The importance of content in one's native language

Zubair Abid
20171076
IIIT Hyderabad
zubair.abid@research.iiit.ac.in

Abstract

This document contains the instructions for preparing a paper submitted to COLING-2020 or accepted for publication in its proceedings. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

A common problem with automated NLP language systems deployed on the internet – especially large-scale ones based on Deep learning – is that they tend to fail when coming across languages that do not contain a large amount of readily digitized training data. This is a rather common problem for many thus-called ”low-resource” languages, as their low quantities of easily available digitized and tagged data makes State of the Art (SOTA) performance impossible with the latest and greatest in Deep Learning, and they are thus restricted to simpler Machine-learning techniques.

Lack of content in native (hereon, it is assumed that English is not a native language for the majority of people, and hence ”native language” will exclude English – especially as it is already de-facto the ’native language of the web’, and its native speakers are thus not hampered in similar means) languages is a problem that plagues not only developers of Natural Language Processing (NLP) systems, but one that also – and, in fact, primarily – impacts the native speakers of the language itself. The problem is not so much technological as it is sociological: the reason for the (attempted) existence of aforementioned NLP tools is in fact to aid in some way to create a social impact in use by native speakers.

Digging deeper, a big part of why people attempt to create such NLP tools – like translators, summarisers, and what not – is to improve the language resources available online for people that speak a language that is not well represented on the internet, to enable accessibility of content on the World Wide Web. It is therefore ironic that the very problem these tools set out to solve – namely, the low representation of ”native” languages – are the reason for the non-functionality of these very tools.

But to conceive of the existence of a problems requires demonstration of it. It is not wont to simply *claim* that one's ”own” languages are important. It is necessary to demonstrate to the plain eye that it is so.

2 Problem description

The problem in itself is simple enough. Most languages apart from English are but mere second-class citizens on the train of the interwebs. In fact! To be a second-class citizen is in itself a privilege; one primarily offered only to prominent European and South-American languages. For the rest are limited to hobbyist domains at best, perhaps Wikipedia may be so kind enough as to give the language its own sub-domain and encyclopaedic homepage. And even so does not guarantee the language an unfettered space on the internet, as we found out recently (Canales, 2020).

This has a severe impact on several things. First, the fact that most people in the world do not speak English – only 1.27 billion out of an estimated 7.7 billion people on Earth (Ethnologue, 2019a), and even fewer as a native tongue - ranking 3rd on the list with only 379 million speakers, behind Spanish

(480 million) and far behind Mandarin (918 million) (Ethnologue, 2019b). This means that for a vast majority of the world, a vast majority of the world wide web is locked off to them, unless someone in their community makes an effort to translate a lot of the information, or set up similar products with more native twists to it. It is a tragedy of sorts, the world's largest ever Library of Alexandria at one's fingertips, indecipherable due to linguistic boundaries.

The goal, therefore, is to see why content in one's own language is vitally important, from multiple aspects - social, political, and economic.

3 Major Insights

We may already know – from intuition, or otherwise – that we are most comfortable conversing with one another in our native tongues. While not necessarily the language spoken by one's parents (the peer group one shares plays just as important a role, if not more than stated), it is undeniable that this language – what some might term as their "language of thought" – is the one in which, given they option, they'd prefer to go about their daily routine engaging in. This is one of they key insights we can bring into play for the observations that we shall tackle later on.

Another key insight we can embrace is the observed role of Google Translate, and other similar translation APIs, over the past few years. From being kludgy, unworkable rule and statistics based highly inaccurate systems they have evolved into a system that is still nowhere near any gold standard, but that can be genuinely considered to be a temporary stand-in, if nothing else, for an actual human translator - for after all the hangling and mangling and wearing about, the API *is* free (or available for use at a nominal charge). The point here is that more and more we have begun to rely, and dare I say trust, automated NLP systems to surpass the language barrier we ourselves do not have the time to.

