Clarkson
Accountability
Transparency
Research Group

Learning From Arabic Corpora But Not Always From Arabic Speakers: A Case Study of the Arabic Wikipedia Editions♔

Saied Alshahrani[1], CS PhD student♕
Advisor: Prof. Jeanna N. Matthews

♔Accepted at the Seventh Arabic Natural Language Processing Workshop (**WANLP 2022**), co-located with the 2022 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2022**), December 7-8, 2022.

♕The co-authors: Esma Wali[2], Jeanna Matthews[3].

Outline

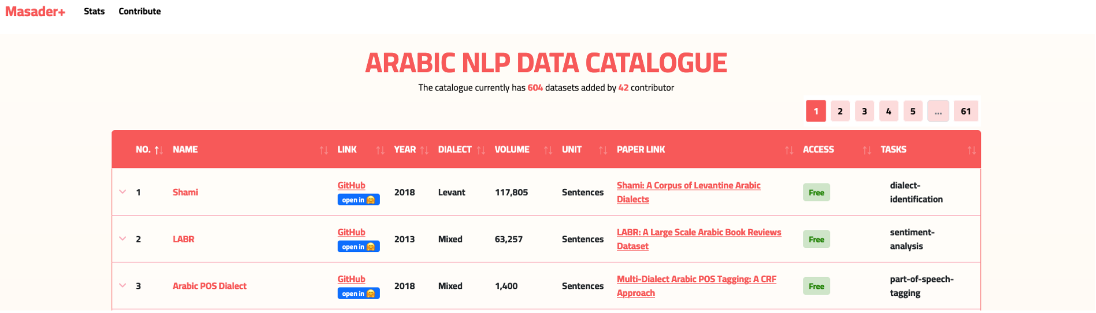
- Abstract
- Background
- Motivation
- Case Study
- Discussion
- Conclusion
- Recommendation
- Future Works
- References

Abstract

- We **present** a case study focusing on differences among the Arabic Wikipedia editions (Modern Standard Arabic (MSA), Egyptian, and Moroccan).
- We **document** issues in the Egyptian Arabic Wikipedia with automatic creation/generation and translation of articles from English without human intervention.

Background

- NLP is a **key** element in decision-making systems like resume parsers that sort lists of job candidates.
- These systems **need** large corpora of human text to understand humans to make recommendations [1].



Masader+ Stats Contribute

ARABIC NLP DATA CATALOGUE

The catalogue currently has 604 datasets added by 42 contributor

1 2 3 4 5 ... 61

NO.	NAME	LINK	YEAR	DIALECT	VOLUME	UNIT	PAPER LINK	ACCESS	TASKS
1	Shami	GitHub open in 🤗	2018	Levant	117,805	Sentences	Shami: A Corpus of Levantine Arabic Dialects	Free	dialect-identification
2	LABR	GitHub open in 🤗	2013	Mixed	63,257	Sentences	LABR: A Large Scale Arabic Book Reviews Dataset	Free	sentiment-analysis
3	Arabic POS Dialect	GitHub open in 🤗	2018	Mixed	1,400	Sentences	Multi-Dialect Arabic POS Tagging: A CRF Approach	Free	part-of-speech-tagging

[23]

Background

- These corpora of human text **convey** many social concepts, including culture, traditions, perspectives, and even historic biases [2, 3, 4, 5].
- Yet, having a corpus of text in a language **does not** necessarily represent the culture of native speakers of that language.

Background

- Human languages are under-represented in **both** corpora development and NLP toolchain support [1].
- This **under-represents** the culture and views of native speakers of those languages in NLP-guided decision-making.

Motivation

- Human corpora **should** be written by native speakers, yet others may be written by non-native speakers or translated from other languages [6].
- Such unrepresentative corpora **could** affect the performance of many NLP tasks such as Large Language Models (LLMs) trained from these corpora.

Motivation

- Wikipedia corpora (content pages) are widely **used** in NLP to train LLMs like ELMo, BERT, and GPT-3 [7, 8, 9].
- Some Wikipedia corpora have been developed/created through bots or automated scripts ; this often involves translation from other languages [10].

Motivation

- This work highlights the differences between text corpora written by native speakers and those translated and generated by automated systems.

