Specialist English: Assignment 3
(solutions)

Rebecca J. Stones
`rebecca.stones82@nbjl.nankai.edu.cn`

October 26, 2017

Here's my solutions; your solutions needn't be identical. There may be problems with the sample snippets I haven't listed.

Problem 1 This is the first snippet:

> To support runs of large-scale parallel applications in which a shared file system abstraction is used, Zhang et. al. [15] developed a scalable MTC data management system to efficiently handle data movement.

In the citation "Zhang et. al. [15]" there's two problems: "et al." is mistyped and the spacing after the periods is too long (LaTeX interprets this as the end of a senence). It should be typeset `Zhang et al.~[15]` or something similar. (I can't find any other problems with this sentence.)

Comment:

- Some students still got this wrong! Please learn how to spell "et al." and never get it wrong again in your whole life. It's a very common mistake.

This is the second snippet:

> To overcome existing problems of current image search engines. A natural solution is to enable users to flexibly express their intentions by providing them a pen and a drawing panel.

Here, the first sentence is incomplete, and should be merged into the second one. The best solution is the obvious solution: to replace "To overcome existing problems of current image search engines. A natural solution ..." with "To overcome existing problems of current image search engines, a natural solution ...". (I can't find any other problems with this sentence, although I'm unsure how "natural" this solution is.)

This is the third snippet:

> The CCM model[9] assumes that user starts the examination of the search results from the top ranked document.

There should be a space between "CCM model" and "[9]", that is, it should be changed to "CCM model [9]". (I can't find any other problems with this sentence.)

Comment:

- Different journals and conferences use different styles (or don't have a fixed format). This particular paper ordinarily adds a space, so this is an example where it is a mistake.

Problem 2 This is the snippet in question:

> The storage for edges in X-Stream is 2.1X-4.5X that of storage in PathGraph. However, the storage of GraphChi for edge relation is 0.72X-1.51X that of PathGraph's storage. This is because GraphChi stores one copy of edges.

There's three things we want to say:

1. The edge-storage cost in X-Stream is 2.1 to 4.5 times as large as that in PathGraph.
2. The edge-storage cost in X-Stream is 0.7 to 1.5 times as large as that in GraphChi.
3. Item 2 above is because GraphChi stores only one copy of each edge.

It is important to use the same wording in items 1 and 2—the use of repetition indicates to the reader that we're making the same comparison but between two different packages.

> ... expressions of similar content and function should be outwardly similar ...
>
> — Strunk & White, Section III.15

There's various ways to combine these:

The edge-storage cost in X-Stream is 2.1 to 4.5 times as large as that in PathGraph. However, the edge-storage cost in X-Stream is 0.7 to 1.5 times as large as that in GraphChi, which is because GraphChi stores only one copy of each edge.

The edge-storage cost in X-Stream is 2.1 to 4.5 times as large as that in PathGraph, and 0.7 to 1.5 times as large as that in GraphChi; this is because GraphChi stores only one copy of each edge.

Comments:

- Most importantly, almost everyone "tweaked" the original snippet (making small improvements here and there), rather than **rewrite** it.

  > ... you must write it, rewrite it, and re-rewrite it several times ...
  >
  > — Halmos, *How to write mathematics*, 2009.

  To write papers well, we need to rewrite from scratch our own writing. We need to lose any personal attachment we have to the original writing—we simply do whatever is needed to optimize it (including rewriting it from scratch).

- It's a mistake to add zeros to say "2.10 to 4.50". (a) We don't know this is correct, and (b) for this sentence, the reader does not need this level of precision.
- The hyphen in "X-Stream" is correct, whereas the dashes in "2.1X-4.5X" and "0.72X-1.51X" are incorrect. I didn't want to be too specific when saying "the dashes have the wrong length".
- Replacing "2.1X-4.5X" with "2.1X–4.5X" or "2.1x–4.5x" does not fix the problem—it's still a letter x. Replacing it with "2.1×–4.5×" (or "2.1–4.5×") is better, but is still not the best choice since × is used as a shorthand for the word "times". (Like how ∀ is shorthand for "for all".)
- We can write "only one copy of each edge" or "only one copy of the set of edges". The first one is best as it's more succinct, and doesn't have the problem of determining if "set" or "list" is a better word.
- It might be even better to be more specific, e.g. "... uses 2.1 to 4.5 times as much memory as ...", but maybe this is clear from the context of the paper.

Problem 3 This is the snippet in question:

Represented by

$$\{(w_i, \mathrm{FID}_{w_i})\}_{w_i \in W} \tag{1}$$

it is a normal per-keyword index that contains $|W|$ keyword fileIDs pairs such as $(w_i, \mathrm{FID}_{w_i})$, where w_i is an arbitrary keyword and FID_{w_i} is the identifier set of files that contain w.

