CS2309 CS Research Methodology

Writing[1]

Lee Wee Sun
School of Computing
National University of Singapore
leews@comp.nus.edu.sg

Semester 1, 2011/12

[1] The material in these slides is mostly extracted from *Writing for Computer Science* by Justin Zobel.

Outline

- Organization
- Content
- Language

Scope

- When writing, plan your content. Ask questions such as
  - Which results are the most surprising?
  - What is the one result that other researchers might adopt in their work?
  - Are the outcomes interesting enough to justify being included?
  - Should you explain the new work first, followed by previous works? Or is the contribution more obvious if old approaches are explained first to set the context?
  - What assumptions and definitions need to be formalized before presenting the main theorem?
  - What are the key background works that need to be be discussed?
  - Who is the readership?

Structure

- Typically, in a scientific paper, you need to:
  - Describe the work in the context of accepted scientific knowledge.
  - State the idea being investigated, often as theory or hypothesis.
  - Explain what is new, what is being evaluated, or what the contribution is.
  - Justify the theory by methods such as proofs or experiments.

- The paper should have a logical flow that feels like a narrative.
- You should re-read and re-organize your writing until it has a logical structure.
- A common structure is a chain
  - Problem statement is first given.
  - Then a review of previous solutions and drawbacks.
  - A new solution is proposed.
  - A demonstration that the new solution is an improvement.
- Another structure is by specificity.
  - Probably the most common in scientific papers.
  - Material is outlined first in general terms.
  - Details are filled in progressively in sections.

- Another structure is by example.
  - Idea or result is initially explained by applying to typical problem.
  - The idea is the explained more formally, filling in the details of the method.
- Another structure is by complexity.
  - A simple case is first given.
  - A complex case explained as an extension.
- Regardless of structure used, always ensure that the writing can be well understood when read as a sequence - every item should be motivated and defined before used.

Paper Components

- Title and author
- Abstract
  - Concise summary of aims, scope and conclusions.
  - Allow readers to judge whether the paper is relevant to them.
- Introduction
  - Expanded version of the abstract.
  - Describe topic, problem, approach, scope, limitation and outcomes.
  - Enough detail for reader to judge whether to read further.

- Body
  - Present the results.
  - Provide the background and terminology, explain the reasoning that led to the conclusions, provide details of proofs, experiments.
  - Usually broken up into sections.
  - Body can be long - clear narrative flow and logical structure essential.
- Literature Review
  - Used to compare new results to similar previous results.
  - Explain how results are extended.
  - Reviews how existing methods differ from one another.
  - Can be early in the paper - in or after the introduction, or after the body.

- Conclusions
  - Concise statement of important results and explanation of significance.
  - State limitations of the work.
  - Look beyond current context to problems not addressed, potential variations, consequences.
- Bibliography
  - Include items cited in text.
  - No other items should be included.
- Appendices
  - Bulky material that would otherwise interfere with the narrative flow of the paper.

Links and Ordering

- The structure of a paper often needs to be communicated.
- Brief summaries at start and/or end of sections are often helpful
  - √ Together these results show that the hypothesis holds for linear coefficients. The difficulties presented by non-linear coefficients are considered in the next section.
- The connection between one paragraph to the next should be obvious - tell the reader what you are going to say, say it, and then tell the reader that you have said it.
- Common error is to include definitions/theorems without indicating why the material is useful - often an ordering problem.
- Motivate the reader at each major step: explain why the theorem/definition is going to be used, why it is interesting or how it fits into the structure of the paper.
- Everything that is not common knowledge (depending on audience) should be explained.

Paragraph

- A paragraph should contain discussion on a single topic or issue.
- First sentence typically summarizes the argument.
- Following sentences elaborates.
- Every sentence should be on the same topic.
- Last sentence often have higher impact than those in the middle of paragraph.

- Long paragraphs may include several ideas - if a long paragraph can be broken, break it.
- Context can be lost between paragraphs, making references hard to follow.
- Link them by reusing key words or phrases and by using expressions that connect one paragraph to the next. Words and phrases like *again*, *therefore* and *for the same reason* are useful for this.
- Example: if a paragraph refers to a fast sorting algorithm, the next paragraph should begin with "The fast sorting algorithm" rather than "This algorithm".

Outline

- Organization
- Content
- Language

- Try to make the paper as self-contained as possible (within space limitation)
  - Include as much of the background, relevant material as possible.
  - All items should be properly motivated and defined.
- Be careful to distinguish existing knowledge and the paper's contributions
  - × Many user interfaces are confusing and poorly arranged. Interfaces are superior if developed according to rigorous principles.
  - √ Many user interfaces are confusing and poorly arranged. We demonstrate that interfaces are superior if developed according to rigorous principles.

