

Faithful NLG in an era of Ethical Awareness: Opportunities for MT

Mona Diab

Facebook AI

George Washington University

Disclaimers

- All source examples are reflective of true use cases but with alteration for privacy and anonymity purposes
- All MT output is authentic from various industrial MT systems
- There will be some profanities and racist language on some of the slides used to illustrate real phenomena encountered in social media

MT's Mission

- MT mission is to transcend language barriers to Facilitate communication and empower people to have unimpeded knowledge/information access
- MT is a utility serving peoples in their daily lives
- MT plays a pivotal role in peoples' lives especially in an era of heightened ethical awareness

Overarching Proposal

Adopt faithfulness as an objective for MT
(*a la* other NLG technologies, eg. summarization)

- Entails creating evaluation metrics that optimize for faithfulness
 - Entails building faithfulness aware models and data sets

What is faithfulness in MT

At a High Level

Generating translations that are “exactly”
equivalent to the source

Other NLG Technologies

Faithful Abstractive Summarization Evaluation

- Task is to produce summaries based on a single or multiple documents in a collection of documents
- Faithful summaries are expected to be reflective of the source and only the source with no hallucination
- Typically in a single language

Mind you it doesn't say much about coverage of all the information in the source as long as it doesn't insert unwarranted information – *different from MT*

An Era of Ethical Awareness

- Equality (Definition inspired by Responsible AI, FB)
 - Catering to all groups and all languages (including dialects and vernaculars) with the same level of service, thereby treat people equally and minimize unjustified differences in outcome
- Equity (Definition inspired by Responsible AI, FB)
 - Understanding whether there are groups that deserve special consideration and make a deliberate decision to prioritize outcomes for the disenfranchised
- Avoid harm and minimize bias
- Minimize misrepresentation going beyond basic equity and equality to address quality
- Increase Transparency
 - Clearly Communicate quality reflecting our confidence in the generated text

Faithfulness

Faithfulness serves ethical considerations better than current common MT mindset optimizing for Adequacy and Fluency only

State of the art MT Today

- **Source: Arabic (MSA)**

هذا البرنامج في أمريكا يجمع معلمين لغة انجليزية محليين مع معلمين من حول العالم لسته اسابيع لتدريبهم على احدث طرق تدريس اللغة الانجليزية.

- **Machine Translation**

This program in America brings together local English language teachers with teachers from around the world for six weeks to train them in the latest methods of teaching English.

State of the art MT Today

- **Source: Arabic (MSA)**

هذا البرنامج في أمريكا يجمع معلمين لغة انجليزية محليين مع معلمين من حول العالم لسته اسابيع لتدريبهم على احدث طرق تدريس اللغة الانجليزية.

- **Machine Translation**

This program in America brings together local English language teachers with teachers from around the world for six weeks to train them in the latest methods of teaching English.

Excellent Translation

State of the art MT Today

- Source: Arabic (MSA)

هذا البرنامج في أمريكا يجمع معلمين لغة انجليزية محليين مع معلمين من حول العالم لسته اسابيع لتدريبهم على احدث طرق تدريس اللغة الانجليزية.

- Machine Translation

This program in America gathers local English teachers with teachers from around the world for six weeks to train them in the latest machine translation technology. **Backtranslation is actually better Arabic than source 😊**

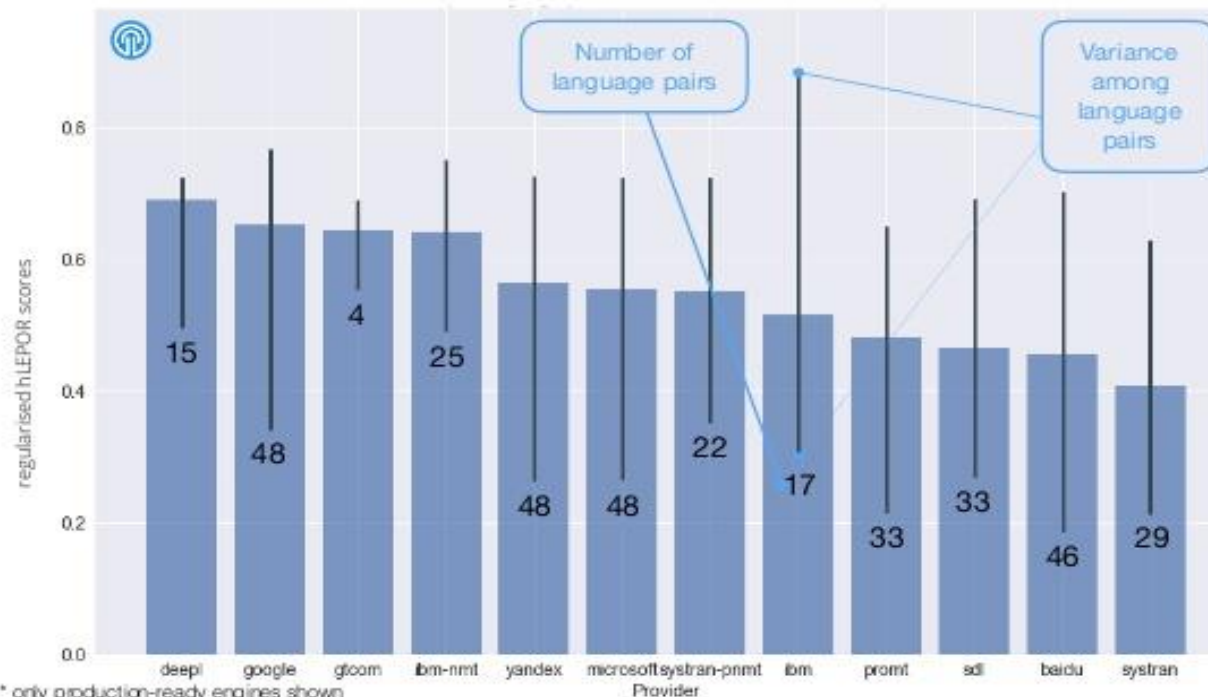
Back translation

يجمع هذا البرنامج في أمريكا بين معلمي اللغة الإنجليزية المحليين ومعلمين من جميع أنحاء العالم لمدة ستة أسابيع لتدريبهم على أحدث طرق تدريس اللغة الإنجليزية.