4 Key Observations

In this section, we will attempt to break down, by the three general categories of Social, Political, and Economic, the various reasons for why language content is important. The breakdowns shall include both the reasons for the need for native language content, and the progress that can be made by the advancements made in language processing due to the increase in aforementioned native language content, opening up whole new worlds previously thought impermeable.

4.1 Social

4.1.1 Accessibility

The first, and probably one of the absolute key factors in the whole equation - is accessibility. Accessibility is not something well standardised on the internet; and even where the W3C, for example (Initiative (WAI),), has setup accessibility rules, they are primarily directed at persons with physical disabilities, rather than at people who might speak different languages than you. It is a bit ambitious, as one might imagine, to regulate the entirety of the internet to providing linguistic accessibility – an child's toy website cannot be expected to pay for translations into 176+ languages.

That being said, the primary advantage of native-language websites is that now anyone who speaks that language can read it. It is limited in its scope - for the widest reach, one would employ English – which as we saw earlier, almost 5/7ths of the world does not know, even as a second or third language. That said for the ones who do not speak English but do Persian, for instance, can now read the websites that have been made with Persian users in mind. It is not the entire vast expanse of the multilingual internet, but a significant portion nonetheless, and an essential step, as we have touched upon earlier but will discuss in detail in just a bit. Native-language content allows users of that language to access a seemingly infinite resource they were locked out of before. And websites can be translated, their reach is not necessarily limited by the geographical limits of their language, but the efforts of those bilinguals willing to put that much more into getting their community to grow one article at a time. A regulatory body in India (noa, 2016) claimed in 2016 that "Internet access" and "local content" are a must for Digital India success.

Access to the internet in a native language brings with it not only the expanse of content provided to the user in said native tongue, but also access to the vast and open nature of the internet in itself.

Collaborative encyclopaedias, open access to the latest state of the art in technological advancements, the wide world of open source. By enabling native language content on the internet, it enriches the lives of both those inducted and those inducting.

A key feature that should not be forgotten here is the incredible growth the vast multitudes of multilingual data will provide to researchers working in Information Retrieval and Natural Language Processing, enabling them to better overcome the problem of data sparsity. And the data will, for a while anyway, definitely never be enough – as seen in recent times, from 340 million parameters in BERT (Devlin et al., 2019), to the 1.5 billion parameters in GPT-2 (Radford et al., 2019), all the way up to 150 billion in GPT-3 (Brown et al., 2020), at which there is still no end in sight, Transformers get better the more data they get.

What that means is that due to the increased amount of language data and more users of said language, the tools to improve accessibility from outside to enable the users to explore far more of the internet than was previously possible, engaging other users to join and generate more content which in turn improves the systems further and further and on and on and so on.

But what does that *actually* mean? In case you have read the fantastic Hitchhikers Guide to the Galaxy, this means we functionality can get to the point of having a babel fish. From the lowered, demeaned status of stragglers on the internet highways suddenly everybody is upgraded to first class status, able to access and read and understand anything written in any language, including the highly abundant English side of the internet. Not to exploit the Library of Alexandria metaphor *too* much, but – I mean – the modern Library of Alexandria, at one's fingertips.

(It should go without saying that the vision described here is as far into the future as Artificial General Intelligence (AGI), and it is incredibly unlikely that the state of AI as it is today will be able to reach those staggering heights. Yet, are we not already there to an extent? Consider how often we use Google translate's ”Translate this page” to read an Arabic news report about our favourite football team. If such small steps are a reality now, it is not impossible to imagine a future – not a perfect one, perhaps, but one that exists *pretty darn well* for what it is.)

4.1.2 Education

- Education: - Language of learning for children. Cite everything you have. (Cummins, 1981) (Hudelson, 1987) (Hakuta and Snow, 1986) Put multiple sentence paragraphs for each. Put theory of knowledge. Put Laltu. (noa, 2020) - Talk about all the information on the internet. Videos of content. Medium blogs of content. Language-locked, still, but the student can access this. The teacher can access this. Scott's tots. - Make it Accessible to the regular non-english user, via translation systems. Now now just specific language, any language. english to any. any to english too. any to any.