Examples

- Use of examples can often help clarify concepts, particularly when the concepts are abstract.
  - √ In a semi-static model, each symbol has an associated probability of representing its likelihood of occurrence. For example, if the symbols are characters in text, then a common character such as "e" might have an associated probability of 12%.
  - √ Large document collections, such as a repository of newspaper articles, can be managed with the same techniques.
  - √ Special cases, such as the empty set, need to be handled separately.
  - √ Algorithms that involve bit manipulation cannot be efficiently implemented in these languages. For example, Huffman coding is impractical because it involves processing a stream one bit at a time.

Straw men

- A straw man is an indefensible hypothesis that an author describes for the sole purpose of criticizing it.
- Contrast should be between the new and the current, not the new and the fictitious.
  - × Query languages have changed over the years. For the first database systems there were no query languages and records were retrieved with programs. Before then data was kept in filing cabinets and indexes were printed on paper. Records were retrieved by getting them from the cabinets and queries were verbal, which led to many mistakes being made. Such mistakes are impossible with new query languages like QIL.
- A straw man is an example of rhetoric - of attempting to win an argument through presentation rather than reasoning.

- Other forms of rhetoric are appeal to authority, appeal to intuitively obvious truth, and presentation of received wisdom as fact.
  - × We did not investigate partial interpretation because it is known to be ineffective.
    - If there is evidence, then it should be cited.
- Unsubstantiated claims should be clearly noted as such.
  - × Most users prefer the graphical style of interface.
  - √ We believe that most users prefer the graphical style of interface.
  - × Another possibility would be a disk-based method, but this approach is unlikely to be successful.
  - √ Another possiblity would be a disk-based method, but our experience suggests that this approach is unlikely to be successful.

Obfuscation

- To obfuscate is to make statements ambiguous or convoluted with the intention of hiding the true meaning.
- For example, it can be used to give the impression of having done something without actually claiming to have done it.
  - × Experiments, with the improved version of the algorithm as we have described, are the step that confirms our speculation that performance would improve. The previous version of the algorithm is rather slow on our test data and improvements lead to better performance.
- It is preferable to be specific.
  - × Amelioration can lead to large savings.
  - √ Amelioration led to savings of 12%-33% in our experiments.

- Use of long-winded sentences can obfuscate.
  - × The status of the system is such that a number of components are now able to be operated.
  - √ Several of the system's components are working.
  - × In respect to the relative costs, the features of memory mean that with regard to systems today disk has greated associated expense for the elapsed time requirements of tasks involving access to stored data.
  - √ Memory can be accessed more quickly than disk.

Reference (cf. Plagiarism Section in Introduction Lecture)

- References are used to explain the relationship of your new work to existing work.
- They demonstrate that the work is new. They demonstrate your knowledge of the area, and they provide pointers to background reading.
- Commonly used style include the ordinal-number style and the name-and-date or Harvard style.
  - √ ... is discussed by Whelks and Babb (1972).
    ... is discussed elsewhere (Whelks and Babb 1972).
    ... is discussed by Whelks and Babb [1972].
    ... is discussed elsewhere [Whelks and Bann 1972].
- When there are more than three authors, all but the first can be replaced by "et al."
  - √ Howers, Mann, Thompson, and Wills [9] provide another example.
  - √ Howers et al. [9] provide another example.

Outline

- Organization
- Content
- Language

Style

- Guidelines and rules of thumb for writing.
- Not to be followed strictly, often a matter of taste.
- However, there are usually good reasons for the rules and you should follow them unless there is a reason not to.
  - Conforming to common styles reduces effort required from readers.
  - Good style makes writing more interesting to read.
- If your writing is not particularly good, you should generally follow the guidelines.

Starting a Paper

- Opening paragraph of a paper can set a reader's attitude.
- Begin well to make the best possible impression.
- Opening sentence should be direct.
  - × The paper does not describe a general algorithm for transactions.
  - √ General-purpose transaction algorithms guarantee freedom from deadlock but can be inefficient. In this paper we desribe a new transaction algorithm that is particularly efficient for a special case, the class of linear queries.
  - × Trees, especially binary trees, are often applied – indeed indiscriminately applied – to management of dictionaries.
  - √ Dictionaries are often managed by a data structure such as a tree, but trees are not always the best choice for this application.

- Establish context of the paper
  - × In this paper we describe a new programming language with matrix manipulation operators.
  - √ Most numerical computation is dedicated to manipulation of matrices, but matrix operations are difficult to implement efficiently in current high-level programming languages. In this paper we describe a new programming language with matrix manipulation operators.
- The opening sentence should indicate the topic clearly
  - × Underutilization of main memory impairs the performance of operating systems.
  - √ Operating systems are traditionally designed to use the least possible amount of main memory, but such designs impairs their performance.
    - The second sentence is clear, states the context and is positive.

