LEVERAGING LLMS FOR RETRIEVAL AUGMENTED TRANSLATION FOR ENGLISH-TO-ARABIC SUBTITLES

Mohamed Hannani
Researcher

Abdelhadi Soudi
Professor

Kristof Van Laerhoven
Professor

ABSTRACT

While machine translation (MT) has made significant progress in domain adaptation, achieving real-time adaptation for subtitles remains a formidable challenge. Large-scale language models (LLMs) have emerged as promising candidates for in-context learning. This study delves into the potential of in-context learning for real-time adaptive MT in the context of English-to-Arabic subtitles, with the utilization of the FAISS index for fuzzy match selection. Our experiments demonstrate the remarkable capacity of LLMs to adapt to in-domain content, surpassing traditional encoder-decoder MT models. Furthermore, we explore the integration of MT techniques, encompassing both robust encoder-decoder models and LLMs, with fuzzy matching. This integration reveals the potential to further elevate translation quality, a particularly valuable asset in the realm of English-to-Arabic subtitle translation, where language support may be limited. To validate our findings, we conducted extensive experiments focused on English-to-Arabic (EN-AR) subtitle translation.

1. INTRODUCTION

The ever-expanding landscape of natural language processing and machine translation has ushered in a new era of communication, significantly reducing linguistic divides and fostering cross-cultural understanding. While large language models, such as GPT, Llama 2, and the recently introduced Falcon have made remarkable strides in handling a wide range of language-related tasks, machine translation goes beyond mere word conversion. It entails the intricate task of preserving nuances, idioms, and the unique stylistic attributes that define human languages.

This brings us to the concept of adaptive translation, a paradigm aimed at enhancing machine translation by tailoring it to specific domains, genres, or styles. A key advantage of adaptive translation is its ability to achieve domain-specific translation goals without the resource-intensive processes of model training and fine-tuning. Our particular emphasis lies in harnessing the capabilities of Llama-2-70b-chat by Meta and GPT-3.5 Turbo by OpenAI with in-context samples.

This report delves into the intricacies of adapting machine translation to domain-specific requirements, utilizing a corpus of approximately 1,500 movie subtitles meticulously translated from English to Arabic. These subtitles encapsulate the nuances of cinematic language and provide a rich contextual source for our translation model. Prior to the inference phase, we leverage the Sentence-Transformer model to compute embeddings for these subtitles, streamlining the retrieval of similar sentences using the FAISS indexing system developed by Facebook. This approach enables us to construct contextually rich prompts, allowing Llama-2-70b-chat and GPT-3.5 Turbo to follow the stylistic cues present in domain-specific examples.

In the following sections, we provide a detailed account of our methodology, experimental setup, results, and engage in a broader discussion of the implications of adaptive translation in the field of machine learning.

2. PAPER LENGTH & FILE SIZE

We adopt a “(6+n)-page policy” for ISMIR 2023. That is, each paper may have a maximum of six pages of technical content (including figures and tables) with additional optional pages that contain only references and acknowledgments. Note that acknowledgments should not be included in the anonymized submission.

Paper should be submitted as PDFs and the file size is limited to 10MB. Please compress images and figures as necessary before submitting.

3. PAGE SIZE

The proceedings will be printed on portrait A4-size paper (21.0cm x 29.7cm). All material on each page should fit within a rectangle of 17.2cm x 25.2cm, centered on the page, beginning 2.0cm from the top of the page and ending with 2.5cm from the bottom. The left and right margins should be 1.9cm. The text should be in two 8.2cm columns with a 0.8cm gutter. All text must be in a two-column format. Text must be fully justified.

4. TYPESET TEXT

4.1 Normal or Body Text

Please use a 10pt (point) Times font. Sans-serif or non-proportional fonts can be used only for special purposes, such as distinguishing source code text.

The first paragraph in each section should not be indented, but all other paragraphs should be.

Attribution: F. Author, S. Author, and T. Author, “Leveraging LLMs for Retrieval Augmented Translation for English-to-Arabic Subtitles”, in *Proc. of the 24th Int. Society for Music Information Retrieval Conf.*, Milan, Italy, 2023.

4.2 Title and Authors

The title is 14pt Times, bold, caps, upper case, centered. Authors’ names are omitted when submitting for double-blind reviewing. The following is for making a camera-ready version. Authors’ names are centered. The lead author’s name is to be listed first (left-most), and the co-authors’ names after. If the addresses for all authors are the same, include the address only once, centered. If the authors have different addresses, put the addresses, evenly spaced, under each authors’ name.