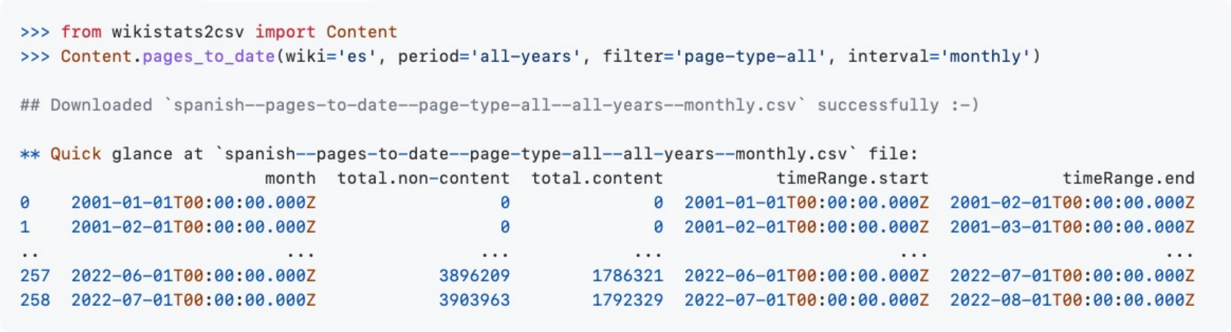
The Case of Wikipedia

- We **compare** the Arabic Wikipedia editions regarding pages to date, new pages, and top editors, besides English Wikipedia as a benchmark.
    - We took a data snapshot of the **four** Wikipedia editions' statistics in July 2022 using the online Wikimedia Statistics service (https://stats.wikimedia.org).

Wikimedia Statistics

The Case of Wikipedia

- We contribute our implementation of the Wikimedia Statistics service as a Python package and Command Line Interface. **Wikistats-to-CSV** (`wikistats2csv`): https://github.com/SaiedAlshahrani/Wikistats-to-CSV.



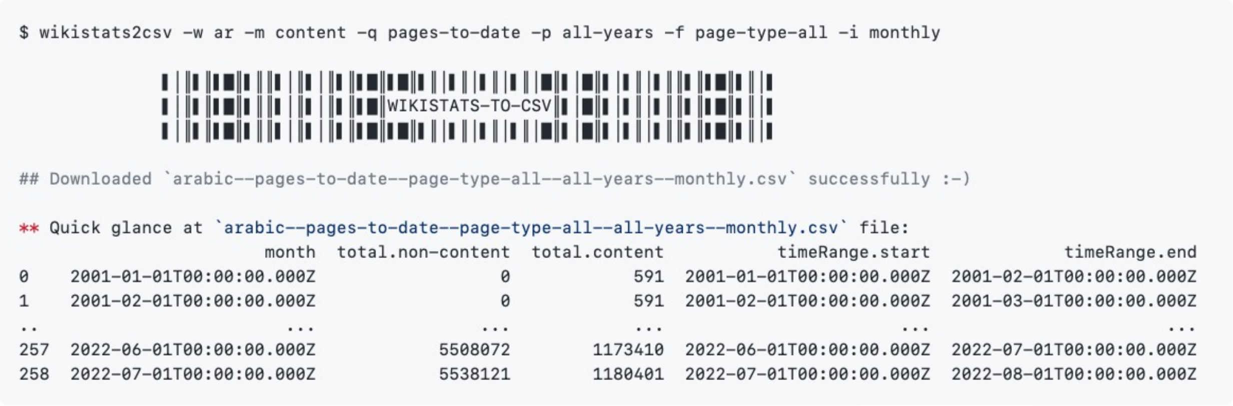
```
>>> from wikistats2csv import Content
>>> Content.pages_to_date(wiki='es', period='all-years', filter='page-type-all', interval='monthly')

## Downloaded `spanish--pages-to-date--page-type-all--all-years--monthly.csv` successfully :-)

** Quick glance at `spanish--pages-to-date--page-type-all--all-years--monthly.csv` file:
                        month  total.non-content  total.content           timeRange.start             timeRange.end
0    2001-01-01T00:00:00.000Z                  0              0  2001-01-01T00:00:00.000Z  2001-02-01T00:00:00.000Z
1    2001-02-01T00:00:00.000Z                  0              0  2001-02-01T00:00:00.000Z  2001-03-01T00:00:00.000Z
..                        ...                ...            ...                       ...                       ...
257  2022-06-01T00:00:00.000Z            3896209        1786321  2022-06-01T00:00:00.000Z  2022-07-01T00:00:00.000Z
258  2022-07-01T00:00:00.000Z            3903963        1792329  2022-07-01T00:00:00.000Z  2022-08-01T00:00:00.000Z
```

The Case of Wikipedia

- We contribute our implementation of the Wikimedia Statistics service as a Python package and Command Line Interface. **Wikistats-to-CSV** (`wikistats2csv`): https://github.com/SaiedAlshahrani/Wikistats-to-CSV.



```
$ wikistats2csv -w ar -m content -q pages-to-date -p all-years -f page-type-all -i monthly

                    WIKISTATS-TO-CSV

## Downloaded `arabic--pages-to-date--page-type-all--all-years--monthly.csv` successfully :-)

** Quick glance at `arabic--pages-to-date--page-type-all--all-years--monthly.csv` file:
                        month  total.non-content  total.content            timeRange.start              timeRange.end
0    2001-01-01T00:00:00.000Z                  0            591  2001-01-01T00:00:00.000Z  2001-02-01T00:00:00.000Z
1    2001-02-01T00:00:00.000Z                  0            591  2001-02-01T00:00:00.000Z  2001-03-01T00:00:00.000Z
..                        ...                ...            ...                        ...                        ...
257  2022-06-01T00:00:00.000Z            5508072        1173410  2022-06-01T00:00:00.000Z  2022-07-01T00:00:00.000Z
258  2022-07-01T00:00:00.000Z            5538121        1180401  2022-07-01T00:00:00.000Z  2022-08-01T00:00:00.000Z
```

