.

2 English Writing

1. **Active and passive voice.** Use the active voice. The number one mistake of inexperienced writers is to use on purpose the passive voice. This makes the text less clear and harder to read. Do not write

   Clustering is sometimes used for dimensionality reduction.

   but instead write

   Practitioners sometimes use clustering for dimensionality reduction.

   There are some cases that passive voice is the better choice, when you want to emphasize the object in the sentence, but this is not very common, in my experience.

2. **Long sentences.** Avoid making sentences that are very long and sentences that are not to the point. The primary goal in scientific writing, whether it regards an article or a thesis, is to convey a message and, unlike in literature writing in which long sentences are sometimes desirable (e.g., Beckett, Faulkner, and Garcia Marquez have written sentences that are longer than one thousand words), using them in technical writing may increase unnecessarily the effort of the reader, who may have a short attention span after having reviewed various papers for the already passed review deadline, leading to a reduced comprehension of the work, and it is for this reason that this sentence must be shortened and rephrased.

3. ***That* and *which*.** *That* and *which* are not interchangeable. We use *that* when the phrase that follows defines the word to which it refers. This is also the reason that *that* is, typically, not preceded by comma. Instead, *which* provides additional information about the object, which is not necessary for defining the word, and for this reason *which* is preceded by a comma. Typically we use *which* when we can remove the secondary phrase containing *which* leaving a sensible sentence. Compare the following two examples:

> *The algorithm that solves the single-source shortest-path problem in a graph $G = (V, E)$ has running time $O(|E| + |V| \log |V|)$.*
>
> *Dijkstra's algorithm, which solves the single-source shortest-path problem in a graph $G = (V, E)$, has running time $O(|E| + |V| \log |V|)$.*

In the first example we use *that* because the phrase that follows specifies the algorithm to which we refer. In the second one we use *which* because we insert some parenthetical information about what Dijkstra's algorithm does. Think that the phrase containing *which* can be omitted and the sentence will still make sense; instead, the phrase containing *that* is necessary to understand to what the verb refers.

4. **Using terms literally.** There are some words that are commonly used in a way that is not very accurate or even wrong. Prefer the correct word. Here are some examples, where the word in the first column is often misused.

Word/phrase	Better alternatives
since	because—*since* is acceptable but *because* is more accurate; *since* means "from the time that"
while	and, whereas, although, even though—*while* is acceptable, but these are better alternatives; *while* means "at the same time"
due to	because of
if	whether—when the meaning is that one of two options or cases are possible, you should use *whether*: *We carried out experiments to test whether our model reflects the system's behavior.* when you want to express "under a condition," you should use *if*: *If your data have a lot of noise, you can use PCA to clean them.*
in order to	to, so that—*in order to* is not wrong, but usually it is unnecessary
like	such as—very often *like* is being used instead of *such as*; you use *like* to compare; you use *such as* to provide an example
continuous	continual—*continuous* means not discrete, *continual* means without interrupting
less	fewer—do not use *less* when referring to discrete items

Note that there are, of course, situations that the word in the first column is the correct one. Consider the following phrase:

> *Since Facebook bought WhatsApp, the number of WhatsApp active users quadrupled.*

This sentence is correct if the intended message is: *From the time that Facebook bought WhatsApp, the number of WhatsApp active users quadrupled.* However, you should not use *since* if the desired meaning is that the increase of WhatsApp users was a result of Facebook

buying it; instead you should, for example, write: *Because Facebook bought WhatsApp, the number of WhatsApp active users quadrupled.*

5. **Acronyms.** If you want to use an acronym then define it the first time that you use it:

   We used a convolutional neural network (CNN) architecture.

   After that you should use the acronym consistently.

   - Some acronyms are so common in the field that you should use them directly (e.g., CPU). If you are not sure, then it is better to define them; for example, are you 100% sure that all the readers of your paper will understand immediately that CNN means a convolutional neural network?
   - Make sure that you do not repeat words implied in the acronym. Do not write

     We used a CNN network architecture.

