Decoding Running Key Ciphers

Victor Redko
260659220

Marie Payne
260686859

Abstract

Artificial intelligence and natural language processing techniques can be used to decode substitution ciphers given the plaintext and keytext follow conventions of the English language. Groupings of n letters for n as large as 6 can be used in decoding methods comprising the Viterbi algorithm and in classifiers such as Multinomial Naive Bayes, logistic regression and support vector machines. We evaluate these methods with a baseline comparison model.

1 Introduction

In cryptography, a letter substitution cipher is constructed by inputting a plaintext and a key into a substitution function to produce a ciphertext. The original plaintext can be recovered by inputting the ciphertext and the key into the inverse of the function. A common choice for the substitution scheme is the *tabula recta,* where c = (p + r) mod 26 for same-case letters in the English alphabet, where c is the ciphertext, p is the plaintext and r is the keystream. If the plaintext or key contain patterns that can be deduced, then they can be decoded through the ciphertext by the means of statistical attacks. In the case that the key is perfectly random, never reused, and kept secret, this results in an unbreakable one-time pad. A variation of the substitution cipher, the running key cipher, uses a stream of characters from, for example, a book to create a long non-repeating key.

2 General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, as well as the authors' names and complete addresses (only in the final version, not in the version submitted for review), which must be centered at the top of the first page (see the guidelines in Subsection 2.5), and any full-width figures or tables. Type single-spaced. Do not number the pages in the camera-ready version. Start all pages directly under the top margin. See the guidelines later regarding formatting the first page.

The maximum length of a manuscript is eight (8) pages for the main conference, printed single-sided, plus two (2) pages for references (see Section 3 for additional information on the maximum number of pages).

By uncommenting `\naaclfinalcopy` 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\naaclpaperid{***}` 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 NAACL HLT 2016 LaTeX style will create a titlebox space of 2.5in for you when `\naaclfinalcopy` is commented out.

2.1 The Ruler

The NAACL HLT 2016 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 `\naaclfinalcopy` 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. Just use fractional references (e.g., the first line on this page is at mark 096.5), although in most cases one would expect that the approximate location will be adequate.

2.2 Electronically-available resources

NAACL HLT provides this description in LaTeX2e (`naaclhlt2016.tex`) and PDF format (`naaclhlt2016.pdf`), along with the LaTeX2e style file used to format it (`naaclhlt2016.sty`) and an ACL bibliography style (`naaclhlt2016.bst`) and example bibliography (`naaclhlt2016.bib`). These files are all available at `naacl.org/naacl-hlt-2016`. A Microsoft Word template file (`naaclhlt2016.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 2016 proceedings.

2.3 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 that 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 processors 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.

2.4 Layout

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

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

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.

2.5 The First Page

Center the title, author name(s) and affiliation(s) across both columns (or, in the case of initial submission, space for the names). Do not use footnotes for affiliations. 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 2.) Long titles 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 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., "Mitchell," not "MITCHELL"). 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 in the camera-ready manuscript.

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 2.3 above). Note that the latter is about 10% less dense than Adobe's Times Roman font.

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 1: Example commands for accented characters, to be used in, e.g., BibTeX names.

2.6 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). Using the provided LaTeX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972); 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 (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved.

References: We recommend including references in a separate `.bib` file, and include an example file in this release (`naalhlt2016.bib`). Some commands for names with accents are provided for convenience in Table 1. References stored in the separate `.bib` file are inserted into the document using the following commands:

```
\bibliography{naaclhlt2016}
\bibliographystyle{naaclhlt2016}
```

References should appear under the heading **References** at the end of the document, but before any

Appendices, unless the appendices contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a reference 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). Authors' full names rather than initials are preferred. You may use **standard** abbreviations for conferences[1] and journals[2].

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.

2.7 Footnotes

Footnotes: Put footnotes at the bottom of the page. They may be numbered or referred to by asterisks or other symbols.[3] Footnotes should be separated from the text by a line.[4] Footnotes should be in 9 point font.

2.8 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:** Figure caption.", "**Table 1:** Table caption." Type the captions of the figures and tables below the body, using 9 point text. Table and Figure labels should be bold-faced.

[1] https://en.wikipedia.org/wiki/List_of_computer_science_conference_acronyms

[2] http://www.abbreviations.com/jas.php

[3] This is how a footnote should appear.

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

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	9 pt	
caption label	9 pt	bold
bibliography	10 pt	
footnotes	9 pt	

Table 2: Font guide.

2.9 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.

3 Length of Submission

The NAACL HLT 2016 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 9 pages with 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 appendices 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.

4 Double-blind review process

As the reviewing will be blind, the paper must not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, e.g., "We previously showed (Smith, 1991) ..." must be avoided. Instead, use citations such as

"Smith previously showed (Smith, 1991) ..." Papers that do not conform to these requirements will be rejected without review. In addition, please do not post your submissions on the web until after the review process is complete (in special cases this is permitted: see the multiple submission policy below).

We will reject without review any papers that do not follow the official style guidelines, anonymity conditions and page limits.

5 Multiple Submission Policy

Papers that have been or will be submitted to other meetings or publications must indicate this at submission time. Authors of papers accepted for presentation at NAACL HLT 2016 must notify the program chairs by the camera-ready deadline as to whether the paper will be presented. All accepted papers must be presented at the conference to appear in the proceedings. We will not accept for publication or presentation papers that overlap significantly in content or results with papers that will be (or have been) published elsewhere.

Preprint servers such as arXiv.org and ACL-related workshops that do not have published proceedings in the ACL Anthology are not considered archival for purposes of submission. Authors must state in the online submission form the name of the workshop or preprint server and title of the non-archival version. The submitted version should be suitably anonymized and not contain references to the prior non-archival version. Reviewers will be told: "The author(s) have notified us that there exists a non-archival previous version of this paper with significantly overlapping text. We have approved submission under these circumstances, but to preserve the spirit of blind review, the current submission does not reference the non-archival version." Reviewers are free to do what they like with this information.

Authors submitting more than one paper to NAACL HLT must ensure that submissions do not overlap significantly ($> 25\%$) with each other in content or results. Authors should not submit short and long versions of papers with substantial overlap in their original contributions.

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.

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

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