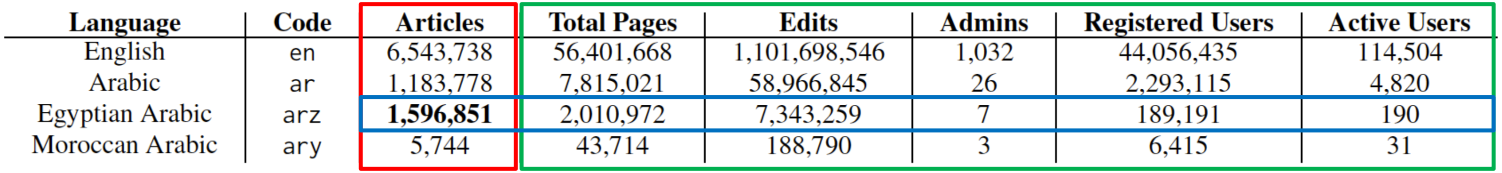
The Case of Wikipedia

- We study the impact of problems in Egyptian Arabic Wikipedia, including large-scale automatic generation and poor translation of content pages from English.
  - Interestingly, over **63%** of the total content pages in Egyptian Wikipedia have been automatically created using template-based translation from English.

The Case of Wikipedia

Arabic Wikipedia Editions

- Wikipedia encyclopedia was launched **20** years ago (in 2001) and was released primarily in English.
- The Arabic, Egyptian, and Moroccan editions appeared on the project in 2004, 2008, and 2019, respectively.



Language	Code	Articles	Total Pages	Edits	Admins	Registered Users	Active Users
English	en	6,543,738	56,401,668	1,101,698,546	1,032	44,056,435	114,504
Arabic	ar	1,183,778	7,815,021	58,966,845	26	2,293,115	4,820
Egyptian Arabic	arz	**1,596,851**	2,010,972	7,343,259	7	189,191	190
Moroccan Arabic	ary	5,744	43,714	188,790	3	6,415	31

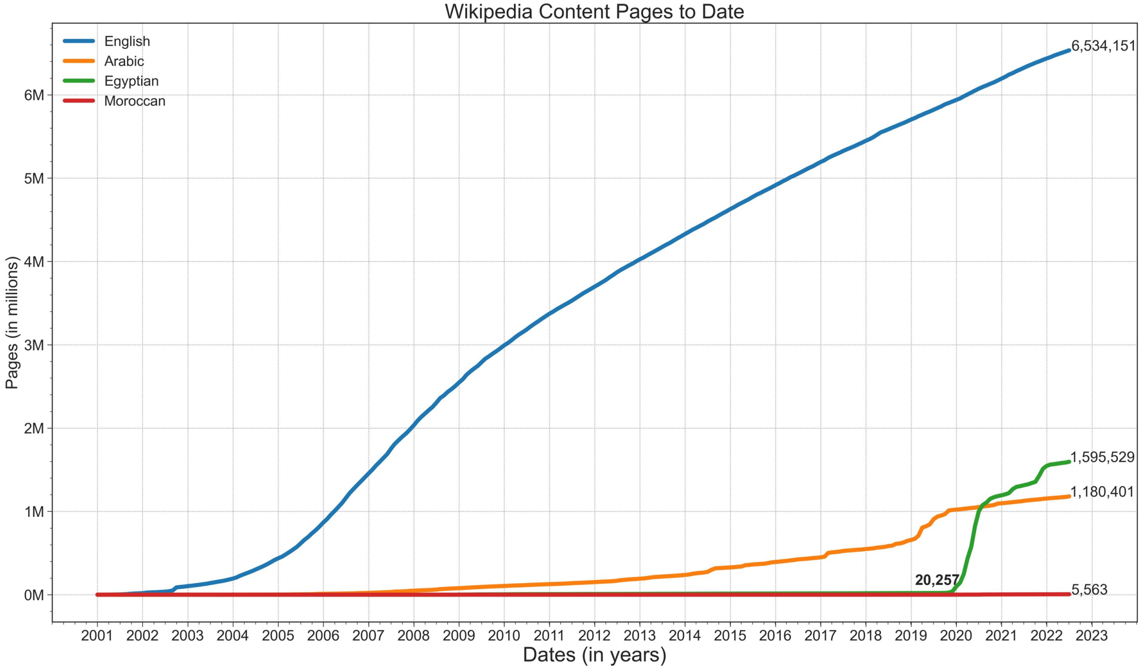
The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date

- English, Arabic, and Moroccan Arabic show **normal** growth in their content pages (articles) over the timeline of Wikipedia.
- The content of Egyptian Arabic has recently grown rapidly and exponentially in the last **two** years.

The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date



The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date

Wikipedia	Articles	Period	Monthly	Daily
Arabic	~1.2 million	19 years	~5,000	~200
Egyptian	~1.6 million	Less than 3 years	~50,000	~2,000

The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date

- This exponential growth of the content pages in the Egyptian Arabic Wikipedia in only **30** months is the result of the large-scale automated creation of the content pages.