   You may also use common Latin acronyms *i.e.*, *e.g.*, and *etc.*; however, reserve them for use inside parentheses (i.e., explaining or providing examples), and you should write a comma afterwards (e.g., as in this phrase). In running text you should use full-phrase alternatives: *that is*, *for example*, *for instance*, *and so on*. The only exception that I believe is fine to use in text is *et al.*, because trying to use alternatives (e.g., *Kempe and colleagues [17] defined the problem of influence maximization in graphs.*) may make the section of *Related Work* cumbersome.

6. **Hyphenation.** In typography there are four types of dashes:

   - Hyphen: - (`-` in LaTeX)
   - En dash: – (`--` in LaTeX)
   - Em dash: — (`---` in LaTeX)
   - Minus sign: − (`$-$` in LaTeX)

   You should learn when to use each of them. Hyphen is the most common, so here I list some examples of its use:

   (a) The hyphen is being used when two words form a compound adjective. Compare the following two phrases.

The project used a machine-learning approach.
Luca has become very strong in machine learning.

In the first example the meaning is

The project used a (machine learning) approach.

so the hyphen makes this clear. Omitting the hyphen could imply that the desired meaning is

The project used a machine (learning approach).

In some cases, the two words are so commonly used together, that I believe it would be acceptable to omit the hyphen. So it would be acceptable to write

The project used a machine learning approach.

as it is clear what we mean; I still prefer to use the hyphen, though.

You should not put a hyphen when the first of the two words finishes in *-ly*:[1]

We modeled the group interactions as a fully connected graph.

There are cases when the compound has more than two words, as in the following structure.

IBM designed a new ((machine learning) based) recommender system.

Then it is correct to use an en dash for the outside parenthesis:

IBM designed a new machine-learning–based recommender system.

However, in such cases it may be better to rephrase:

IBM designed a new recommender system based on machine learning.

[1]Some editors recommend to avoid using the hyphen also if the first word is one of is one of *more*, *most*, *less*, *least*, and I also tend do that: *The algorithm that Francesca designed keeps in memory the most frequent items.*

(b) Other uses of hyphens regard prefixes, that is, adding parts such as *bi*, *multi*, *co*, *non*, and many others. There you can use the rule of the thumb not to use the hyphen but unite the prefix as a single word, unless this would create confusion in the meaning or the reading of the word, when two i's will be next to each other, or when the second word is capitalized. For example, you should write the words as follows:

nonnegative, biweekly, multiparty, coclustering, semiconductor, superpower, semi-induced, anti-Trump, co-op, re-sent, re-cover, re-creation

In the last three examples, we use *co-op* (cooperative) instead of *coop* (cage), *re-sent* (sent again) instead of *resent* (bitterness), *re-cover* (cover again) instead of *recover* (get better), and *re-creation* (a new creation) instead of *recreation* (pleasant activity), assuming of course that these are the meanings we want to convey.

(c) Some authors use the hyphen incorrectly in phrases like the following:

The algorithm selects the i-th element of the list.

You should write ith element, or, if you really do not like it, i'th element.

When you connect two words that are at the same level, you should use the en dash and not the hyphen:

We designed a new algorithm for predicting gene–disease associations.

The Rabin–Karp algorithm is a string-searching algorithm based on hashing.

In the last example, the first dash is an en dash; the second is a hyphen.

7. **Capitalization.** *Capitalization* refers to writing the first letter of a word in capital, independently of its location in a sentence. Capitalization in running text creates clutter and should be avoided. There is no reason to capitalize any of the words—except for the first one—in the following example:

Convolutional Neural Networks have been very successful in Computer Vision.

Instead you should write:

Convolutional neural networks have been very successful in computer vision.

I recommend capitalization in the titles of papers and sections, as most publishers request this. Here are some general guidelines to follow:

- Capitalize the first and last word of the title, unless, for example, it is a lower-case variable name.
- Capitalize words with at least five letters.
- Do not capitalize prepositions and articles, connectives (e.g., *and*, *or*) but do capitalize other small words (e.g., *Is*, *You*, *It*).
- Consider hyphenated compounds as separate words for the purpose of capitalization (e.g., *Depth-First Search*).

