EN.600.461 Computer Vision
Final Project
Recognizing and Translating Text from Images

Joon Hyuck Choi
(Senior Undergraduate)
The Johns Hopkins University
3400 N Charles Street
Baltimore, MD 21218, USA
jchoi100@jhu.edu

Abstract

This work explores optical character recognition (OCR) in photos of printed and handwritten documents. It first explores basic preprocessing of photos of documents using the OpenCV library such as thresholding and denoising. It next discusses the use of the tesseract-ocr library to perform OCR. It finally discusses the incorporation of the Google Translate API to translate the output text from tesseract-ocr into multiple different languages.

1 Introduction

The main goals of this project are 1) Given a photo of a document (either hand-written or printed), convert it into a clean scanned version; 2) Take the clean scanned version and output OCR results. Optional goal mentioned in the original proposal that was met in this work is taking the OCR output and translating the text into different languages using the Google Translate API.

2 How to Run the Code

Assuming that the user has Python 2.7+, pytesseract, and tesseract installed, run the following command on the command line in the same directory as `driver.py`.

```
$python driver.py
{path-to-image-file} {list of
target languages separated by
single space}
```

Example usage:

`$python driver.py doc7.png es fr`

performs OCR on `doc7.png`, outputs text in the original input language, and outputs the text translated in Spanish and French.

2.1 Electronically-available resources

NAACL HLT provides this description in LaTeX2e (`naaclhlt2010.tex`) and PDF format (`naaclhlt2010.pdf`), along with the LaTeX2e style file used to format it (`naaclhlt2010.sty`) and an ACL bibliography style (`naaclhlt2010.bst`). These files are all available at `http://naaclhlt2010.isi.edu`. A Microsoft Word template file (`naaclhlt2010.dot`) is also available at the same URL. We strongly recommend the use of these style files, which have been appropriately tailored for the NAACL HLT 2010 proceedings.

2.2 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). This format can be generated from postscript files: on Unix systems, you can use `ps2pdf` for this purpose; under Microsoft Windows, you can use Adobe's Distiller, or if you have cygwin installed, you can use `dvipdf` or `ps2pdf`. Note that some word processing programs generate PDF which may not include all the necessary fonts (esp. tree diagrams, symbols). 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 processor may generate very large postscript/PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the postscript and/or 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.

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.

Additionally, it is of utmost importance to specify the **US-Letter format** (8.5in $\times$ 11in) when formatting the paper. When working with `dvips`, for instance, one should specify `-t letter`.

Print-outs of the PDF file on US-Letter 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 above as soon as possible.

2.3 Layout

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

- Left and right margins: 1in
- Top margin:1in
- Bottom margin: 1in
- Column width: 3.15in
- Column height: 9in
- Gap between columns: 0.2in

Papers should not be submitted on any other paper size. Exceptionally, authors for whom it is *impossible* to format on US-Letter paper, may format for *A4* paper. In this case, they should keep the *top* and *left* margins as given above, use the same column width, height and gap, and modify the bottom and right margins as necessary. Note that the text will no longer be centered.

2.4 The First Page

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. 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 guide to font sizes and styles, see Table 1.) Long title should be typed on two lines without a blank line intervening. Approximately, put the title at 1in from the top of the page, followed by a blank line, then the author's names(s), and the affiliation 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., "Leacock," not "LEACOCK"). The affiliation should contain the author's complete address, and if possible an electronic mail address. Leave about 0.75in between the affiliation and the body of the first page.

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 0.25in on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. 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 abstract text should be in 10 point font.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers.

Indent when starting a new paragraph. For reasons of uniformity, use Adobe's **Times Roman** fonts, with 11 points for text and subsection headings, 12 points for section headings and 15 points for the title. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default; see section 2.2 above). Note that the latter is about 10% less dense than Adobe's Times Roman font.

2.5 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.

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972).

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain 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* (American Psychological Association, 1983). 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* (Association for Computing Machinery, 1983).

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.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

Acknowledgment sections should go as a last (unnumbered) section immediately before the references.

2.6 Footnotes

Footnotes: Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2] Footnotes should be in 9 point font.

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
document text	11 pt	
abstract text	10 pt	
captions	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

2.7 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns and should be placed at the top of a page. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

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 10 point text.

3 Length of Submission

The NAACL HLT 2010 main conference accepts submissions of long papers and short papers. The maximum length of a long paper manuscript is eight (8) pages of content and one (1) additional page of references *only* (appendices count against the eight pages, not the additional one page). The maximum length of a short paper manuscript is four (4) pages including references. For both long and short papers, all illustrations, references, and appendices must be accommodated within these page limits, observing the formatting instructions given in the present document. Papers that do not conform to

[1] This is how a footnote should appear.

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

the specified length and formatting requirements are subject to be rejected without review.

Acknowledgments

Do not number the acknowledgment section.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.