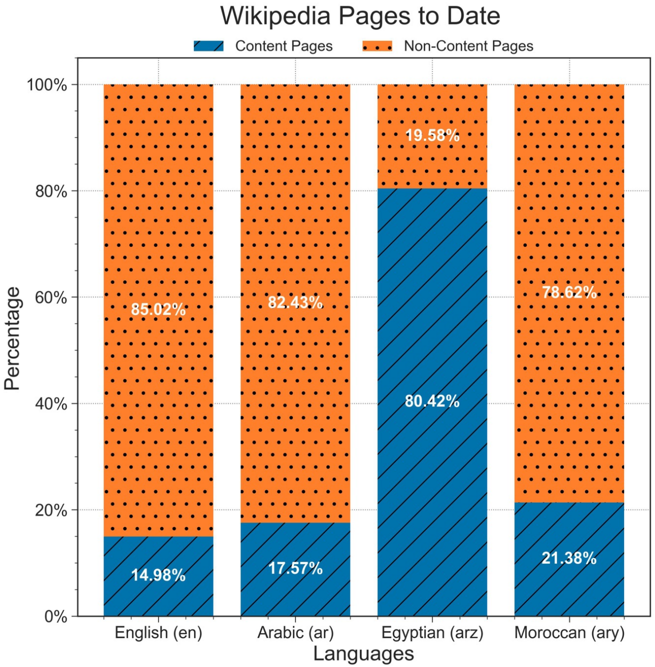
The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date

- We visualize the percentage of all page types to date for the **four** Wikipedia editions, displaying the difference in percentage between page types to study the characteristics of each Wikipedia within itself.

The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date



The Case of Wikipedia

Arabic Wikipedia Editions: Pages to Date

- Egyptian Arabic Wikipedia has approximately **20%** of non-content pages and **80%** of content pages, which is a consequence of the large-scale automation of content creation.

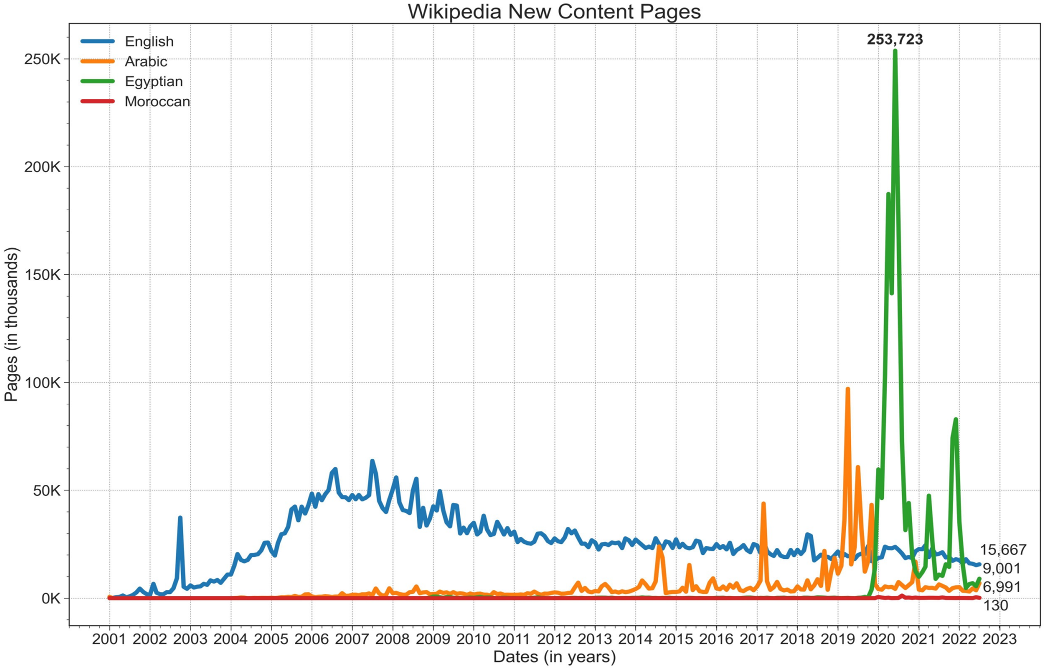
The Case of Wikipedia

Arabic Wikipedia Editions: New Pages

- In June 2020, approximately **253,000** new content pages were created in the Egyptian Arabic Wikipedia.
- Nearly 23,700 new content pages were created on English Wikipedia, nearly 4,280 on Arabic Wikipedia, and nearly 50 on Moroccan Wikipedia, all in June 2020.

The Case of Wikipedia

Arabic Wikipedia Editions: New Pages



The Case of Wikipedia

Arabic Wikipedia Editions: Top Editors

- Wikipedia has **four** types of editors:
  - Registered Users:
    - logged-in users but not in group-bot nor name-bot sets.
  - Group-bots:
    - logged-in users who are part of a bot group.
  - Name-bots:
    - logged-in users whose name contains ‘bot’.
  - Anonymous Users:
    - unlogged-in users (tracked by IPs or devices fingerprints).

