Instruction for Camera-Ready Submission

May 14, 2020

These instructions are adapted from the ones for ACL 2020, EMNLP 2020, and NAACL 2021.

To submit your camera-ready version, please go to the softconf site, click the submission id, and then click "Final Camera Ready Submission". This document is organized based on the sections in the Final Submission Page. Please note that the following applies to the ACL-IJCNLP main conference papers and Findings of ACL papers. It will also apply to ACL-IJCNLP workshop papers unless the workshops have their own instructions.

Please remember to use the template files for the conference and follow the instructions in this document. In addition, you need to run a tool called aclpub_check developed by the NAACL2021 publication chairs to automate the checking for some frequently encountered formatting issues. For details of the tool, see Section 3 and the Appendix. Camera-ready copies that fail to meet these instructions will be returned for re-submission.

1. Metadata (Part A in the final submission form)

Can I make changes to the author list in the camera-ready version?

No, you cannot change the author list, as explained in the FAQ at
https://2021.aclweb.org/blog/faq-decision/

In fact, the authors' field will not be editable in the final submission form.

How should I enter metadata on the START system?

The metadata (title, author, abstract) that you enter into START is very important, because it is used on the conference website, proceedings, handbook, mobile app, and the ACL Anthology (and propagates to Google Scholar, etc.).

Before the metadata is entered, please have all authors ensure that the name in their START profile (User Console, Update Profile) appears exactly the way that they want it to appear.

- Unicode (UTF-8) can be used for accented or special characters.

- Ordinarily, names are **NOT** written in all caps or all lowercase.

- The "Last Name" is the name(s) by which your paper is to be cited. It is usually a family name, even for authors from cultures where the family name is written first.

- The "First Name" is usually a given name or names, including middle names/initials.

The metadata should be written using Unicode (UTF-8) with LaTeX commands. Please try to follow these guidelines:

- In titles, please capitalize the first word, the first word after a colon (:), and all other words (including hyphenated words like Mixed-Case), except the function words: articles, pronouns, conjunctions, prepositions, particles, subordinating conjunctions, and the infinitive marker "to".

- BibTeX (in many bibliography styles, including ACL's) lowercases the titles of conference papers, and needs to be told which letters not to lowercase. So if your title has letters that should always be capitals, please protect them with curly braces, like this: *{E}nglish, {C}homsky, {IBM}, {CFG}s, {HMM}s*. Please also protect the first letter after a sentence-final punctuation mark. For example:

  Can {LSTM} Learn to Capture Agreement? {T}he Case of {B}asque

  Named Entity Extraction from Noisy Input: {S}peech and {OCR}.

  It is important to only protect those first letters that belong to the categories exhibited above, i.e., proper names (including language names), acronyms, abbreviations, the first letter after punctuation (such as a colon) and the first letter in a sentence. These curly braces will *not* appear in the online conference program or proceedings. They will only appear in the BibTeX file that others will use to cite your paper.

- If you need literal curly braces, please escape them like this: \{ \}

- Please don't use any nonstandard LaTeX commands, and there should be no \footnote or citations using \cite or related commands.

- You can use LaTeX math mode where appropriate: An $O(n^2)$ Algorithm for n-gram Smoothing.

- You can use Unicode (UTF-8) for accented or special characters.

- If you copy-and-paste from your PDF file, please be sure to rejoin words broken by hyphenation.

2. Copyright (Part B in the form)

What about copyright?

When you submit the paper, you will be asked to sign the ACL Copyright Transfer Agreement on behalf of all authors, electronically (via the START Conference Manager). Authors retain many rights under this agreement and it is appropriate in the vast majority of cases. Please contact

the publication chairs at acl2021publicationchairs@googlegroups.com, with any concerns regarding copyright.

Who should sign the copyright form?

Before signing this form, please confirm with your co-authors (and, if applicable, your and their employers) that they authorize you to sign on their behalf. Only the authorized representative needs to sign the copyright form. Please sign your full name (not just your first or last initials).

3. Main paper (Part C in the form)

What are the page limits?

For both long and short papers, ACL-IJCNLP 2021 allows one extra page to help address reviewer comments. Please use the extra space to help address reviewer comments. Long papers are permitted at most **9 pages** of text and short papers may use up to **5 pages** of text, plus unlimited space for references and the impact statement. Acknowledgements count toward the page limit.

What is the max size of the PDF file?

The max size of the PDF file (including the Appendices, see Section 4) should be no more than 20MB.

What is the format for the camera-ready copy?