Economy

- Remove unnecessary sentences. In the following, highlighted sentences can be removed without affecting the meaning.

  The volume of information has been rapidly increasing in the past few decades. While computer technology has played a significant role in encouraging the information growth, the latter has also had a great impact on the evolution of computer technology in processing data throughout the years. Historically, many diffferent kinds of databases have been developed to handle information, including the early hierarchical and network models, the relational model, as well as object-oriented and deductive databases. However, no matter how much these databases have improved, they still have their deficiencies. Much information is in textual format. This unstructured style of data, in contrast to the old structured record format data, cannot be managed properly by the traditional database models. Furthermore, since so much information is available, storage and indexing are not the only problems. We need to ensure that relevant information can be obtained upon querying the database.

- Revise frequently to delete superfluous words and simplify your sentences.
- Avoid showing off - information should be conveyed without unnecessary dressing.
- Do not condense too far - the writing can become hard to understand.
  - × Bit-stream interpretation requires external description of stored structures. Stored descriptions are encoded, not external.
  - √ Interpretation of bit-streams requires external information such as descriptions of stored structures. Such descriptions are themselves data, and if stored with the bit-stream become part of it, so that further external information is required.

Tone

- Scientific writing should be accurate and objective - avoid nuance, ambiguity, sensuality.
- A direct, simple style is most appropriate:
  - One idea per sentence or paragraph, one topic per section.
  - Simple and logical organization.
  - Short words.
  - Short sentences with simple structure.
  - Short paragraphs.
  - Avoid buzzwords, cliches, and slang.
  - Avoid excess, in length or style.
  - Omit any unnecessary material.
  - Be specific.

- Do not overqualify - this is a natural tendency of scientists but can be taken too far and make the meaning unclear.
  - × The results show that, for the given data, less memory is likely to be required by the new structure, depending on the magnitude of the numbers to be stored and the access pattern.
  - √ The results show that less memory was required by the new structure. Whether this result holds for other data sets will depend on the magnitude of the numbers and the access pattern, but we expect that the new structure will usually require less memory than the old.

- Technical writing is not a good outlet for artistic impulses.
  - × We have already seen, in our consideration of *what is*, that is the usual simplified assumptions lead inexorably to a representation that is desirable, because a solution is always desirable; but repugnant, because it is false. And we have presented *what should be*, assumptions whose nature is not susceptible to easy analysis but are the only tenable alternative to ignorance (absence of solution). Our choice is then Hobson's choice, to make do with what material we have – viable assumptions – and to discover whether the intractable can be teased into a useful form.
  - √ We have seen that the usual assumptions lead to a tractable model, but this model is only a poor representation of real behaviour. We therefore proposed better assumptions, which however are difficult to analyze. Now we consider whether there is any way in which our assumptions can be usefully applied.

- Should not imitate popular science writing.
  - × As each value is passed to the server, the "heart" of the system, it is checked to see whether it is in the appropriate range.
  - √ Each value passed to the central server is checked to see whether it is in the appropriate range.

Voice

- Avoid excessive use of indirect statements and passive voice.
  - × The following theorem can now be proved.
  - √ We can now prove the following theorem.
- Another common indirect style is to use verbs like “perform”, “utilized”, “achieved”, “carried out”, “conducted”, “done”, “occurred”, and “effected”.
  - × Tree structures can be utilized for dynamic storage of terms.
  - √ Terms can be stored in dynamic tree structures.
  - × Local packet transmission was performed to test error rates.
  - √ Error rates were tested by local packet transmission.
- Passive voice can change meaning or emphasis. If necessary, use it.

- Using “we” is valuable when distinguishing between your contributions and prior works.
  - × It is shown that stable graphs are closed.
  - √ We show that stable graphs are closed.
- By convention, “we” is commonly used in single authored works in place of “I”.

Ambiguity

- Check for ambiguity and rewrite your sentences.
  - × The compiler did not accept the program because it contained errors.
  - √ The program did not compile because it contained errors.
  - × In addition to skiplists we have tried trees. They are superior because they are slow in some circumstances but have lower asymptotic cost.
  - √ In addition to skiplists we have tried trees. Skiplists are superior because, although slow in some circumstances, they have lower asymptotic cost.
  - × The machine crashed and it was necessary to reboot it.
  - √ The machine crashed and a reboot was necessary.