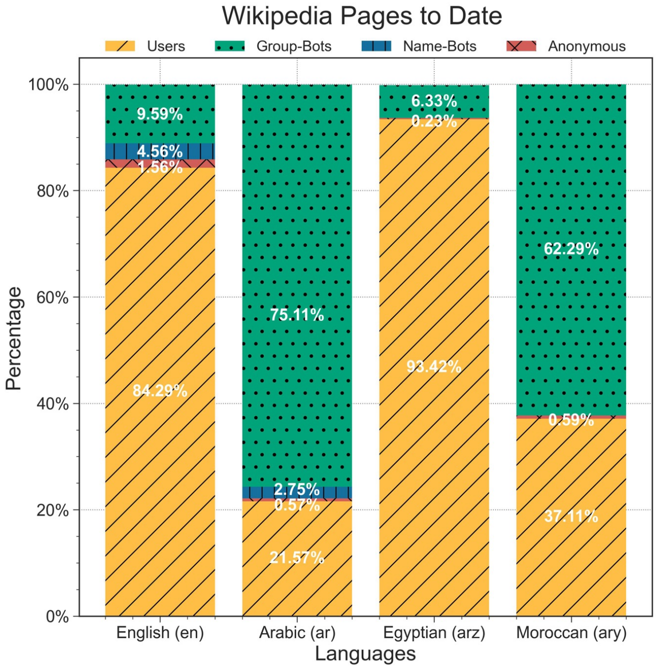
The Case of Wikipedia

Arabic Wikipedia Editions: Top Editors

- We visualize the percentage of all pages to date for the four Wikipedia editions by displaying the difference in percentage between **editor types** to study the characteristics of each Wikipedia within itself.

The Case of Wikipedia

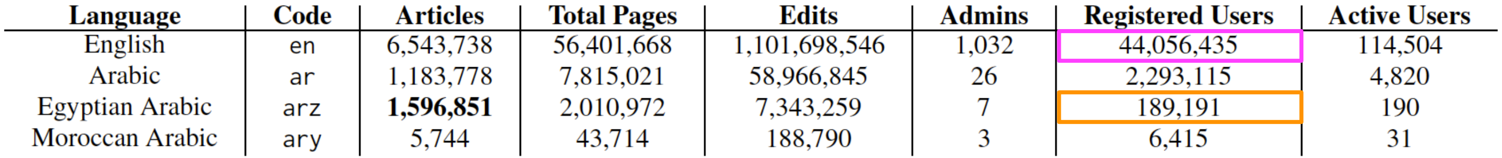
Arabic Wikipedia Editions: Top Editors



The Case of Wikipedia

Arabic Wikipedia Editions: Top Editors

- We hypothesize that the reason behind the 84% of pages created by registered users in English Wikipedia is the huge gap in the total number of registered users (~**233X** bigger than Egyptian Wikipedia).



Language	Code	Articles	Total Pages	Edits	Admins	Registered Users	Active Users
English	en	6,543,738	56,401,668	1,101,698,546	1,032	44,056,435	114,504
Arabic	ar	1,183,778	7,815,021	58,966,845	26	2,293,115	4,820
Egyptian Arabic	arz	**1,596,851**	2,010,972	7,343,259	7	189,191	190
Moroccan Arabic	ary	5,744	43,714	188,790	3	6,415	31

The Case of Wikipedia

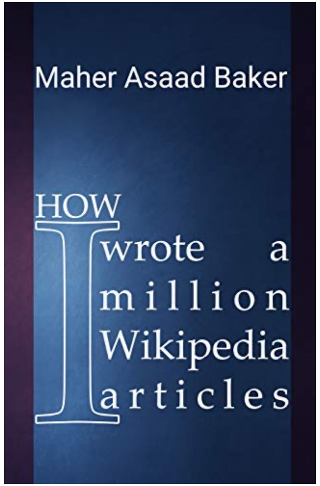
Egyptian Wikipedia Problems

- We found that over **one** million articles (63% of the total articles) in Egyptian Arabic Wikipedia have been created by one registered user called "HitomiAkane".
- This user has made more than 1,562,615 new creations, made nearly 1,615,216 edits, and created over **1** million auto-generated articles without human revision [11].