- Culture - (Geser and Mulrenin, 2002)

4.2 Political

- Information should not be only english, in the language of the bourgeoisie. The proletariat must rise, and this must be in the native tongue.

4.3 Economic

5 Possible Approaches

6 Conclusions

7 Outline

- Generally, increased accessibility
  - Accessibility for people who do not speak, or read, english
  - Parallel, and general corpora for training translation models
  - Document stores that can be used to fine-tune NLP tools
- Betterment of Society and Education

  - Language of Learning for children
  - Educators accessing the internet for material
  - Students accessing the internet for material

- Increased interaction on websites
  - Bigger share of market is engaging with you

The following instructions are directed to authors of papers submitted to COLING-2020 or accepted for publication in its proceedings. All authors are required to adhere to these specifications. Authors are required to provide a Portable Document Format (PDF) version of their papers. **The proceedings are designed for printing on A4 paper.**

Authors from countries in which access to word-processing systems is limited should contact the publication co-chairs Derek F. Wong (`derekfw@umac.mo`), Yang Zhao (`yang.zhao@nlpr.ia.ac.cn`) and Liang Huang (`liang.huang.sh@gmail.com`) as soon as possible.

We may make additional instructions available at `http://coling2020.org/`. Please check this website regularly.

8 Problem Description

Manuscripts must be in single-column format. **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The lengths of manuscripts should not exceed the maximum page limit described in Section 10. Do not number the pages.

8.1 Electronically-available Resources

We strongly prefer that you prepare your PDF files using LaTeX with the official COLING 2020 style file (coling2020.sty) and bibliography style (acl.bst). These files are available in coling2020.zip at `http://coling2020.org/`. You will also find the document you are currently reading (coling2020.pdf) and its LaTeX source code (coling2020.tex) in coling2020.zip.

You can alternatively use Microsoft Word to produce your PDF file. In this case, we strongly recommend the use of the Word template file (coling2020.dotx) in coling2020.zip. If you have an option, we recommend that you use the LaTeX2e version. If you will be using the Microsoft Word template, you must anonymise your source file so that the pdf produced does not retain your identity. This can be done by removing any personal information from your source document properties.

8.2 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 for non-Latin 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`.

If you cannot meet the above requirements for the production of your electronic submission, please contact the publication co-chairs as soon as possible.

Place licence statement here for the camera-ready version. See Section 8.9 of the instructions for preparing a manuscript.

8.3 Layout

Format manuscripts with a single column 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
- Width: 16.0 cm
- Height: 24.7 cm

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

8.4 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.

The **Times New Roman** font, which is configured for us in the Microsoft Word template (coling2020.dotx) and which some Linux distributions offer for installation, can be used as well.

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	
captions	11 pt	
sub-captions	9 pt	
abstract text	11 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

8.5 The First Page

Centre the title, author's name(s) and affiliation(s) across the page. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. Do not include the authors' names or affiliations in the version submitted for review.

Title: Place the title centred at the top of the first page, in a 15 pt bold font. (For a complete 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'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 "Schlangen" not "SCHLANGEN"). 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 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 co-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 between addresses and main body. The width of the abstract text should be smaller than main body by about 0.6 cm on each side. Centre the word **Abstract** in a 12 pt 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 11 pt font.

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

Indent when starting a new paragraph. Use 11 pt for text and subsection headings, 12 pt for section headings and 15 pt for the title.

Licence: Include a licence statement as an unmarked (unnumbered) footnote on the first page of the final, camera-ready paper. See Section 8.9 below for details and motivation.

8.6 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. Do not number subsubsections.

Citations: Citations within the text appear in parentheses as (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Append lowercase letters to the year in cases of ambiguity. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved. Collapse multiple citations as in (**?**; **?**). 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 `\newcite` to get "author (year)" citations.

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

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

Please do not use anonymous citations and do not include any of the following when submitting your paper for review: acknowledgements, project names, grant numbers, and names or URLs of resources or tools that have only been made publicly available in the last 3 weeks or are about to be made public and would compromise the anonymity of the submission. Papers that do not conform to these requirements may be rejected without review. These details can, however, be included in the camera-ready, final paper.

