A Transformer-based Pretrained Language Model for Automatic Document Summarization

Anonymous ACL submission

Abstract

Document summarization is an essential task in the information retrieval. User intention for specific documents is hard to satisfy for the information overload, thus there are rarely golden standards for matching a certain document. I propose a multi-grained summarization model for documents that uses attention on paragraphs, supervised learning on the sentence importance, and sequence-to-sequence learning on characters and words. Tested on the 2020 Chinese Law Summarization Dataset, the proposed model achieved better performance than other modern summarization models.

1 Introduction

Document matching is an essential task in daily work and study. Because the matching algorithm is not capable of choosing important words from the massive documents, to use the content of a document as the matching criteria often fails in returning unintended documents. Document summarization condenses long documents into short paragraphs by retaining core information. Thus, a much wiser way for the document matching is using the summarization instead of the content.

2 Related Work

2.1 Sequence by Sequence in Summarization

3 Length of Submission

The 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.

The conference encourages the submission of additional material that is relevant to the reviewers but not an integral part of the paper. There are two such types of material: appendices, which can be read, and non-readable supplementary materials, often data or code. Additional material must be submitted as separate files, and must adhere to the same anonymity guidelines as the main paper. The paper must be self-contained: it is optional for reviewers to look at the supplementary material. Papers should not refer, for further detail, to documents, code or data resources that are not available to the reviewers. Refer to Appendices A and B for further information.

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

4 Anonymity

As reviewing will be double-blind, papers submitted for review should not include any author information (such as names or affiliations). Furthermore, self-references that reveal the author's identity, *e.g.*,

> We previously showed (Gusfield, 1997)
> ...

should be avoided. Instead, use citations such as

Gusfield (1997) previously showed...

Please do not use anonymous citations and do not include acknowledgements. **Papers that do not conform to these requirements may be rejected without review.**

Any preliminary non-archival versions of submitted papers should be listed in the submission form but not in the review version of the paper. Reviewers are generally aware that authors may present preliminary versions of their work in other venues, but will not be provided the list of previous presentations from the submission form.

Once a paper has been accepted to the conference, the camera-ready version of the paper should include the author's names and affiliations, and is allowed to use self-references.

LaTeX-specific details: For an anonymized submission, ensure that `\aclfinalcopy` at the top of this document is commented out, and that you have filled in the paper ID number (assigned during the submission process on softconf) where `***` appears in the `\def\aclpaperid{***}` definition at the top of this document. For a camera-ready submission, ensure that `\aclfinalcopy` at the top of this document is not commented out.

5 Multiple Submission Policy

Papers that have been or will be submitted to other meetings or publications must indicate this at submission time in the START submission form, and must be withdrawn from the other venues if accepted by ACL 2020. Authors of papers accepted for presentation at ACL 2020 must notify the program chairs by the camera-ready deadline as to whether the paper will be presented. We will not accept for publication or presentation the papers that overlap significantly in content or results with papers that will be (or have been) published elsewhere.

Authors submitting more than one paper to ACL 2020 must ensure that submissions do not overlap significantly (¿25%) with each other in content or results.

6 Formatting 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 Section 6.5). **Type single-spaced.** Start all pages directly under the top margin. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section 3. Pages should be numbered in the version submitted for review, but **pages should not be numbered in the camera-ready version**.

LaTeX-specific details: The style files will generate page numbers when `\aclfinalcopy` is commented out, and remove them otherwise.

6.1 File Format

For the production of the electronic manuscript you must use Adobe's Portable Document Format (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. 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.

LaTeX-specific details: PDF files are usually produced from LaTeX using the `pdflatex` command. If your version of LaTeX produces Postscript files, `ps2pdf` or `dvipdf` can convert these to PDF. To ensure A4 format in LaTeX, use the command `\special{papersize=210mm,297mm}` in the LaTeX preamble (below the `\usepackage` commands) and use `dvipdf` and/or `pdflatex`; or specify `-t a4` when working with `dvips`.

6.2 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact

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	10 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

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.

6.3 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. If Times Roman is unavailable, you may use Times New Roman or **Computer Modern Roman**.

Table 1 specifies what font sizes and styles must be used for each type of text in the manuscript.

LaTeX-specific details: To use Times Roman in LaTeX2e, put the following in the preamble:

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

6.4 Ruler

A printed ruler (line numbers in the left and right margins of the article) should be presented in the version submitted for review, so that reviewers may comment on particular lines in the paper without circumlocution. 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.

Reviewers: note that the ruler measurements may 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.*, this line ends at mark 295.5).

LaTeX-specific details: The style files will generate the ruler when `\aclfinalcopy` is commented out, and remove it otherwise.

