Automatic text summarization, 2018

Anonymous Author(s)
Affiliation
Address
email

Abstract

Today there are many documents, articles, papers and reports available in digital form. These volumes of text are invaluable sources of information and knowledge that need to be effectively summarized to be useful. In automatic text summarization machine learning techniques are often used to generate summaries. A prior step to the generation of summaries is usually the extraction of nuggets. This paper presents the two approaches we use for the extraction of nuggets, as well as a description of their effectiveness and shortcomings.

1 Introduction

With the dramatic growth of the internet, people are overwhelmed by the tremendous amount of online information and documents. This expansion in availability of data has demanded extensive research in the automatic generation of summaries from a collection of different type of text.

Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document.

In general, there are two different approaches for text summarization:*extraction* and *abstraction*

`https://cmt.research.microsoft.com/NIPS2018/`

Please read the instructions below carefully and follow them faithfully.

1.1 Style

$\sim 15\%$ more words in the paper compared to earlier years.

Authors are required to use the NIPS LaTeX style files obtainable at the NIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

1.2 Retrieval of style files

The style files for NIPS and other conference information are available on the World Wide Web at

`http://www.nips.cc/`

The file `nips_2018.pdf` contains these instructions and illustrates the various formatting requirements your NIPS paper must satisfy.

The only supported style file for NIPS 2018 is `nips_2018.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

New preprint option for 2018 If you wish to post a preprint of your work online, e.g., on arXiv, using the NIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit. Please **do not** use the `final` option, which should **only** be used for papers accepted to NIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `nips_2018.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 3, 4, and 5 below.

2 Evaluation

2.1 Manual evaluation

The summaries are given to human annotators for evaluation. The annotators are students who attend the same course but are in another work group (?). For evaluation Likert Scales are used. Since refernce summaries don't exist it can't be evaluated by comparing a summary with a gold standard. Furthermore the annotators shouldn't have to read all ... source documents of a summary to judge the summary itself. This process woud be too time-consuming. Instead items are used on the Likert Scale which can be judged by only reading the summary itself. In total there are eleven categories: "Grammaticality", Non-redundancy", Referential clarity", "Focus", "Structure", "Coherence", "Readability", "Information Content", "Spelling", "Length" and "Overall Quality". For each category the annotators should assign a score from 1 (= very poor) to 5 (= very good), a weight and a confidence (both scales also from 1 to 5) of their grading. Each summary is evaluated by four to five different annotators.

Besides the summaries of all groups summaries created by two simple approaches (footnote) are evaluated as well. These summaries serve as baseline summaries. The first approach is ... The second approach is ...

Most categories seem like any text evaluation categories like "Spelling" and "Grammaticality". Other categories seem especially summary-related. These are the categories "Information Content" and "Focus". They represent the goal of a summary very well which is to present the most important content of the summarized texts. Since all summarized texts in this corpus are about a certain query the focus should be visible, too.

The resulting evaluations can be used for assessing the quality of the summaries produced by our system. It is important for the evaluation that we only work at the nugget extraction. This input is given to another group which then produced the summaries. In this way we are completely responsible for the results in some evaluation categories while other evaluation results also depend on the steps of building the hierarchy and actually creating a summary. The output which we after the nugget extraction are whole sentences (more about the output in section ...). The summary is then only built out of these sentences. In this way all categories which just operate on a sentence level are completely our responsibility. Among these categories are strictly only the two categories "Spelling" and "Grammaticality". We are also highly responsible for the categories "Information Content", "Focus" and "Non-Redundancy". All extracted sentences should ideally contain importannt information related to the query. Furthermore it can be argued that in the step of nugget extraction nuggets with the same meaning as another nugget are ignored. The categories "Referential Clarity", "Structure" and "Coherence" in comparison are very dependent on the ordering of the sentences. It can be argued that "Referential clarity" is also influenced by the nugget extraction. For sentences with a pronoun the system should also extract the reference sentence. Otherwise the sentence is not well usable in the next steps. This isn't done in the step of nugget extraction, but in later steps. The

category "Length" especially depends on the last step, the summary creation. "Readability" and of course "Information Content" are very general categories which can't be assigned to any particular step. The focus of our analysis will be all steps which can be influenced by our work, the nugget extraction. Thus the categories "Structure", "Length" and "Coherence" will only be shortly discussed.

In the following we compare the results of our group with the results of the other groups and the two baseline approaches. Our average overall score is 2.86. The average overall scores of the other groups are 0.39 to 0.74 points better. In contrast to the baseline approaches our summaries are much better. The baseline approaches only have an average overall score of 1.61 and 1.62. So our approach is more than one point better than the baselines. Now we take a closer look at the different categories. "Overall Qualiy" isn't discussec here because it does not highlight a particular aspect of a summary. Compared to the other groups our summaries are worst in all categories except for "Referential Clarity". In the category "Information Content" which is very important for summaries we outdo both baseline approaches significantly at least. The categories we are best at are "Spelling" with ..., "Non-Redundancy" with ... and "Grammaticality" with .... The other groups also perform best at "Grammaticality" and "Spelling"??? This is not surprising since all groups extracted whole sentences for the summarization. These sentences should be mostly grammatical, correctly spelled sentences. Perhaps there are some exceptions since the sentences are taken from forum posts. categories we are worst in are "Structure" with 2.86 points, "Coherence" with 2.88 points and "Information Content" with 2.9 points. "Structure and "Coherence" are also the categories the other groups perforn worst.???

3 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 5 regarding figures, tables, acknowledgments, and references.

4 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

4.1 Headings: second level

Second-level headings should be in 10-point type.

4.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

5 Citations, figures, tables, references

These instructions apply to everyone.

5.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `nips_2018` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{nips_2018}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

5.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

5.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

5.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

[1] Sample of the first footnote.

[2] As in this example.

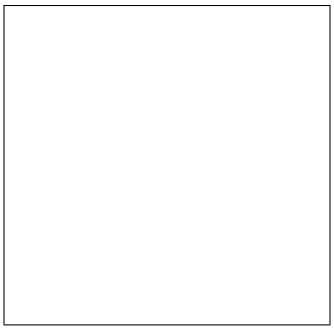

Figure 1: Sample figure caption.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

7 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

```
\usepackage{amsfonts}
```

followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!\!\!C} %complex numbers
```

Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

7.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use more than eight pages as long as the additional pages contain *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.