Doing Resarch in Natural Language Processing

Session 3: Scientific Writing: Precision and Clarity

Frank Keller

27 September 2023

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Overview

Balancing Precision with Clarity

Avoiding Needless Complexity

Measuring Complexity

Illustrations and Captions

Reading: Alley (2018), Chapter 2.

Please also look at Alley's web site, which has a lot of videos and additional materials: https://www.craftofscientificwriting.com/

Balancing Precision with Clarity

Choose the Right Word

- Use the right technical terms: you wouldn't say *weight* when you mean *mass*.
- But everyday words have a precise meaning too (don't use fancy words like *plethora* unless you're sure what they mean).
- Take care with easily confused words such as *continuously* and *continually*.
- Alley has a whole list of them, Appendix D.

Synonyms

In creative writing, the use of synonyms is encouraged (they keep your prose interesting).

In scientific writing, synonyms are mostly not a good thing.

For example, *development set* and *validation set* are synonyms, but stick to one to avoid confusion. (The reader may wonder whether you are using two different sets to tune your model.)

Also, there's many *near*-synonyms, which can also cause confusion. For example *image descriptions* and *image captions* are closely related, but not exactly the same.

Don't hesitate to repeat a word if it's the right word!

Connotations, Exaggerations

Avoid words with negative connotations, e.g., *cheap*, *obvious*.

Avoid exaggerations, e.g., *countless activities*, *a thorough literature search*.

Be careful with words such as *prove* and *optimal*, which have precise meaning in most scientific fields.

Avoiding Needless Complexity

Needless Complexity

Avoiding needless complexity is the most important advice to scientific writers (according to Alley). Avoid needlessly complex:

- paragraphs
- words
- phrases
- sentences

Consider the following paragraph, written by Niels Bohr.

Complex Paragraphs

The Correspondence Principle. So far as the principles of the quantum theory are concerned, the point which has been emphasized hitherto is the radical departure from our usual conceptions of mechanical and electrodynamical phenomena. As I have attempted to show in recent years, it appears possible, however, to adopt a point of view which suggests that the quantum theory may, nevertheless, be regarded as a rational generalization of ordinary conceptions. As may be seen from the postulates of the quantum theory, and particularly the frequency relation, a direct connection between the spectra and the motion of the kind required by the classical dynamics is excluded but at the same time, the form of these postulates leads us to another relation of a remarkable nature.

Complex Paragraphs

The Correspondence Principle. So far as the principles of the quantum theory are concerned, the point which has been emphasized hitherto is the radical departure from our usual conceptions of mechanical and electrodynamical phenomena. As I have attempted to show in recent years, it appears possible, however, to adopt a point of view which suggests that the quantum theory may, nevertheless, be regarded as a rational generalization of ordinary conceptions. As may be seen from the postulates of the quantum theory, and particularly the frequency relation, a direct connection between the spectra and the motion of the kind required by the classical dynamics is excluded but at the same time, the form of these postulates leads us to another relation of a remarkable nature.

Complex words

Complex Paragraphs

The Correspondence Principle. So far as the principles of the quantum theory are concerned, the point which has been emphasized hitherto is the radical departure from our usual conceptions of mechanical and electrodynamical phenomena. As I have attempted to show in recent years, it appears possible, however, to adopt a point of view which suggests that the quantum theory may, nevertheless, be regarded as a rational generalization of ordinary conceptions. As may be seen from the postulates of the quantum theory, and particularly the frequency relation, a direct connection between the spectra and the motion of the kind required by the classical dynamics is excluded but at the same time, the form of these postulates leads us to another relation of a remarkable nature.

Complex words

Complex sentences: on average 40 words per sentence

Complex Paragraphs

The Correspondence Principle. Many people have stated that the quantum theory is a radical departure from classical mechanics and electrodynamics. However, the quantum theory may be regarded as nothing more than a rational extension of classical concepts. Although no direct connection exists between quantum theory and classical dynamics, the form of the quantum theory's postulates, particularly the frequency relation, leads us to another kind of relation, one that is remarkable.

This revised version is shorter and less complex.