4.3 First Page Copyright Notice

Please include the copyright notice exactly as it appears here in the lower left-hand corner of the page. It is set in 8pt Times. After your paper is accepted, you will need to insert the appropriate author names and paper title in the copyright notice when submitting the camera-ready version. For LaTeXusers, this will be handled by the template automatically. For Word users, this has to be done manually.

4.4 Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

4.5 Line Numbers

Line numbers should be included in your originally submitted manuscript, for reference during reviewing. However, after your paper is accepted, you must remove all line numbers from the final camera-ready version. This can be done in LaTeXby commenting out the command `\linenumbers` on Line 22. This can be done in Microsoft Word by selecting Layout > Line Numbers > None.

5. FIRST LEVEL HEADINGS

First level headings are in Times 10pt bold, centered with 1 line of space above the section head, and 1/2 space below it. For a section header immediately followed by a subsection header, the space should be merged.

5.1 Second Level Headings

Second level headings are in Times 10pt bold, flush left, with 1 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

5.1.1 Third and Further Level Headings

Third level headings are in Times 10pt italic, flush left, with 1/2 line of space above the section head, and 1/2 space below it. The first letter of each significant word is capitalized.

Using more than three levels of headings is highly discouraged.

String value	Numeric value
Hello ISMIR	2023

Table 1. Table captions should be placed below the table.

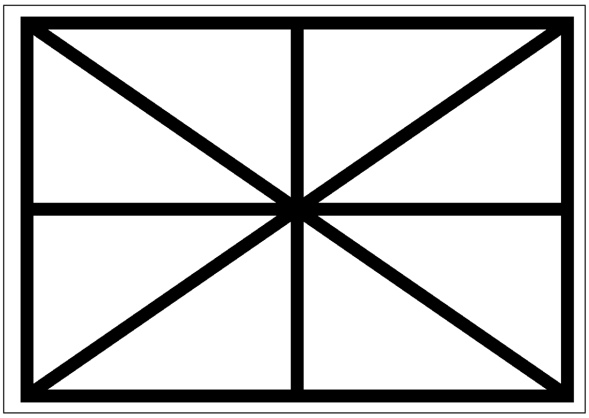

Figure 1. Figure captions should be placed below the figure.

6. FOOTNOTES AND FIGURES

6.1 Footnotes

Indicate footnotes with a number in the text. [1] Use 8pt type for footnotes. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a 0.5pt horizontal rule.

6.2 Figures, Tables and Captions

All artwork must be centered, neat, clean, and legible. All lines should be very dark for purposes of reproduction and art work should not be hand-drawn. The proceedings are not in color, and therefore all figures must make sense in black-and-white form. Figure and table numbers and captions always appear below the figure. Leave 1 line space between the figure or table and the caption. Each figure or table is numbered consecutively. Captions should be Times 10pt. Place tables/figures in text as close to the reference as possible. References to tables and figures should be capitalized, for example: see Figure 1 and Table 1. Figures and tables may extend across both columns to a maximum width of 17.2cm.

7. EQUATIONS

Equations should be placed on separate lines and numbered. The number should be on the right side, in parentheses, as in Eqn (1).

$$E = mc^2 \tag{1}$$

8. CITATIONS

All bibliographical references should be listed at the end of the submission, in a section named “REFER-

[1] This is a footnote.

ENCES," numbered and in the order that they first appear in the text. Formatting in the REFERENCES section must conform to the IEEE standard (`https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf`). Approved IEEE abbreviations (Proceedings → Proc.) may be used to shorten reference listings. All references listed should be cited in the text. When referring to documents, place the numbers in square brackets (e.g., [1] for a single reference, or [2–4] for a range).

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of [1]," not "In our previous work [1]." 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."

9. REFERENCES

[1] A. Author and B. Author, "The title of the conference paper," in *Proc. of the 18th Int. Society for Music Information Retrieval Conf.*, Suzhou, China, 2017, pp. 111–117.

[2] A. Someone, B. Someone, and C. Someone, "The title of the journal paper," *Journal of New Music Research*, vol. A, no. B, pp. 111–222, September 2010.

[3] O. Person, *Title of the Book*. Montréal, Canada: McGill-Queen's University Press, 2021.

[4] F. Person and S. Person, "Title of a chapter this book," in *A Book Containing Delightful Chapters*, A. G. Editor, Ed. Tokyo, Japan: The Publisher, 2009, pp. 58–102.