The Case of Wikipedia

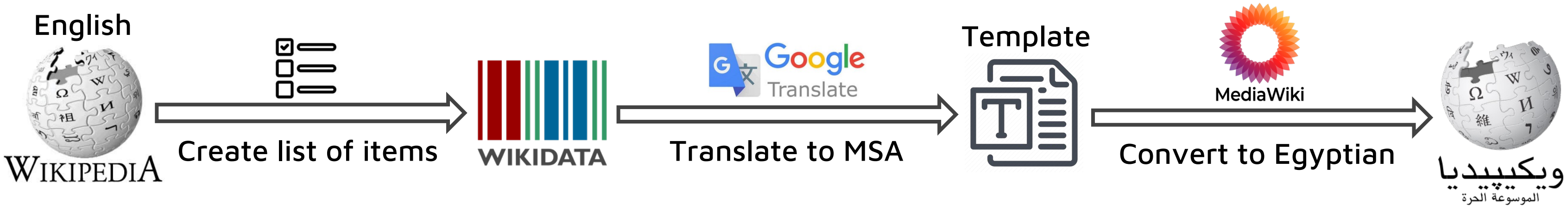
Egyptian Wikipedia Problems

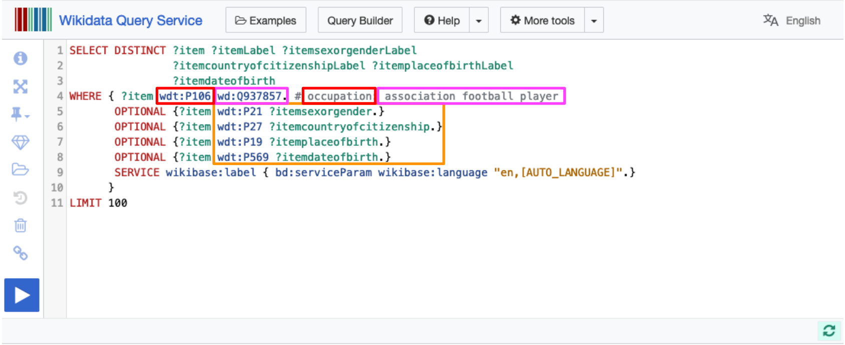
- This user described the large-scale content creation process using template-based translation from English in his book: **How I Wrote a Million Wikipedia Articles** [10].



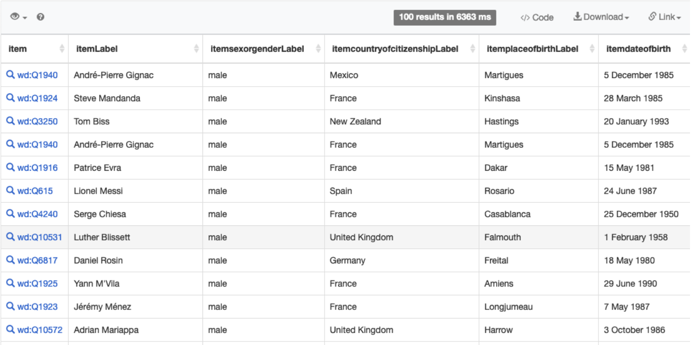
The Case of Wikipedia

Egyptian Wikipedia Problems





```
SELECT DISTINCT ?item ?itemLabel ?itemsexorgenderLabel
                ?itemcountryofcitizenshipLabel ?itemplaceofbirthLabel
                ?itemdateofbirth
WHERE { ?item wdt:P106 wd:Q937857. # occupation  association football player
       OPTIONAL {?item wdt:P21 ?itemsexorgender.}
       OPTIONAL {?item wdt:P27 ?itemcountryofcitizenship.}
       OPTIONAL {?item wdt:P19 ?itemplaceofbirth.}
       OPTIONAL {?item wdt:P569 ?itemdateofbirth.}
       SERVICE wikibase:label { bd:serviceParam wikibase:language "en,[AUTO_LANGUAGE]".}
      }
LIMIT 100
```

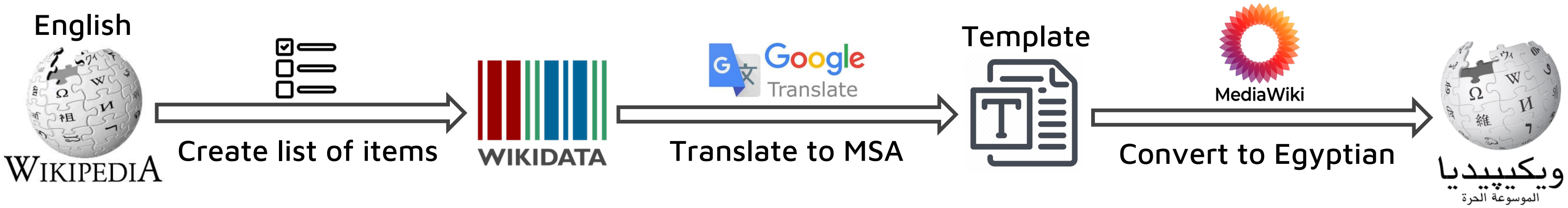


100 results in 6363 ms

item	itemLabel	itemsexorgenderLabel	itemcountryofcitizenshipLabel	itemplaceofbirthLabel	itemdateofbirth
wd:Q1940	André-Pierre Gignac	male	Mexico	Martigues	5 December 1985
wd:Q1924	Steve Mandanda	male	France	Kinshasa	28 March 1985
wd:Q3250	Tom Biss	male	New Zealand	Hastings	20 January 1993
wd:Q1940	André-Pierre Gignac	male	France	Martigues	5 December 1985
wd:Q1916	Patrice Evra	male	France	Dakar	15 May 1981
wd:Q615	Lionel Messi	male	Spain	Rosario	24 June 1987
wd:Q4240	Serge Chiesa	male	France	Casablanca	25 December 1950
wd:Q10531	Luther Blissett	male	United Kingdom	Falmouth	1 February 1958
wd:Q6817	Daniel Rosin	male	Germany	Freital	18 May 1980
wd:Q1925	Yann M'Vila	male	France	Amiens	29 June 1990
wd:Q1923	Jérémy Ménez	male	France	Longjumeau	7 May 1987
wd:Q10572	Adrian Mariappa	male	United Kingdom	Harrow	3 October 1986