Complex Words

Avoid words that are long and infrequent, but don't add precision and clarity, e.g.:

- *elucidate:* use *show*, *reveal* instead
- many *-ize* words: *prioritize* or *utilize*; use *rank* and *use* instead
- some *-ize* words have precise meaning: *minimize* or *maximize*

Individual word substitutions may not make a difference, but overall, the effect can be substantial.

Complex Words

The objective of this study is to develop an effective commercialization strategy for solar energy systems by analyzing the factors that are impeding early commercial projects and by prioritizing the potential government and industry actions that can facilitate the viability of the projects.

This study will consider why current solar energy systems have not yet reached the commercial stage and will evaluate the steps that government and industry can take to make these systems commercial.

Complex Words

Other sources of complexity that should be used sparingly or avoided:

- abbreviations: use as sparingly as possible
- all caps for names: avoid if possible
- slashed terms: replace by a single, better term, or a conjunction

Complex Phrases

Also phrases can become unwieldy, e.g., this compound noun:

Solar One is a 10-megawatt solar thermal electric central receiver Barstow power pilot plant.

Best to break this up into multiple phrases, each with a one or two modifiers, or even into a several sentences:

Solar One is a solar-powered pilot plant located near Barstow, California. This plant produces 10 megawatts of electric power by capturing solar energy in a central receiver design.

Complex Sentences

The object of the work was to confirm the nature of electrical breakdown of nitrogen in uniform fields at relatively high pressures and interelectrode gaps that approach those obtained in engineering practice, prior to the determination of the processes that set the criterion for breakdown in the above-mentioned gases and mixtures in uniform and non-uniform fields of engineering significance.

This sentence is long (61 words) and it tries to communicate multiple ideas at once.

Instead, try to use short sentences (in the teens). And express *one idea per sentence.*

Complex Sentences

Also, the sentence contains 11 prepositional phrases. It just strings one PP after the next, making it hard to for the reader to figures out when the sentence will end.

Rewritten version with only 3 PPs:

At relatively high pressures and typical electrode gap distances, the electrical breakdown of nitrogen was studied in uniform fields.

Another version that uses two sentences:

This study tests the electrical breakdown of nitrogen in uniform fields. For these tests, the electrode gap distances were typical, while the pressures were relatively high.

This version emphasizes what was novel in this study (in sentence 2).

Over to You

Exercise 1

Let's look at some text from the intro of Vaswani et al. (2017) (next page):

- Does it use complex words, phrases, sentences?
- What about synonyms, exaggerations, abbreviations, needlessly complex verbs?
- What is the overall balance of precision and clarity?

Would you re-write these paragraphs? How?

Exercise 1

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [35, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [38, 24, 15].

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples.

Measuring Complexity

Measuring Complexity

There are ways to quantify the complexity of a text. A number of *reading indices* have been proposed, which predict the *reading level* of a text.

A typical example is the *Gunning Fog Index:*

$$F_i = 0.4\left(\frac{N_w}{N_s} + P_{lw}(100)\right)$$

where N_w is the number of words per paragraph, N_s is the number of sentences per paragraph, and P_{lw} is the percentage of long word (3 or more syllables).

Measuring Complexity

The *reading level* indicates how many years of reading experience is needed to understand a text, ranging from 6 to 12 (high school) to 17 (college graduate).

Example for reading levels:

- newspapers: $F_i = 10$
- *Scientific American:* $F_i = 12$
- Einstein's *Special Theory of Relativity:* $F_i = 12$
- Niels Bohr's paragraph on p8: $F_i = 24$

Does that mean a 12th grader would understand *Special Theory of Relativity?* No, it just means they'd be comfortable with the lengths of the words and sentences.

In your own writing, aim for reading levels between 10 and 13.