The file must be in Portable Document Format (PDF) on A4 paper. We require the use of ACL LaTeX style files or Microsoft Word style files tailored for this year's ACL conference. You can access the style files and detailed formatting instructions here: https://2021.aclweb.org/downloads/acl-ijcnlp2021-templates.zip or https://www.overleaf.com/latex/templates/instructions-for-acl-ijcnlp-2021-proceedings/mhxffkjdwymb.

If you are using LaTeX, please create the PDF file with pdflatex or xelatex. This ensures use of the proper Type 1 fonts and also takes advantage of other PDF features. You will have the best results using a modern LaTeX distribution, in particular, TeX live. Using the geometry package to set the A4 format is recommended.

How do I check the format of the camera-ready version before submitting?

You should check the paper format according to this instruction page and use the style template files with the links above. The NAACL2021 publication chairs have developed a package called aclpub_check to automate much of the format checking (See the Appendix for

details). Please run this package before submitting your paper to softconf in order not to miss some subtle formatting details.

How should the final copy differ from the original submission?

The camera-ready version of your paper should incorporate the comments of the reviewers as well as other changes you see fit to make. In addition, be sure to do all of the following:

- Ensure that your paper conforms to the [provided styles, font and page size](#).

- Include the authors' names and affiliations under the title. Note that the list of authors should be identical to the list specified when submitting the paper and should be in the same order.

- De-anonymize references to your own work in the body of the paper.

- Where appropriate, add acknowledgments for colleagues, reviewers, and grants. Do not number the Acknowledgements section. Please note that the acknowledgement section should fit within the allowed page limits (9 pages for long and 5 pages for short papers, plus unlimited pages for the impact statement and references) and be in the same font as the rest of the paper.

- Ensure that all tables, graphs, and figures are readable at standard resolutions.

- The Appendix (if exists) should appear after the references, as part of the same PDF file (see Section 4); in contrast, the supplemental materials should appear as a separate zip file (see Section 5).

What are the tips to make my final version more accessible?

As a central venue of publication for our community, please prioritise the accessibility of your final version. The Diversity & Inclusion committee for ACL 2020 has outlined some tips on how to do this: https://acl2020.org/blog/accessibility-for-camera-ready/

How do I ensure that my file is correctly formatted?

You need to run the aclpub_check package (see the Appendix), and pay special attention to things like the following:

- **The paper size:** Your paper needs to be formatted to A4. Here are a couple of ways to check this:
  - Using pdfinfo. The pdfinfo command should include "Page size: 595.276 x 841.89 pts" in its output.
  - Using Apple's Preview.app. Open the PDF, and type Ctrl-I. It should report the correct page size.
  - Using Adobe Acrobat. Open the PDF, navigate to File, Properties…, Description. The field labeled "Page Size" should read 8.27 × 11.69 inches in.

- **Embedding fonts:** You can check your final PDF with the command "pdffonts mypaper.pdf" and confirm that all the fonts say "yes" under "emb". START will not let you upload your final PDF otherwise. If you are including graphics with the PDF extension, these files must also have embedded fonts. If your paper uses Asian fonts, they must be embedded in the PDF file so that they can be displayed by non-Asian versions of the PDF reader (Asian versions ship with a larger set of default fonts.)
- **Long titles:** The title should **NOT** exceed the left & right margins. Fold a long title into multiple lines if necessary.
- **Margins:** The margins of the text area must be reserved as specified in the style file.
- **Page numbers:** DO **NOT** provide page numbers for your PDF file. The page numbers will be generated automatically while compiling the proceedings.
- **Consistent Author Names:** The author names must be consistent everywhere and be the same as registered to the submission page(s), and have the same spelling style in all accepted papers. This can avoid generating non-unique Author Index entries for authors with more than one accepted paper. Be sure to ask your co-authors whether you enter their names correctly.

What if my paper includes graphics?

Remember that you are providing a camera-ready copy. Thus, artwork and photos should be included directly in the paper in their final positions. Ideally, you should use vector graphic formats (PDF, EPS), which allow the graphics to scale arbitrarily. Avoid GIF or JPEG images that are low resolution or highly compressed.

Your paper must look good both when printed (A4 size) and when viewed on screen as PDF (zoomable to any size, color okay). Thus, you may want to use color high-resolution graphics, allowing readers to zoom in on a graph and study it. However, please check that the same graph or photograph is legible when printed and in a PDF viewer at different resolutions. Don't go overboard on resolution; keep file sizes manageable. Note that vector graphics (e.g., encapsulated PostScript) look good at any scale and take up little space (unless you are plotting many thousands of data points).

