Specialist English: Assignment 4 (solutions)

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

November 12, 2018

Here's my solutions; your solutions needn't be identical. To begin with, here's the snippet from the assignment:

Problem 1 The title is

> Finding effective classifier for malicious URL detection

The countable noun phrase "effective classifier" is singular and thus missing an article (in this case "an"). The word "finding" is not needed. Thus, we change it to:

> An effective classifier for malicious URL detection.

Another possibility is:

> Effective classifiers for malicious URL detection.

Problem 2 The sentence is:

> Besides, we adjust the number of iterations of random forest and random choice characteristic number of random forest in experiment.

Suitable replacements for "Besides, ..." (which are not likely to be confused with "Anyway, ...") are "Moreover, ...", "Further, ...", or "Furthermore, ...". I discourage "What's more, ..." as it's informal.

The part of the sentence in blue doesn't make sense to me—I think it's safe to say it's nonsensical. Realistically, the sentence needs completely rewriting.

Problem 3 The sentence is:

> Beyond blacklisting heuristics, domestic and foreign scholars have conducted a lot of research work on malicious URL detection.

The phrase "a lot of" (while correct) is something a child would say. More appropriate for a research publication is "much" (in this context, "much research"). Alternatives are "a large amount of research" and "substantial research", among others.

Some students said "variety" but this is equivalent to "many kinds of", which has a different meaning. Some students said "massive" but this implies large in quantity, not in number (e.g., "this mountain is massive" does not imply there are many mountains); "massive" is also better suited to describe physical quantities. Some students said "plenty", but this is not suitable—"plenty" implies "enough", and if there's enough research already, why are we writing a paper on the topic?

Some students said "numerous", but "research" is uncountable (i.e., we can't have "one research", "two researches", etc.), so we can't have "numerous research". (Last year, students

went to the dictionary and found it says "research" is countable—this was true maybe 100 years ago, but it is no longer true. Treating "research" as countable is almost universally considered a grammar error.)

> In addition to blacklisting heuristics, much research has been conducted on malicious URL detection.

This also gets rid of other problems:

- "Beyond ..." is difficult to parse in this context, "In addition to ..." is clearer, and
- "domestic and foreign scholars" is mismatched with the authors publishing in an international conference; everyone is "foreign" to somewhere.

Problem 4 The snippet is:

> accuracy. Different from [1], Ma, J et al.[2] analyze the vocabulary of suspicious URL (Lexical Features) and host
>
> ⋮
>
> [2] Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious urls. In: IV, J.F.E., Fogelman-Souli_e, F., Flach, P.A., Zaki, M.J. (eds.) *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1245-1254. ACM (2009)

1. Suitable alternatives for "Different from [1], ..." (which feels incorrect) are "Differing from [1], ..." and "Unlike [1], ...". Alternatively, we can just delete "Different from [1]".
2. We correct "Ma, J et al.[2]" to

   > Ma et al. [2]

   I don't believe there is an alternative way of writing this.
3. Retaining the same reference format, we fix reference [2] to

   > Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proc. SIGKDD*. pp. 1245–1254. (2009)

   or alternatively

   > Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proc. Conference on Knowledge Discovery and Data Mining*. pp. 1245–1254. (2009)

Please note that **online references (including BibTeX references) are full of errors**. Many students lost marks as a result of using online references with errors.

Problem 5 The snippet is:

> malicious URL at a much higher cost and lower efficiency. In this paper, a feature extraction scheme is proposed, which combines the character frequency and structural features with small acquisition cost, and performs the model contrast experiment and parameter optimization experiment.