Illustrations and Captions

Illustrations

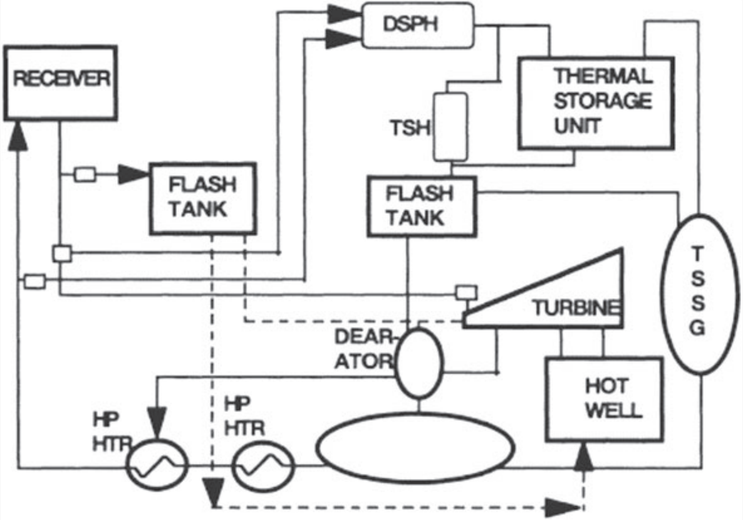

Figure 2-1. Thermal storage system.

Illustrations

Balance precision and clarity also in your illustrations:

- Don't use figures that are more complex than the text used to explain them.
- Use figures to illustrate the most important aspects of what you want to explain, leave out unnecessary details.
- When a figure provides information that's not in the text, it needs to be explained (or be self-explanatory).

Illustrations

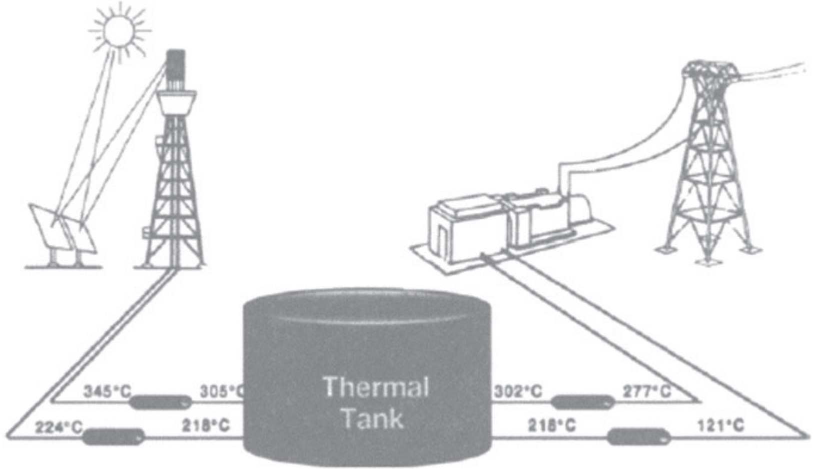

Figure 2-2. Thermal storage system. This storage system takes excess energy from the solar receiver and stores it for later use when the sun is no longer providing solar radiation to the mirrors.

Captions

Every figure needs a caption:

- Reader are automatically drawn to figures, and will try to understand them, often before reading the main text.
- The caption needs to contain everything that's required to understand the figure.
- Start with a phrase that identifies the illustration; formulate it using the same consideration as for document titles.
- Then explain what the figure shows in more detail, expand any abbreviations, label all the parts, etc.

Over to You

Exercise 2

Let's look at an illustration from Vaswani et al. (2017) (next page):

- Does the figure balance clarity and precision?
- Does the caption contain a meaningful title?
- Are figure and caption taken together self-contained?

How would you modify the figure and the caption to improve it?

Exercise 2

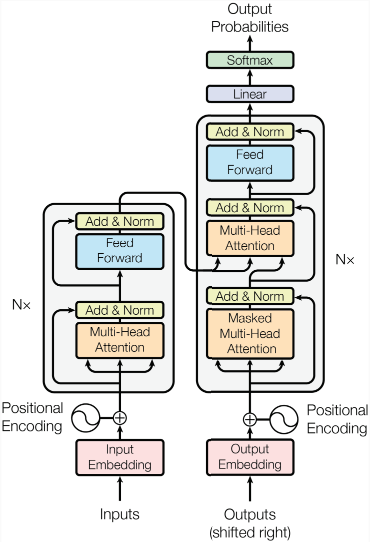

Figure 1: The Transformer - model architecture.

References

Alley, Michael. 2018. *The Craft of Scientific Writing*. Springer, New York, NY, 4 edition.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Red Hook, NY, pages 5998–6008.