Trying Tries: Stochastic Morpheme Segmentation

Denizhan Pak
Dept. of Linguistics - Indiana University
denpak@iu.edu

Abstract

There are many reasons to apply morphological analysis for complex linguistic tasks. The application of computational tools to corpus data is heavily improved by increased information. Morphological analysis can provide novel information that can benefit downstream tasks. Semantics and dependencies can also be captured more easily given more morphological information. To determine the list of morphemes in a language however is a time consuming task and an unsupervised algorithm could relieve quite a few researchers and grad students. In this paper we propose a potentially useful unsupervised machine learning to accomplish just this task.

1 Introduction

The process of finding meaningful sub-sequences in a larger sequential dataset is a difficult and important task. This is especially true in linguistics. Morphology is a crucial element of any coherent study of language.

2 General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors' names and complete addresses, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection 3.6). **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 5. Pages are numbered for initial submission. However, **do not number the pages in the camera-ready version**.

By uncommenting `\aclfinalcopy` at the top of this document, it will compile to produce an example of the camera-ready formatting; by leaving it commented out, the document will be anonymized for initial submission. When you first create your submission on softconf, please fill in your submitted paper ID where `***` appears in the `\def\aclpaperid{***}` definition at the top.

The review process is double-blind, so do not include any author information (names, addresses) when submitting a paper for review. However, you should maintain space for names and addresses so that they will fit in the final (accepted) version. The ACL 2018 LaTeX style will create a titlebox space of 6.35 cm for you when `\aclfinalcopy` is commented out.

2.1 The Ruler

The ACL 2018 style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document without the provided style files, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\aclfinalcopy` command in the document preamble.)

Reviewers: note that the ruler measurements do not align well with lines in the paper – this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. In most cases one would expect that the approximate location will be adequate, although you can also use fractional references (*e.g.*, the first paragraph on this page ends at

mark 117.5).

2.2 Electronically-available resources

ACL provides this description in LaTeX2e (`acl2018.tex`) and PDF format (`acl2018.pdf`), along with the LaTeX2e style file used to format it (`acl2018.sty`) and an ACL bibliography style (`acl_natbib.bst`) and example bibliography (`acl2018.bib`). These files are all available at http://acl2018.org/downloads/acl18-latex.zip.

A Microsoft Word template file (`acl18-word.docx`) and example submission pdf (`acl18-word.pdf`) is available at http://acl2018.org/downloads/acl18-word.zip. We strongly recommend the use of these style files, which have been appropriately tailored for the ACL 2018 proceedings.

2.3 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from LaTeX using the *pdflatex* command. If your version of LaTeX produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one where it was created.** Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying "Output to a file", then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. When working with `dvips`, for instance, one should specify `-t a4`. Or using the command `\special{papersize=210mm,297mm}` in the latex preamble (directly below the `\usepackage` commands). Then using `dvipdf` and/or `pdflatex` which would make it easier for some.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs as soon as possible.

2.4 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Column width: 7.7 cm
- Column height: 24.7 cm
- Gap between columns: 0.6 cm

Papers should not be submitted on any other paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

2.5 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

2.6 The First Page

Center the title, author name(s), and affiliation(s) across both columns (or, for the initial submission, **Anonymous ACL submission** for names and affiliations). Do not use footnotes for affiliations. Include the paper ID number assigned during the submission process in the header. Use the two-column format only when you begin the abstract.

Title: Place the title centered at the top of the first page, in a 15-point bold font. (For a complete

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
subsection titles	11 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	11 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

guide to font sizes and styles, see Table 1) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author name(s), and the affiliation(s) on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (*e.g.*, use "Mitchell" not "MITCHELL"). Do not format title and section headings in all capitals as well except for proper names (such as "BLEU") that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronic paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side. Center the word **Abstract** above the body of the abstract using the font size and style shown in Table 1. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The font size of the abstract text should be as shown in Table 1.

Text: Begin typing the main body of the text

Command	Output
{\"a}	ä
{\^e}	ê
{\`i}	ì
{\.I}	İ
{\o}	ø
{\'u}	ú
{\aa}	å

Command	Output
{\c c}	ç
{\u g}	ğ
{\l}	ł
{\~n}	ñ
{\H o}	ő
{\v r}	ř
{\ss}	ß

Table 2: Example commands for accented characters, to be used in, *e.g.*, BIBTEX names.

immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers in the final version.

Indent: Indent when starting a new paragraph, about 0.4 cm.

2.7 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections (*i.e.*, use `\subsubsection*` instead of `\subsubsection`).

Citations: Citations within the text appear in parentheses as (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Using the provided LATEX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (**??**); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved.

Also refrain from using full citations as sentence constituents. We suggest that instead of

"(**?**) showed that ..."

you use

"Gusfield (**?**) showed that ..."

If you are using the provided LATEX and BibTEX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

You can use the command `\citealp` (alternative cite without parentheses) to get "author year" citations (which is useful for using citations within parentheses, as in **?**).

output	natbib	previous ACL style files
(**?**)	`\citep`	`\cite`
?	`\citet`	`\newcite`
(**?**)	`\citeyearpar`	`\shortcite`

Table 3: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

If the BibTeX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LaTeX package. To disable the hyperref package, load the style file with the `nohyperref` option: `\usepackage[nohyperref]{acl2018}`.

Compilation Issues: Some of you might encounter the following error during compilation:

"`\pdfendlink` *ended up in different nesting level than* `\pdfstartlink`."

This happens when `pdflatex` is used and a citation splits across a page boundary. To fix this, disable the `hyperref` package (see above), recompile and see the problematic citation. Next rewrite that sentence containing the citation. (See, *e.g.*, `http://tug.org/errors.html`)

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. We are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use BibTeX records that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at `http://aclanthology.info/`.

As examples, we cite (**?**) to show you how papers with a DOI will appear in the bibliography. We cite (**?**) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, *e.g.*,

"We previously showed (**?**) ..."

should be avoided. Instead, use citations such as

"**?** (**?**) previously showed ... "

Please do not use anonymous citations and do not include acknowledgments when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (**?**). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (**?**).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

2.8 Footnotes

Footnotes: Put footnotes at the bottom of the page and use the footnote font size shown in Table 1. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

2.9 Figures and Tables

Placement: Place figures and tables in the paper near where they are first discussed, as close as possible to the top of their respective column.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1: Caption of the Figure." "Table 1: Caption of the Table." Type the captions of the figures and tables below the body, using the caption font size shown in Table 1.

[1] This is how a footnote should appear.

[2] Note the line separating the footnotes from the text.

2.10 Equation

An example equation is shown below:

$$A = \pi r^2 \quad (1)$$

The numbering (if any) and alignment of the equations will be done automatically (using `align` or `equation`).

2.11 Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. A simple criterion: All curves and points in your figures should be clearly distinguishable without color.

3 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration "translation".

4 Length of Submission

The ACL 2018 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for references – so that reviewers' comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references.

For both long and short papers, all illustrations and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

Workshop chairs may have different rules for allowed length and whether supplementary material is welcome. As always, the respective call for papers is the authoritative source.

5 Supplementary Material

ACL 2018 also encourages the submission of supplementary material to report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Essentially, supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data.

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.

The supplementary material does not form part of the paper, does not count towards the page limit, and should not be included in an "Appendix" section following the references in this template. The "container" for supplementary materials is a separate document, and such materials should be submitted separately from the paper using the appropriate fields on the review form and the camera-ready upload form.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section (*i.e.*, use `\section*` instead of `\section`). Do not include this section when submitting your paper for review.