What if my paper's title or abstract has changed?

In addition to edit the title/abstract in the main paper, please also remember to edit those metadata fields when you upload the camera-ready version, so that they will appear correctly in the table of contents, author index, conference schedule, etc. **Please also note that your name will appear in conference metadata as you have configured it in START**, so make sure that it is correct there (e.g., capitalization, full name, etc.). You can change this on the user settings page of the START conference manager, under "User" -> "Account Information" -> "Update Profile".

4. The Appendices (As the last part of the pdf file for Part C in the form)

Appendices are for things such as lemmas, hyperparameters, formulas, proofs, and tables that are informative but not critical to the understanding of the paper.

Where do the Appendices go?

Appendices should be part of the PDF file uploaded to Part C in the form. It should appear at the end of the PDF file, after the references.

What's the page limit of the Appendices?

The Appendices do not count towards the page limit of the main paper, but they need to be no more than 4 pages, as recommended by the publication chairs.

What template should the Appendices use?

The Appendices should use exactly the same template as the main paper (See Section 3).

5. Additional document for conditionally accepted papers (Part D in the form)

Who needs to submit a document for Part D?

That depends on the status of your paper:

- If the status is Accept-P1-condition or Accept-P2-condition, you must submit an additional document (explained below).
- If the status is Accept-P1 or Accept-P2, please do NOT upload any documents to Part D even if your paper receives an EAC review and/or an EAC metareview. You can skip this section.

What should the document look like?

In this document, authors need to provide a short document that explains how they have made the changes requested by the EAC meta-reviews (which appears at the bottom of all other reviews and has a special heading). The document can be either a plain text file or a PDF file, and should list each of the concerns from the original EAC meta-review along with a short summary of what was done to remedy the concern.

To make this as straightforward as possible, we recommend copying and pasting the contents of the EAC meta-review into the document and then describing the action taken after each item

in the EAC metareview. If there are additional issues mentioned in EAC reviews that were not included in the EAC metareview, authors may choose to address those points as well; however, only the points mentioned in the EAC meta-review will be the basis of whether the final version of the paper will be accepted. This explanatory document will not be made public; it is just to expedite the conditional acceptance workflow for the EAC's review process.

The size of this document should be no more than 10MB.

What will happen after June 1?

The EAC chairs will go over this document and check the camera-ready version to determine whether the required changes have been included in the camera-ready version. If so, the condition will be removed, the paper will be accepted to the main conference or Findings, and the status of having been conditionally accepted will not be publicly visible. If not, the paper will be rejected. The PCs will inform the authors of the final EAC decisions by June 7, 2021.

If you have any questions about EAC reviews or the changes required by EAC, you can contact the ACL-IJCNLP 2021 Ethics Chairs <acl-2021-ethics-chairs@googlegroups.com>. You can also check the Ethics FAQ at https://2021.aclweb.org/ethics/Ethics-FAQ/.

6. Supplementary material (Part E in the form)

Where do supplemental materials go?

Supplemental material contains material such as data, software, all other material and is uploaded separately as a zip file. The maximum size of the zip file is 20MB.

Do I need to submit latex source files?

No. Latex files will not be used in ACL Anthology, thus please do NOT submit them.

7. Deadlines

When and where do I send my final camera-ready paper?

You must submit the final version of your paper by **June 1, 2021** (11:59pm, UTC-12 hours, "anywhere on Earth") by navigating to the ACL 2021 START login page and following the internal links. This is a firm deadline, i.e., no change will be accepted after that.

8. Questions or Comments

If you have any questions about:

- the camera-ready version including running the aclpub_check package: please contact our publication chairs at acl2021publicationchairs@googlegroups.com.
- the EAC issues and additional document in Part D: please contact the EAC chairs at acl-2021-ethics-chairs@googlegroups.com

Please do **NOT** contact ACL-IJCNLP PCs unless your questions cannot be answered by the publication chairs or the EAC committee.

Appendix: Checking the format with the aclpub_check package

This appendix is modified from the document of the NAACL 2021 publication chairs at:
https://docs.google.com/document/d/1MYKGS8x17Mvl4fKYz_ZHIkre9KBhzB_9W87bu4rbtqI/edit

Following NAACL 2021, we are using a package called "aclpub_check" that automatically detects author formatting errors, margin violations as well as many other common formatting errors. Before submitting your paper to softconf, please run the package first and fix the detected errors.