In LaTeX, for an anonymized submission, ensure that `\colingfinalcopy` at the top of this document is commented out. For a camera-ready submission, ensure that `\colingfinalcopy` at the top of this document is not commented out.

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* (**?**). 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.

- Example citing an arxiv paper: (**?**).

- Example article in journal citation: (**?**).

- Example article in proceedings: (**?**).

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**.

8.7 Footnotes

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

8.8 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Colour 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 11 pt text.

Narrow graphics together with the single-column format may lead to large empty spaces, see for example the wide margins on both sides of Table 1. If you have multiple graphics with related content, it may be preferable to combine them in one graphic. You can identify the sub-graphics with sub-captions below the sub-graphics numbered (a), (b), (c) etc. and using 9 pt text. The LaTeX packages wrapfig, subfig, subtable and/or subcaption may be useful.

8.9 Licence Statement

As in COLING-2014, COLING-2016, and COLING-2018, we require that authors license their camera-ready papers under a Creative Commons Attribution 4.0 International Licence (CC-BY). This means that authors (copyright holders) retain copyright but grant everybody the right to adapt and re-distribute their paper as long as the authors are credited and modifications listed. In other words, this license lets researchers use research papers for their research without legal issues. Please refer to `http://creativecommons.org/licenses/by/4.0/` for the licence terms.

Depending on whether you use British or American English in your paper, please include one of the following as an unmarked (unnumbered) footnote on page 1 of your paper. The LaTeX style file (coling2020.sty) adds a command `blfootnote` for this purpose, and usage of the command is prepared in the LaTeX source code (coling2020.tex) at the start of Section “Introduction”.

[1] This is how a footnote should appear.

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

We strongly prefer that you licence your paper as the CC license above. However, if it is impossible you to use that license, please contact the COLING-2020 publication co-chairs Derek F. Wong (derekfw@umac.mo), Yang Zhao (yang.zhao@nlpr.ia.ac.cn) and Liang Huang (liang.huang.sh@gmail.com), before you submit your final version of accepted papers. (Please note that this license statement is only related to the final versions of accepted papers. It is not required for papers submitted for review.)

9 Major Insights

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".

10 Core Arguments

The maximum submission length is 9 pages (A4) of content for long papers and 4 pages (A4) of content for short papers, plus an unlimited number of pages for references (for both long and short papers). Authors of accepted papers will be given additional space in the camera-ready version to reflect space needed for changes stemming from reviewers comments.

Papers that do not conform to the specified length and formatting requirements may be rejected without review.

11 Overall Analysis

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

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July. arXiv: 2005.14165.

Katie Canales. 2020. For years, an American has been writing articles in a stereotypical Scottish accent on the official Scots Wikipedia, and some people online are not happy.

Jim Cummins. 1981. *Bilingualism and Minority-Language Children. Language and Literacy Series.* The Ontario Institute for Studies in Education, 252 Bloor Street West, Toronto, Ontario M5S 1V6 ($3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Ethnologue. 2019a. English.

Ethnologue. 2019b. What are the top 200 most spoken languages?

Guntram Geser and Andrea Mulrenin. 2002. *The DigiCULT report: technological landscapes for tomorrow's cultural economy: unlocking the value of cultural heritage: executive summary: January 2002.* Office for official publications of the European Communities.

K Hakuta and C Snow. 1986. *Compendium of Papers on the Topic of Bilingual Education of the Committee on Education and Labor, House of Representatives, 99th Congress, 2nd Session.* U.S. Government Printing Office. Google-Books-ID: yboZAAAAMAAJ.

Sarah Hudelson. 1987. The Role of Native Language Literacy in the Education of Language Minority Children. *Language Arts*, 64(8):827–841. Publisher: National Council of Teachers of English.

W3C Web Accessibility Initiative (WAI). W3C Accessibility Standards Overview.

2016. Internet access, local content must for Digital India success: IAMAI - The Economic Times.

2020. Knowledge and Language.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.