The Case of Wikipedia

Egyptian Wikipedia Problems



- We quote the example of football player that he used to demonstrate the automation process in the book:

[label] [date of birth], **[gender]** is a football player from **[citizenship]**, **[gender]** was born at **[date of birth]** in **[place of birth]**.

The Case of Wikipedia

Egyptian Wikipedia Problems

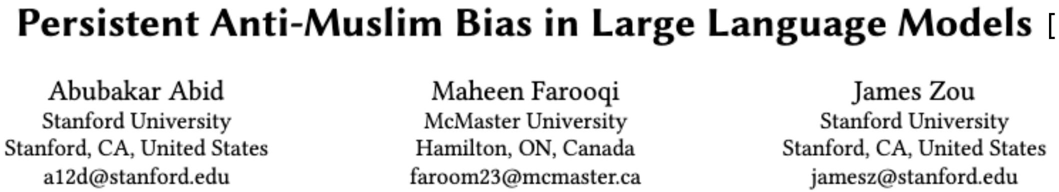
- The process of converting the translated MSA articles to the Egyptian Arabic dialect remains **mysterious**.
  - We hypothesize that the user used a lexicon-based conversion between MSA and Egyptian to make it look like it was produced organically by native speakers.

- Overall, the used process represents a relatively **shallow**, template-based content translation.

The Case of Wikipedia

Egyptian Wikipedia Problems

- Wikipedia articles should **only** be written, contributed to, edited, and maintained by the people.
- This lack of representativeness and cultural richness **holds** many potential problems that could negatively impact society when deploying AI systems or NLP tools trained on such unrepresentative corpora [12].



Persistent Anti-Muslim Bias in Large Language Models [22]

Abubakar Abid
Stanford University
Stanford, CA, United States
a12d@stanford.edu

Maheen Farooqi
McMaster University
Hamilton, ON, Canada
faroom23@mcmaster.ca

James Zou
Stanford University
Stanford, CA, United States
jamesz@stanford.edu

Discussion

- Arabic poses many **challenges** in NLP that prevent simply translating from another language due to its morphological richness and high ambiguity [13, 14].
  - Arabic verbs have ~5,400 inflected forms, English verbs have 6, and Chinese verbs have only one [15].
  - Diacritical marks vs. multiple variants of a word [5].

Discussion

- Arabic has **many** dialectal variants, like Egyptian and Moroccan Arabic, that are different from MSA, which are primarily spoken, do not have orthographic rules, and have few resources [16, 17].

Discussion

- Despite all these challenges, translating from English to enrich low-resource languages' content like Arabic is a **common** practice, mainly done using Machine Translation models (MTs) [18].

Discussion

Representativeness

- The heart of the lack of representativeness problem, specifically in the Arabic language, can be discussed from **two** different perspectives:
    1. The template-based large-scale auto-generation of content, especially in the Wikipedia encyclopedia.
    2. The translation of content from English to Arabic using direct translation methods or off-the-shelf tools.

Discussion

Representativeness: Generation

- We have documented that the Egyptian Arabic Wikipedia content **does not** genuinely represent the Egyptian people, their culture, traditions, or views.

Discussion

Representativeness: Generation

- We, humans, express our values, traditions, or perspectives through writing, and when using auto-generated text corpora, we **miss** the people's culture, sentiments, stances, opinions, etc.

Discussion

Representativeness: Generation

- Even though Wikipedia articles are factual entries, the choice of topics, choice of facts, and choice of words **reflect** the perspective and values of the authors.
- It **matters** whether these choices are made by native speakers or by translation from other languages.

Discussion

Representativeness: Translation

- The other face of the lack of representativeness problem is **translating** the content of the English language to other low-resource languages like Arabic using direct translation or off-the-shelf tools.

Discussion

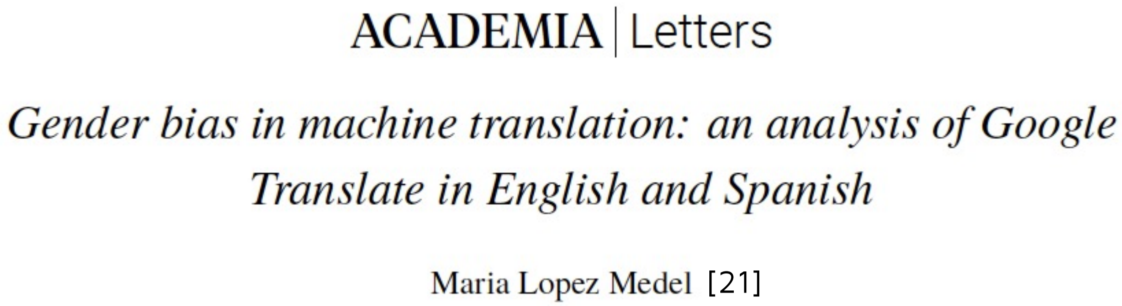
Representativeness: Translation

- Wikimedia Foundation has encouraged users and contributors to use MTs to translate and create the initial content of more than **400,000** articles on Wikipedia project using Google Translate [19].

