CS 454 Signature Project, Spring 2024
Report on the News Story, "A 'Shocking' Amount of the Web Is Already AI-Translated Trash, Scientists Determine"
Sample Assignment Response by Henry M. Walker

Tech News from the highly-regarded Association for Computing Machinery (ACM) reports an intriguing story, 'Shocking' Amount of the Web Is Already AI-Translated Trash", that discusses the wide-spread presence of Web pages generated through automate machine translation (MT) techniques. The article itself, reported by the Vice Media Group, has the subtitle, "Researchers warn that most of the text we view online has been poorly translated into one or more languages—usually by a machine" [4]. Overall, this news story raises several interesting and important issues.

1. Where did the reported story come from, and to what extent might these groups have potential interests or biases?
2. What patterns have been identified regarding Web pages and their translation into other languages?
3. To what extent is the quality of translations dependent upon the source and the target languages?
4. Why does the article call the current state of translated Web material "shocking"?

The following paragraphs address each of these questions in turn.

News Story Sources: When reading any news story or other report, it is important to determine the author(s) perspectives and interests in the material. Of course, everyone has a perspective, and readers need to appreciate the extent to which that perspective might impact the reporting. In this case, two groups might be considered.

1. The researchers are identified as working at the Amazon Web Services AI lab. This affiliation suggests that these people likely are quite knowledgable regarding both techniques in artificial intelligence and the content of the Web—this group likely can be considered as experts in the field. On the other hand, this Amazon group also is selling its services, so may have a vested interest in approaching this application in a certain way. Such an interest might color their research report or maybe not.

2. The story itself appears as a "Motherboard Tech by VICE" page, published by the Vice Media Group—a company that has been described as a "controversial counter-culture publisher that focuses on lifestyle, arts, culture, and news/politics." Further, according to Pull Rank, a "pioneering content marketing and enterprise SEO agency"", the Vice Media Group is particularly good at connecting with its audience, and PullRank ranks Vice Media as its number five successful media company, after Insider, USA Today, CNET, and the Huffington Post. [3]. Headlines and story content certainly may help connect with an audience, but again the nature of any coloring of story content might require additional analysis.

Patterns in Web Pages and their Translation: The basis for the reporting in this news story began with a collection of 6.38 billion sentences gathered from Web pages. Within that data, the research looked for "patterns of multi-way parallelism, which describes sets of sentences that are direct translations of one another in three or more languages" [4]. This study concluded that 57.1 percent of the sentences collected

consisted of these multi-way parallel sentences in three or more languages. Although the article does not clarify this count, it seems likely that in each case, one of the sentences would be in original form (without translation), and the remaining two or more versions were translated. Thus, removing the original version, the percentage of translated sentences would still be high, but likely under 50%—perhaps by a considerable amount.

Further discussion in this article involves the notion of high-resource and low-resource languages. "Technically speaking, whenever a language is lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical [natural language processing] NLP applications, it is considered a low-resource language." [2] Correspondingly, a high-resource language is one for which much source text is available. (A simple Web search for low- and high-resource languages yields many articles that related to machine translation and technical challenges.] Interestingly, at least two additional patterns were noted concerning these different categories of languages.

1. Sentences studied in low-resource languages (often from Africa or Asia) had an average parallelism of 8.6—that is, a sentence appeared in original form in one language, and an average of 7.6 sentences were translations. The corresponding average for high-resource languages was 4. Generally, the research showed that high-resource sentences are original or are translations from a high-resource language, such as English or French. If a sentence is translated into one low-resource language, it is likely to be translated to other such languages as well.
2. Sentences appearing in low-resource languages tend to be short (5-10 words) and "more predictable". Although not mentioned in the news story, short sentences typically must be reasonably simple—a sentence can have few multiple clauses, compound sentences, or dependent clauses within a 5-10 word limit. Further, the story indicated that these sentences often came from text that required little expertise, and suggested the topics often involved little depth.

Factors Impacting The Quality of Translation: According to the news story, the effectiveness of machine translation depends upon having vast data sets (e.g., billions or trillions of words and/or sentences). Further, for much work in artificial intelligence, the results of translation depend upon the quality of the underlying data.

By definition, extensive text is available for high-resource languages. However, even here, quality could be an issue. For example, if data come directly from original sources, quality may be strong. However, machine translation or other factors may impact quality, and some material on the Web may have deficiencies. For high-resource languages, one might expect a relatively high percentage of available text to be a reasonable quality.

On the other hand, the quality of machine translation for low-resources languages is impacted by a relative lack of original source material. Accessing text in these languages also may include numerous materials that have been machine translated. Altogether, there seems to be a potential downward cycle for machine translation: fewer original sources yields poorer machine-translated text; more translated text

yields a lower percentage of original sources in the overall source pool, and the lower percentage weakens what might be obtained by scraping the Web.

Is the Current State of Translated Material on the Web 'Shocking': The beginning of the news story discusses how this study of Web pages arose. Several native speakers of low-resource languages at the Amazon Web Services AI lab observed that much of the Web materials in their native languages seemed to be machine translated and of low quality. The purpose of the study was to gather data that might inform discussions of the amount and the quality of machine-translated material on the Web. In considering the data presented, it would seem that text in high-resource languages often will be either original or machine translated using extensive data. Either way, the text (original or translated) may be of reasonable quality.

However, text in low-resource languages often has neither of these characteristics: the text often is translated from another source, and the translation itself may not be particularly good (based on a relative lack of original material).

Certainly this situation will be of considerable concern to people speaking low-resource languages, and to those concerned with issues of diversity, equity, inclusion, etc. Whether this situation is "shocking" might be a matter of perspective and audience, but certainly this seems to be an issue within the research community and speakers of low-resource languages.

Conclusion

In conclusion, a Web search yields many articles describing challenges for the machine translation of materials involving low-resource languages. Certainly, the problem reported in this news story seems well documented.

Further, a Web search suggests that researchers are proposing a range of approaches to address the issue of quality of machine translation for text in low-resource languages. Of course, adding materials from native speakers can be helpful, but realistically work along this line can be limited. Looking forward, it seems important to follow additional research and development in this field.

References

1. Association for Computing Machinery (ACM), "'Shocking' Amount of the Web Is Already AI-Translated Trash", ACM Tech News, URL: https://technews.acm.org/archives.cfm?fo=2024-01-jan/jan-22-2024.html (accessed January 27, 2024).
2. Laumann, Felix, "Low-resource language: what does it mean?", medium.com, URL: https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5 (accessed January 24, 2024).
3. McDermott, Andrew, "13 Examples of Media Companies that Know their Audience" PullRank, URL: https://ipullrank.com/13-examples-of-media-companies-that-know-their-audience (accessed January 27, 2024).
4. Roscoe, Jules, "A 'Shocking' Amount of the Web Is Already AI-Translated Trash, Scientists Determine", Vice Media Group, January 12, 2024, URL: https://www.vice.com/en/article/y3w4gw/a-shocking-amount-of-the-web-is-already-ai-translated-trash-scientists-determine (accessed January 27, 2024).