Learning From Arabic Corpora But Not Always From Arabic Speakers:

A Case Study of the Arabic Wikipedia Editions

Saied Alshahrani Esma Wali Jeanna Matthews

INTRODUCTION:

- NLP is a key element in decision-making systems that need large human text corpora to understand humans.
- Human languages are under-represented in both corpora development and NLP toolchain support.
- Wikipedia corpora are widely used and could be developed/created through bots or automated scripts, or translated from other languages like English.

THE CASE OF WIKIPEDIA:

- Wikipedia has three Arabic Wikipedia editions: Modern Standard Arabic, Egyptian Arabic, and Moroccan Arabic.
- Egyptian Wikipedia's content pages (corpora) increased exponentially and rapidly in the last two years, while Arabic and Moroccan Wikipedia have grown normally.
- In June 2020, approximately 253,000 new content pages (articles) were created in the Egyptian Arabic Wikipedia.
- Over one million articles, 63% of the total content pages, in Egyptian Wikipedia have been automatically created by one registered user using template-based translation from English using Google Translate API.

DISCUSSION:

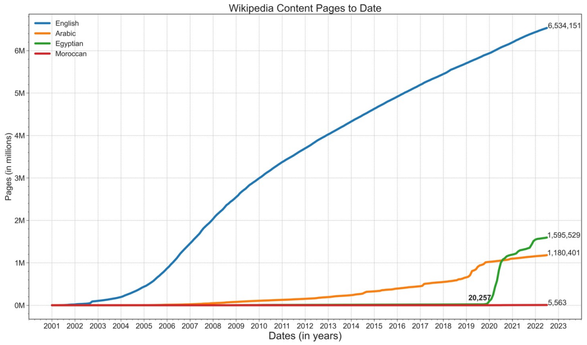
- The Arabic language poses many challenges to NLP tools due to its morphological richness and high ambiguity.
- Direct translation or off-the-shelf translation tools not only perform poorly but also show ethical problems.
- Wikipedia article is a factual entry, yet the choice to write an article on one topic over another reflects the author's perspective and values.
- It matters whether such a choice is made by native speakers or by translation from other languages.

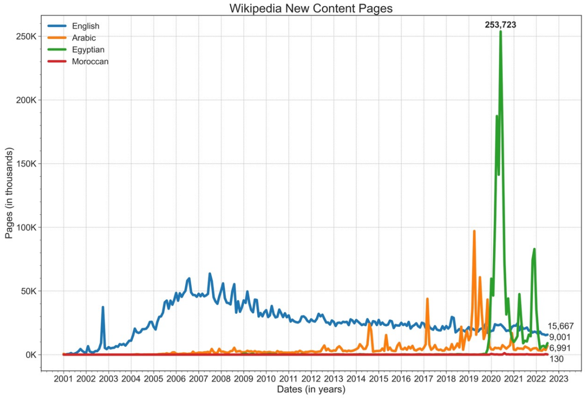
Natural Language Processing corpora, e.g., Wikipedia corpora (content pages) that are automatically created using shallow templated-based, poorly translated using direct translation, auto-generated by advanced Large Language Models, or even the assembled corpora using text augmentation techniques do not echo the complex structure of the Arabic language and its dialects, do not express the views of the Arabic speakers, and do not represent the cultural richness and historical heritage of the Arabic language and its people.

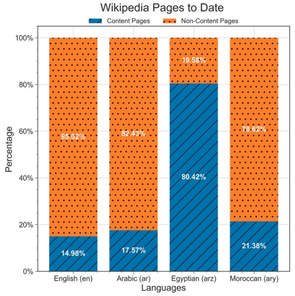
Visit the Project:
https://webspace.clarkson.edu/~alshahsf/Representativeness.html

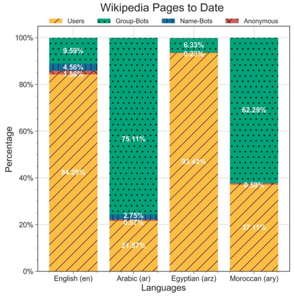
SCAN ME!

Clarkson









REPRODUCIBILITY:

Analysis of Arabic Wikipedia Editions

Implementation of Wikistats2csv