Discussion

Representativeness: Translation

- Yet, we **should** consider the quality of these translation tools, the quality of the translation work conducted by non-expert Wikipedia users, and what they could bring to the content of Wikipedia from potential issues like biases and sexism [20, 21].



ACADEMIA | Letters

Gender bias in machine translation: an analysis of Google Translate in English and Spanish

Maria Lopez Medel [21]

Discussion

Final Statement

- NLP corpora that are automatically created using shallow templated-based, poorly translated using direct translation, auto-generated by advanced LLMs, 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.

Conclusion

- We studied the three Arabic Wikipedia editions besides English Wikipedia and shed light on the problem of the Egyptian Arabic Wikipedia.
    - We found that **one** registered user has automated the creation of over **1** million articles in less than three years and used a shallow, template-based translation method.

Recommendation

- We recommend that NLP practitioners **avoid** the inorganic, unauthentic, unrepresentative corpora in their applications when the goal is to learn from past human behavior and to investigate how the corpora they do use were created, generated, or assembled.
- We recommend that when registered users employ automated translation processes, their contributions should be marked differently than "registered user"; perhaps "**registered user (automation-assisted)**".

Future Works

- We plan to **construct** a representative analysis of the Arabic Wikipedia editions in terms of corpus and similarity metrics (*work in progress*).
- We plan to study the **implications** of using such unrepresentative corpora that are naively auto-created, poorly translated, or automatically generated on the downstream applications of the NLP.

References

[1] Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. 2020. Is Machine Learning Speaking my Language? A Critical Look at the NLP-Pipeline Across 8 Human Languages. arXiv preprint arXiv:2007.05872.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Advances in neural information processing systems, 29.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.

[4] Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying Gender Bias in Different Corpora. In Companion Proceedings of the Web Conference 2020,WWW'20, page 752–759, New York, NY, USA. Association for Computing Machinery.

[5] Saied Alshahrani, Esma Wali, Abdullah R Alshamsan, Yan Chen, and Jeanna Matthews. 2022. Roadblocks in Gender Bias Measurement for Diachronic Corpora. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 140–148, Dublin, Ireland. Association for Computational Linguistics.

[6] Sergiu Nisioi, Ella Rabinovich, Liviu P. Dinu, and Shuly Wintner. 2016. A Corpus of Native, Non-native and Translated Texts. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4197–4201, Portorož, Slovenia. European Language Resources Association (ELRA).

[7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[9] 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, ClemensWinter, 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. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

References

[10] Maher Asaad Baker. 2022. How I Wrote a Million Wikipedia Articles. 2 Edition. BookRix GmbH Co. KG., Munich, Germany.

[11] Wikimedia Foundation. 2022b. Wikimedia Statistics. Last accessed on 2022-09-11.

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

[13] Khaled Shaalan, Sanjeera Siddiqui, Manar Alkhatib, and Azza Monem. 2018. Challenges in Arabic Natural Language Processing. World Scientific.

[14] Ali Farghaly and Khaled Shaalan. 2009. Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing, 8(4).

[15] Nizar Habash. 2020. A Short Introduction to Arabic Natural Language Processing. Women to Impact, King Abdullah University of Science and Technology, Saudi Arabia.

[16] Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 426–432, Atlanta, Georgia. Association for Computational Linguistics.

[17] Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov, and Stephan Vogel. 2014. Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pages 207–216, Doha, Qatar. Association for Computational Linguistics.

[18] Ahmed El-Kholy and Nizar Habash. 2010. Orthographic and morphological processing for English–Arabic statistical machine translation. Machine Translation, 26:25–45.

[19] Runa Bhattacharjee and Pau Giner. 2022. You can now use Google Translate to translate articles on Wikipedia. Last accessed on 2022-09-11.

[20] Stefanie Ullmann and Danielle Saunders. 2021. Google Translate is sexist. What it needs is a little gender sensitivity training. Last accessed on 2022-09-11.

[21] Maria Lopez-Medel. 2021. Gender bias in machine translation: an analysis of Google Translate in English and Spanish. Academia.edu.

[22] Abid, A., Farooqi, M., and Zou, J., 2021. Persistent Anti-Muslim Bias in Large Language Models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 298-306).

[23] Alyafeai, Z., Masoud, M., Ghaleb, M. and Al-shaibani, M.S., 2021. Masader+: Metadata sourcing for Arabic text and speech data resources. arXiv preprint arXiv:2110.06744.

Visit the Project:

https://webspace.clarkson.edu/~alshahsf/Representativeness.html

SCAN ME!