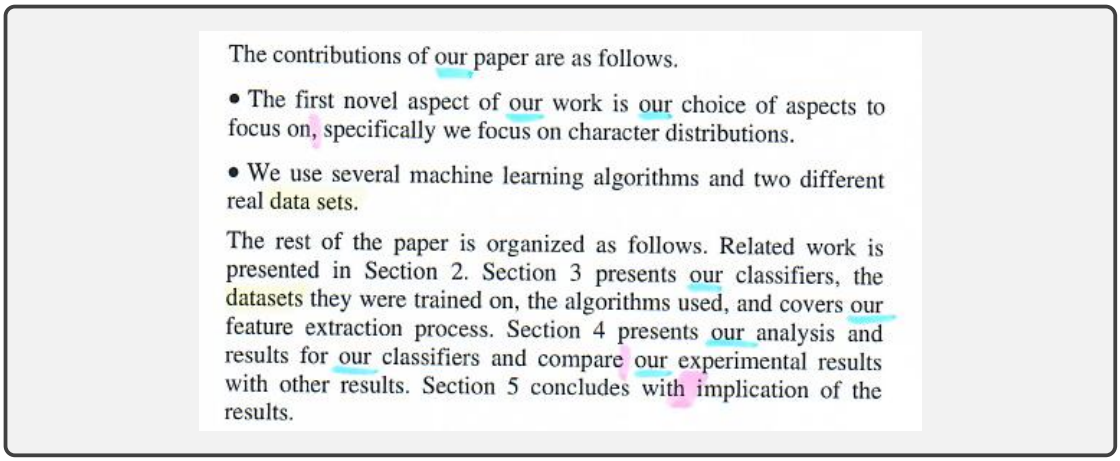
The phrase "the character frequency and structural features with small acquisition cost" is difficult to parse since the reader needs to distinguish between:

1. "the character frequency" and "structural features with small acquisition cost" (which would be the natural interpretation), which implies only the structural features which have a small acquisition cost are included (and those with a high acquisition cost are excluded); and
2. "the character frequency and structural features" with "small acquisition cost", which implies all structural features are included, and that it results in a small acquisition cost.

I'm not sure which is the correct interpretation.

Several students mentioned the vague word "small", but vagueness (i.e., being imprecise) is different to ambiguity (i.e., having multiple logical interpretations).

Problem 6 The snippet is:



The contributions of our paper are as follows.

- The first novel aspect of our work is our choice of aspects to focus on, specifically we focus on character distributions.
- We use several machine learning algorithms and two different real data sets.

The rest of the paper is organized as follows. Related work is presented in Section 2. Section 3 presents our classifiers, the datasets they were trained on, the algorithms used, and covers our feature extraction process. Section 4 presents our analysis and results for our classifiers and compare our experimental results with other results. Section 5 concludes with implication of the results.

1. In the first bulleted item, there are two separate clauses: one ending "... choice of aspects to focus on" followed by the one beginning "specifically ...". Moreover, there's no words used to adjoin them (e.g., "so" or "and"). The most appropriate punctuation mark is a semi-colon ";", since the context of the first clause is used in the second clause. Also possible is a colon ":", although it would be better to omit "specifically" in this case.

   Students also suggested "—", and starting a new sentence instead. Neither of these are great, but they're correct (as far as I can tell). Having two short sentences, as proposed, is undesirable.

2. We avoid using "our" in all of the cases in the snippet as follows:

   - we change "our paper" to "this paper" (or just delete "of our paper", since it's clear from context),
   - we change "our work" to "this work" (some students wrote "the proposed work", but the work is not proposed, the work is actually done, and so this is inaccurate),
   - we change "our choice" to "the choice",
   - we change "our classifies" to e.g. "the proposed classifiers" or "the chosen classifiers",
   - we change "our feature extraction process" to e.g. "the proposed feature extraction process",
   - we change "our analysis and results" to "an analysis and the results",
   - we change "our classifies" as above, and
   - we change "our experimental results" to "the experimental results".

Problem 7 There are many problems; some that were mentioned repeatedly in student assignments are:

- The department names don't need repeating for each member of the same department.
- The first words in the abstract "Malicious URL" is a singular countable noun phrase without an article. (It should be "Malicious URLs".)
- The abbreviation "char" (instead of "character") in the keywords is unsuitable.
- We don't write "although ..., but ..." in English.
- The adverb is misplaced in "the users inadvertently will be caught" and should be "the users will inadvertently be caught" or "the users will be inadvertently caught" or "the users will be caught inadvertently".
- It is peculiar to start a paragraph with "Then, let's look at some other examples" which is a sentence fragment, and changes the tone of the writing: the author has changed to a narrator (suitable for novels, but not scientific research). The word "then" is totally irrelevant. This short sentence is better merged into the paragraph naturally, e.g., "Other examples include Garera et al. [5], who ...".
- Writing both "data sets" (with a space) and "datasets" (without a space) is inconsistent. Both are acceptable, as long as they're consistently used.