A.1: Access the package

The package is written in Python and there are four steps to using it:

1: `git clone https://github.com/yz-joey/ACLPUB.git`
2: `cd ACLPUB`
3: `pip install -e .`
4: `python3 aclpub_check/formatchecker.py --paper_type PAPER_TYPE PAPER_NAME.pdf`

You should also be able to use it via:

1: `pip3 install git+https://github.com/yz-joey/ACLPUB.git`
2: `python3 -m aclpub_check.formatchecker --paper_type PAPER_TYPE PAPER_NAME.pdf`

PAPER_TYPE should be the word *short* or *long.* PAPER_NAME can be any valid file name.

Below, we are walking you through how to use the tool by considering three papers published by publication chairs at NAACL 2021.

A.2. Example 1

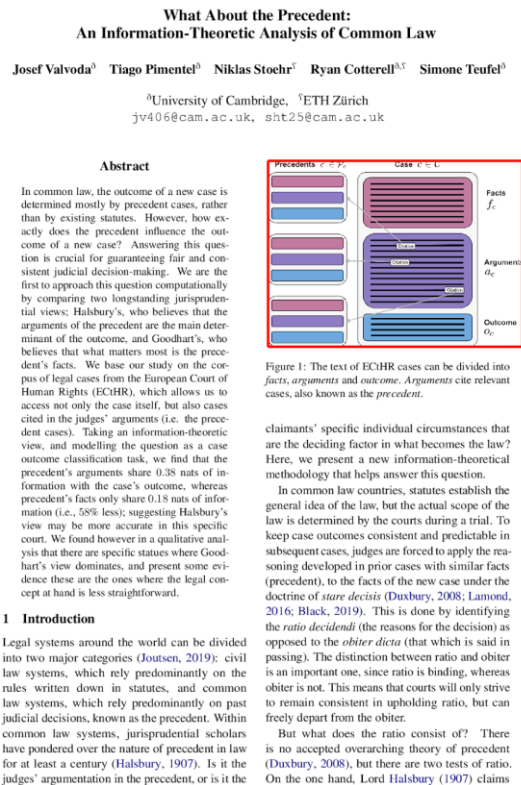First, consider Josef Valvoda's paper "What About the Precedent: An Information-Theoretic Analysis of Common Law"

```
> python3 aclpub_check/formatchecker.py --paper_type long precedent.pdf

Checking precedent.pdf
Error (Margin): An image on page 1 bleeds into the margin.
```

```
We detected 1 error and 0 warnings in your paper.
```

Thus, we see this paper has a margin violation. The script output a png called "errors-precedent-page-1.png" which you can see below:

What About the Precedent:
An Information-Theoretic Analysis of Common Law

Josef Valvoda[◊]   Tiago Pimentel[◊]   Niklas Stoehr[ς]   Ryan Cotterell[◊,ς]   Simone Teufel[◊]

[◊]University of Cambridge, [ς]ETH Zürich
jv406@cam.ac.uk, sht25@cam.ac.uk

It shows that Josef has to shrink the first-page picture to make the paper compliant with the NAACL formatting rules. This is easily remedied with the adjustbox https://www.ctan.org/pkg/adjustbox package or a similar bit of LaTeX.

A.3: Example 2

Next, consider Tiago Pimentel and Irene Nikkarinen's paper: "How (Non-)Optimal is the Lexicon?"

```
> python3 aclpub_check/formatchecker.py --paper_type long lexicon.pdf

Checking lexicon.pdf
```

Error (Margin): Text on page 13 bleeds into the margin.

This time, the automatically generated png shows me that the authors didn't keep some math equations out of the margin in a proof in the appendix.

	Train	Validation	Test
English	242,030	66,668	66,243
Finnish	466,745	109,232	110,378
Hebrew	311,860	104,555	104,478
Indonesian	243,118	69,792	70,079
Tamil	479,668	116,196	115,422
Turkish	308,419	84,300	83,871
Yoruba	47,740	12,877	12,877

Table 3: The number of word types used in training, validation and testing.

B Model training

As mentioned in the main text, we cannot directly infer the parameters of our model and we use a solution similar to expectation maximization (Wei and Tanner, 1990). We freeze our LSTM generator while learning the PYCRP parameters and then freeze the PYCRP to train the LSTM model.