8. **Serial (or Oxford) comma.** When you itemize a list of items of at least three items use a comma before the *and* or the *or*:

   > *We divided our data set into training, validation, and test set.*
   > *Broder [2] classified web-search queries as navigational, informational, or transactional.*

9. **Citations.** A citation is not considered part of the text. You should not write

   > *Given the increased performance of attention mechanisms, [7] introduced the transformer architecture.*

   but

   > *Given the increased performance of attention mechanisms, Vaswani et al. [7] introduced the transformer architecture.*

   Sometimes in conference submissions, in extreme cases of space you may want to do it, but you should be aware that is a mistake and, you should try to fix it in the camera-ready version.

10. **Addressing yourself and the reader.** Do not address to you as *the authors* and to the reader as *the reader*:

    > *The authors recommend to execute the code with at least 16 GB of RAM.*

    Instead, use *we* when you refer to the authors.

    > *We recommend to execute to code with at least 16 GB of RAM.*

    In the case of a single author you can use *we* or *I*; most authors prefer *we*.

    Most of the times you do not need to refer directly to the reader, but in case you do, I think it is fine to use *you*.

    > *We recommend that you execute the code with at least 16 GB of RAM.*

11. **Gender.** As the world becomes more politically correct, you may want to avoid using only one gender in descriptions and examples. I like to use both *he* and *she* when referring, for example, to the *user* (but be consistent: If you use *she* then continue using *she* if you refer to the same user.). Sometimes this is useful as it can also help disambiguate: you may use *he* for the *seller* and *she* for the buyer (or vice versa).

12. **Tenses.** Be consistent with the tense that you use when you describe the work carried out; use either present or past, not both. Avoid descriptions like the following:

    > *We study the effect of cache size on our algorithm's performance. Our experiments showed that the running time decreases linearly in the cache size.*

3 LaTeX

The general advice regarding LaTeX is to invest some time to learn how to use it properly, by going through some book and then consulting it when you write. In Section 4 I recommend you one. Here I have listed a few specific suggestions for new LaTeX users, based on common errors I have observed:

1. Do not finish a paragraph with \\. Instead, leave an empty line.
2. Use `\emph{}` instead of `\textit{}` or `{\it }`.
3. To write opening and closing double quotes you should type twice ` to open quotes and twice ' to close them. For example, to type "food" you must type ``food''. These are different than the keyboard's double straight quotes symbol ".
4. Learn about the tilde symbol ~ and use it. In particular, don't combine it with a space character, tilde itself is a space.
5. Study the *amsmath* package and use it to display equations properly.
6. Sooner or later you will need to have subfigures in a figure. There exist various packages, and some old ones created compilation problems. I recommend the advice given at: `http://www.peteryu.ca/tutorials/publishing/latex_captions`. Most of the times, I use:

```
\usepackage{caption}
\usepackage{subcaption}

\captionsetup[subfigure]{labelformat=simple}
\renewcommand\thesubfigure{(\alph{subfigure})}
```

4 Books on Writing

There are some books that you can consult to improve your writing:

- *BUGS in Writing: A Guide to Debugging Your Prose* by Lyn Dupré: This books collects multiple errors that scientists do when they write and gives some excellent advice. It is annoying at times but, nevertheless, it is easy to read and your writing will improve a lot if you read it. I have based a lot of my advice on information from this book.
- *Writing for Computer Science, Third Edition*, by Justin Zobel: This book is mostly targeted to computer scientists. It provides excellent advice on various topics related to writing, including style, how to write mathematics, how to describe algorithms, and how to handle figures; highly recommended.
- *The Elements of Style, Fourth Edition*, by William Strunk Jr. and E. B. White: This is very short, so you can finish it in a few hours. It is addressed to writers in general, and it is older (the first edition was published in 1935), but most of the advice is useful for current scientific writing.

- *The Chicago Manual of Style, 17th Edition*: Sometimes we want to check how to write a particular phrase: *Do I need to put a hyphen between these two words? Should I put the period before or after the closing double quotation?* This book is a reference (1146 pages), with generally accepted rules by (American) publishers, and you can consult it when in doubt.

For LaTeX, I like the book of Kopka and Daly, *A Guide to LaTeX, Fourth Edition*, which you can find online at Daly's website: `https://www2.mps.mpg.de/homes/daly/GTL/gtl_20030512.pdf`