Problems:

1. This snippet is passive; writing actively is generally recommended: “We do X” instead of “X is done”.
2. “Represented by ... it is ...” is ungrammatical.
3. The equation number is unused—it should not be there.
4. The phrase “keyword fileIDs pairs” is ungrammatical; it’s best to use a compound adjective “keyword-fileID pairs”.
5. The phrase “such as” is not needed.
6. The mathematical notation implies it has size $|W|$, so we don’t need to restate this.
7. The mathematical notation implies a typical element is denoted $(w_i, \mathrm{FID}_{w_i})$.
8. We have a subscript of a subscript in “FID_{w_i}”; it’s best to avoid these for readability, if possible.
9. The index i is unused, and missed in w at the end.
10. The words “normal” and “arbitrary” seem to be used needlessly. Moreover, saying “w_i is an arbitrary keyword” is problematic, as w_i is restricted to elements in W.
11. The phrase “identifier set of files that contain w” is ungrammatical; it should be “the set of fileIDs for files that contain w” or “the set of fileIDs of files that contain w”. Particularly important, is that FID_{w_i} contains fileIDs (not files) and writing “[blah] set of files” is ordinarily going to be interpreted as a set of files.

What we want to tell the reader is:

1. We use $\{(w, \mathrm{FID}_w)\}_{w \in W}$ to denote an index.
2. In this notation w is a keyword and FID_w is the set of fileIDs for files that contain w.

So we write:

For a set of keywords W, we use $\{(w, \mathrm{FID}_w)\}_{w \in W}$ to denote its index, where FID_w is the set of fileIDs for files that contain w.

Or if we prefer:

For a set of keywords W, we use $\{(w, \mathrm{FID}_w) \colon w \in W\}$ to denote its index, where FID_w is the set of fileIDs for files that contain w.

In this case, to be consistent, we’d have to change the notation throughout the whole paper.

Comments:

- In this problem, if we “tweak” instead of “rewrite”, we won’t even get close to optimized writing (at best, we’ll find a “local minimum”).
- It seems a lot of students were unfamiliar with the notation $\{(w_i, \mathrm{FID}_{w_i})\}_{w_i \in W}$ as an alternative for $\{(w_i, \mathrm{FID}_{w_i}) \colon w_i \in W\}$; this resulted in introducing all sorts of errors.
- Some students made a mistake in replacing FID_{w_i} (typeset e.g. `$\mathrm{FID}_{w_i}$`) with FID_{w_i}, which is equal to $F \times I \times D_{w_i}$.

  Compare this to $\log n$ and $\sin x$, typeset `$\log n$` and `$\sin x$`, respectively.
- In my examples, I don’t say that w is a keyword; we don’t need to since we say $w \in W$, and W is a set of keywords.
- Some students changed “keyword fileIDs pairs” to “keyword and fileID pairs”. To illustrate how this is ambiguous, consider these two sets:

  - $\{(\text{boy}_1, \text{boy}_2), (\text{girl}_1, \text{girl}_2)\}$, and
  - $\{(\text{boy}_1, \text{girl}_1), (\text{boy}_2, \text{girl}_2)\}$.

  Both can be described as consisting of "boy and girl pairs".

  And possibly this could also be considered as consisting of "boy and girl pairs":

  - $\{\text{boy}, (\text{girl}_1, \text{girl}_2), (\text{girl}_3, \text{girl}_4)\}$.

- Doug West recommends $\{(w, \mathrm{FID}_w)\colon w \in W\}$ with a colon (which is clearest when typeset with a `\colon`) instead of $\{(w, \mathrm{FID}_w) \mid w \in W\}$ (typeset `\mid`) or $\{(w, \mathrm{FID}_w)|w \in W\}$ (typeset `|`). This is because the notation $|$ is used in several other ways, such as cardinality (e.g. $|W|$), divisibility (e.g. $5|n$)[1], and evaluation (e.g. $\frac{dy}{dx}|_{x=0}$). (I agree with West's advice here.)

Some additional tips:

- When introducing notation it's best to use the word "denote": both "notation" and "denote" have the same root. However, sometimes using "let" or "define" results in a simpler sentence, and should be preferred. (The word "represented" has several meanings in mathematics[2], and is best avoided except in these circumstances)

- Likewise, when introducing a table, I recommend the word "tabulate".

- For figures, I find it best to say "plots", "illustrates", or "depicts". (I avoid saying a figure "shows" something, as this word can also mean "demonstrate".)

[1] Although I strongly recommend using the word "divides" instead of $|$.

[2] E.g., an integer divisible by 5 can be represented by $5k$ for some integer k.