Expectation step. This step uses a Gibbs sampling procedure to estimate the parameters of the PYCRP. For each token in our dataset, we fix all cluster assignments $\mathbf{z}_{-n}$ except for the given token's one z_n. We then re-sample this token's cluster based on the marginal probability $p(z_n | \mathbf{z}_{-n}, \boldsymbol{\ell}, \boldsymbol{w}_n)$. We do this for 5 epochs, and use the assignments which result in the best development set cross-entropy. This process can both remove clusters and create new ones by replacing tokens. The set of populated clusters (together with their wordform labels) then allows creating a new wordform dataset of size K', where the distribution of the token frequencies is expected to be less skewed. In practice this can be done by using the resulting set of cluster labels $\{\boldsymbol{\ell}_k\}_{k=1}^{K'}$, i.e. a word will appear in this new dataset as many times as it was assigned as a cluster label.

Maximization step. We use the set of populated cluster labels to train the generator LSTM—assuming that this allows learning a more representative model of a language's graphotactics as the irregular common words are less dominant in its training set. In other words, at each epoch, the generator will be trained in a wordform as many times as it has been assigned as a cluster label.

Hyperparameters and implementation details. For the PYCRP, we fix hyper-parameters $a = 0.5$ and $b = 10,000$, and we use the optimized Gibbs

sampling algorithm designed by Blunsom et al. (2009). As our generator, we use a three layers LSTM with an embedding size of 128, a hidden size of 512 and dropout of .33. This LSTM is trained using AdamW (Loshchilov and Hutter, 2019) and we hotstart it by initially training on the set of word types in the training set (the set of unique wordforms in it).

C Proof of Bounded Entropy for Proposition 2

In this section, we present the proof of Proposition 2. This proposition is repeated here for convenience:

Proposition 2. *If a language model $p(\boldsymbol{w})$ is ε-smooth, then its entropy is finite, i.e. $\mathrm{H}(W) < \infty$.*

Proof. To prove this, we first break the entropy of a string in two parts. The entropy of the first character, plus the entropy of the following ones given the first.

$$\mathrm{H}(W) = \mathrm{H}(W_1) + \mathrm{H}(W_{>1} \mid W_1) \qquad (19)$$

We first bound the entropy of the first character using a uniform distribution upperbound:

$$\mathrm{H}(W_1) = -\sum_{\boldsymbol{w}_1 \in \Sigma} p(\boldsymbol{w}_1) \log p(\boldsymbol{w}_1) \qquad (20)$$
$$\leq -\log|\Sigma|$$

We now use the ε-smoothness property to upperbound the entropy of the following characters given the first.

$$\mathrm{H}(W_{>1} \mid W_1) \qquad\qquad (21)$$
$$= \sum_{\boldsymbol{w}_1 \in \Sigma} p(\boldsymbol{w}_1)\mathrm{H}(W_{>1} \mid W_1 = \boldsymbol{w}_1)$$
$$= p(\text{EoW})\mathrm{H}(W_{>1} \mid W_1 = \text{EoW})$$
$$+ \sum_{\boldsymbol{w}_1 \in \Sigma, \boldsymbol{w}! = \text{EoW}} p(\boldsymbol{w}_1)\mathrm{H}(W_{>1} \mid W = \boldsymbol{w}_1)$$
$$= \sum_{\boldsymbol{w}_1 \in \Sigma, \boldsymbol{w}! = \text{EoW}} p(\boldsymbol{w}_1)\mathrm{H}(W_{>1} \mid W_1 = \boldsymbol{w}_1)$$
$$\leq \sum_{\boldsymbol{w}_1 \in \Sigma, \boldsymbol{w}! = \text{EoW}} p(\boldsymbol{w}_1)\mathrm{H}(W_{>1})$$
$$\leq (1 - \varepsilon)\,\mathrm{H}(W)$$

Now that we have both this upperbounds, we can use them to bound the initial equation:

$$\mathrm{H}(W) = \mathrm{H}(W_1) + \mathrm{H}(W_{>1} \mid W_1) \qquad (22)$$
$$\leq -\log|\Sigma| + (1 - \varepsilon)\,\mathrm{H}(W)$$

Here, it's likely that the equations have to be broken over two lines.

A.4: Example 3

Finally, consider Jennifer White's paper "A Non-Linear Structural Probe"

```
> python3 aclpub_check/formatchecker.py --paper_type short structural.pdf
```

```
Checking structural.pdf
All Clear!
```

So, there were no mistakes!

The script checks for many violations, e.g. page-limit violations, font and font size violations and even a few common typos. The script even makes recommendations about citing non-arXiv versions of papers. These, however, are simply warnings:

Warning (Bibliography): It appears you are using arXiv links more than you should (18/55). Consider using ACL Anthology DOIs instead.