Sentence

- Don't say too much in a single sentence.
  - × When the kernel process takes over, that is when in the default state, the time that is required for the kernel to deliver a message from a sending application process to another application process and to recompute the importance levels of these two application processes to determine which one has the higher priority is assumed to be randomly distributed with a constant service rate R.
  - √ When the kernel process takes over, one of its activities is to deliver a message from a sending application process to another application process, and to then recompute the importance levels of these two application processes to detemine which has the higher priority. The time required for this activity is assumed to be randomly distributed with a constant service rate R.
    - That the kernel process is the default state is irrelevant here.

- Avoid nested sentences.
  - × In the first stage, the backtracking tokenizer with a two-element retry buffer, errors, including illegal adjacencies as well as unrecognized tokens, are stored on an error stack for collation into a complete report.
  - √ The first stage is the backtracking tokenizer with a two-element retry buffer. In this stage possible errors include illegal adjacencies as well as unrecognized tokens; when detected, errors are stored on a stack for collation into a complete report.
- Do not fracture the conditional part of “if” expressions.
  - × If the machine is lightly loaded, then speed is acceptable whenever the data is on local disks.
  - √ If the machine is lightly loaded and data is on local disks, then speed is acceptable.
  - √ Speed is acceptable when the machine is lightly loaded and data is on local disks.

- Avoid double negatives - they are difficult to parse and ambiguous.
  - × There do not seem to be any reasons not to adopt the new approach.
  - √ The new approach is at least as good as the old and should be adopted.

Tense

- Most scientific writing is in present or past tense.
- Use the present tense for eternal truths.
- Past tense is used to describe the work and outcomes.
  - √ Although theory suggests that the Klein algorithm has asymptotic complexity $O(n^2)$, in our experiments the trend observed was $O(n)$.
- References can be discussed in present or past tense. Presence tense is preferable unless forced by context.
  - √ Willert (1999) shows that the space is open.
  - √ Haast (1986) postulated that the space is bounded, but Willert (1999) has since shown that it is open.
- Future tense is rarely used in scientific writing other than in the conclusions.

Definitions

- Terminology, variables, abbreviations, and acronyms should be defined or explained the first time they are used.
- Emphasis on the first occurrence of new terminology is helpful.
  - × We use homogeneous sets to represent these events.
  - √ We use *homogeneous* sets to represent these events.
  - √ To represent these events we use homogeneous sets, whose members are all of the same type.

Qualifiers

- Don't put more than one qualifier in a sentence. Otherwise the text will look timid.
  - × It is perhaps possible that the algorithm might fail on unusual inputs.
  - √ The algorithm might fail on unusual input.
  - √ It is possible that the algorithm would fail on unusual input.
  - × We are planning to consider possible options for extending our result.
  - √ We are considering how to extend our results.
- Writing is more forceful without qualifiers such as "very", "quite" and "simply".
  - × There is very little advantage to the networked approach.
  - √ There is little advantage to the networked approach.

Misused words

Which, that: Use "that" in preference to "which": use "which" only when it cannot be replaced by "that"

- × There is one method which is acceptable.
- √ There is one method that is acceptable.
- √ There are three options, of which only one is is tractable.

Absence of the word "that" can make sentence unclear.

- × It is true the result is hard to generalize.
- √ It is true that the result is hard to generalize.

Less, fewer: Use "less" for continuous quantities and "fewer" for discrete quantities. e.g. less space and fewer errors.

Affect, effect: The "effect" is the *consequence* of an action. An action "affects" or *influences* the outcome.
Alternate, alternative: The word "alternate" means *other*. The word "alternative" means something that can be chosen.
Assume, presume: "Assume" means *take as being true*. "Presume" means *take for granted*.
May, can: "May" indicates choice. "Can" indicates capability.
Will, shall: The word "shall" can seem archiac and is rarely preferable to "will". Both are often used unnecessarily and can be deleted in many cases.
Conversely, inversely, similarly, likewise: Only use "conversely" if the statement that follows really is the opposite of the preceding material. Don't use "similarly" or "likewise" unless whatever follows has a strong parallel to the preceding material. Some authors use "inversely", but the meaning is rarely clear; avoid it.

Padding

- Padding is the use of unnecessary phrases such as “the fact that” or “in general”.
- Adjectives are another form of padding.
  - × A well-known method such as the venerable quicksort is a potential alternative in instances of this kind.
  - √ A method such as quicksort is a potential alternative.
- Some examples:
  adding together → adding,
  after the end of → after,
  in the region of → approximately,
  cancel out → cancel,
  let us now consider → consider,
  cooperate together → cooperate,
  divided up → divided,
  give a description of → describe,
  during the course of → during,

totally eliminated → eliminated,
first of all → first,
in view of the fact → given,
joined up → joined,
merged together → merged,
in the vast majority of → most,
separate into partitions → partition,
completely unique → unique,
whether or not → whether.