6.5 Title and Authors

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Place the title centered at the top of the first page, in a 15-point bold font. Long titles should be typed on two lines without a blank line intervening. Put the title 2.5 cm 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.*, use "Mitchell" not "MITCHELL"). Do not format title and section headings in all capitals 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.

The title, author names and addresses should be completely identical to those entered to the electronical 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.

Start the body of the first page 7.5 cm from the top of the page. **Even in the anonymous version of the paper, you should maintain space for names and addresses so that they will fit in the final (accepted) version.**

6.6 Abstract

Use two-column format when you begin the 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 0.6 cm 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.

6.7 Text

Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document.

Indent 0.4 cm when starting a new paragraph.

6.8 Sections

Format section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals.

6.9 Footnotes

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

6.10 Graphics

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. Color is allowed, but adhere to Section 7's guidelines on accessibility.

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. Captions should be placed below illustrations. Captions that are one line are centered (see Table 1). Captions longer than one line are left-aligned (see Table 2).

LaTeX-specific details: The style files are compatible with the caption and subcaption packages; do not add optional arguments. **Do not override the default caption sizes.**

[1]This is how a footnote should appear.

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

Command	Output	Command	Output
`{\"a}`	ä	`{\c c}`	ç
`{\^e}`	ê	`{\u g}`	ğ
`{\'i}`	ì	`{\l}`	ł
`{\.I}`	İ	`{\~n}`	ñ
`{\o}`	ø	`{\H o}`	ő
`{\'u}`	ú	`{\v r}`	ř
`{\aa}`	å	`{\ss}`	ß

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

6.11 Hyperlinks

Within-document and external hyperlinks are indicated with Dark Blue text, Color Hex #000099.

6.12 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).

Refrain from using full citations as sentence constituents. Instead of

> "(Gusfield, 1997) showed that ..."

write

> "Gusfield (1997) showed that ..."

LaTeX-specific details: Table 3 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations as in Gusfield (1997). You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations as in (Gusfield, 1997). 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 Gusfield, 1997).

6.13 References

Gather the full set of references together under the heading **References**; place the section before any Appendices. Arrange the references alphabetically by first author, rather than by order of occurrence in the text.

Output	natbib command	Old ACL-style command
(Gusfield, 1997)	`\citep`	`\cite`
Gusfield, 1997	`\citealp`	no equivalent
Gusfield (1997)	`\citet`	`\newcite`
(1997)	`\citeyearpar`	`\shortcite`

Table 3: Citation commands supported by the style file. The 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.

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 full names for authors, not just initials.

Submissions should accurately reference prior and related work, including code and data. If a piece of prior work appeared in multiple venues, the version that appeared in a refereed, archival venue should be referenced. If multiple versions of a piece of prior work exist, the one used by the authors should be referenced. Authors should not rely on automated citation indices to provide accurate references for prior and related work.

The following text cites various types of articles so that the references section of the present document will include them.

- Example article in journal: (Ando and Zhang, 2005).
- Example article in proceedings, with location: (Börschinger and Johnson, 2011).
- Example article in proceedings, without location: (Andrew and Gao, 2007).
- Example arxiv paper: (Rasooli and Tetreault, 2015).

LaTeX-specific details: The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `acl2020.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliographystyle{acl_natbib}
\bibliography{acl2020}
```

You can obtain the complete ACL Anthology as a BibTeX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the anthology and your own bib file, use the following instead of the above.

```
\bibliographystyle{acl_natbib}
\bibliography{anthology,acl2020}
```

6.14 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.

All camera-ready references are required to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Appropriate records should be found for most materials in the current ACL Anthology at `http://aclanthology.info/`. As examples, we cite (Goodman et al., 2016) to show you how papers with a DOI will appear in the bibliography. We cite (Harper, 2014) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

LaTeX-specific details: Please ensure that you use BibTeX records that contain DOI or URLs for any of the ACL materials that you reference. 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.

6.15 Appendices

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

7 Accessibility

In an effort to accommodate people who are colorblind (as well as those printing to paper), grayscale readability is strongly 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.

8 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"

9 LaTeX Compilation Issues

You may 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, the style file contains a patch consisting of two lines: (1) `\RequirePackage{etoolbox}` (line 455 in `acl2020.sty`), and (2) A long line below (line 456 in `acl2020.sty`).

If you still encounter compilation issues even with the patch enabled, disable the patch by commenting the two lines, and then disable the `hyperref` package by loading the style file with the `nohyperref` option:

```
\usepackage[nohyperref]{acl2020}
```

Then recompile, find the problematic citation, and rewrite the sentence containing the citation. (See, *e.g.*, `http://tug.org/errors.html`)

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

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.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Benjamin Börschinger and Mark Johnson. 2011. A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 10–18, Canberra, Australia.

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

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

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

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here.

LaTeX-specific details: Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may 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.** 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, pseudocode or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

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.