State of the art commercial MT (03/2018) – Intento, Inc.

Overall Performance

48 language pairs, 900-3000 sentences per pair*



* only production-ready engines shown

© Intento, Inc.

Detailed data on each language pair provided in the full report

Notes

1. Data mostly edited text including WMT data collections to date
2. Metric is human LEPOR which is a variant of BLEU with length and brevity penalties

State of the art commercial MT (03/2018) – Intento, Inc.

Available
MT
Quality



Detailed data on
each language
pair provided in
the full report

* base pricing tier
** up to 5% worse than the leader

- In general, XX-> EN achieved higher performance than EN-> XX (except JA ->EN, TR-> EN)
- EN-> XX much lower for low resource languages CS, TR, FI, KO, AR
- Non EN directions (16/18) show much poorer performance
- Out of 12 MT systems considered, only a max of 6 systems are considered competitive (FR->ES). The majority of the directions only had 2 systems that are competitive (22/48)

State of the art MT Today

- Source

هذا البرنامج في أمريكا يجمع معلمين لغة انجليزية محليين مع معلمين من حول العالم لسته اسابيع لتدريبهم على احدث طرق تدريس اللغة الانجليزية.

- Machine Translation

This program in America gathers local language teachers with teachers from around the world for six weeks to train them in the latest methods of teaching English.

This is edited (well formed)
Modern Standard Arabic data

Back translation

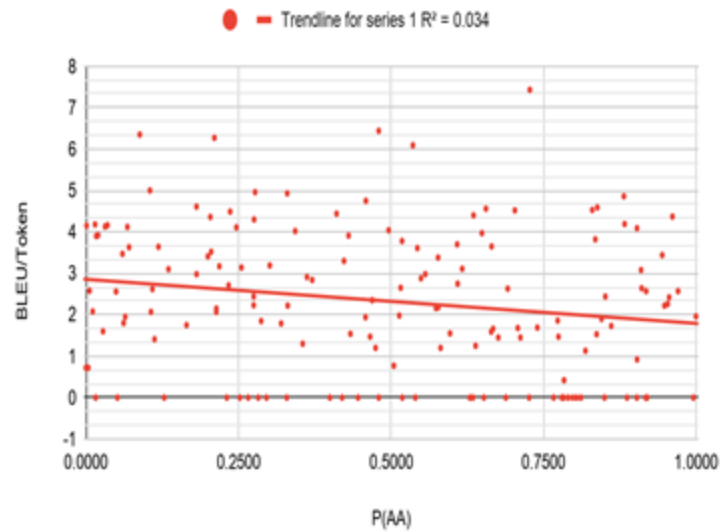
يجمع هذا البرنامج في أمريكا بين معلمي اللغة الإنجليزية المحليين ومعلمين من جميع أنحاء العالم لمدة ستة أسابيع لتدريبهم على أحدث طرق تدريس اللغة الإنجليزية.

MT State of the art on social media

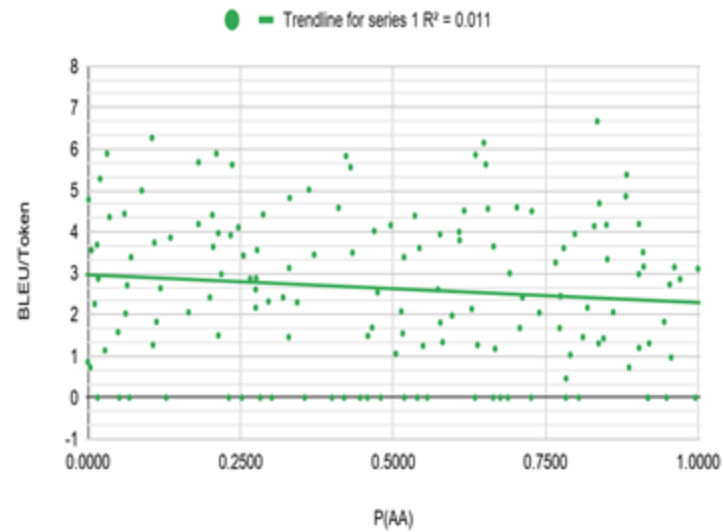
Direction	Source	MT	HT
AR→EN	قدموا هون من خلال الرابط	They provided here through the link	Apply here through the link
EN→FR	Don't forget to hit me up.	N'oublie pas à me frapper . BT: Don't hesitate to hit me.	N'oublie pas à me contacter .
TR→EN	30 derecede sıkmadan ve bastırmadan yıkanabilir	30 degrees can be washed without pressure and fucking	washable at 30 degrees without wringing or pressing
EN→AR	Super relate. Silent treatment to the max	سوبر تتصل . معاملة صامتة إلى أقصى الحدود BT: Super is calling. Silent treatment to the extreme	أنا معاك بالظبط . أتجاهل لأقصى الحدود
AR→EN	سوفجارديت الفيشي	Souvardite Vichy	Saved the file
EN→AR	Vote him out!	صوت لصالحه BT: Vote for him	خرجوه بالانتخابات
AR→EN	مشيت الكلبة بتاعتي في الشارع	My bitch walked down the street	Walked my dog in the street

Translations from vernaculars is bad quality

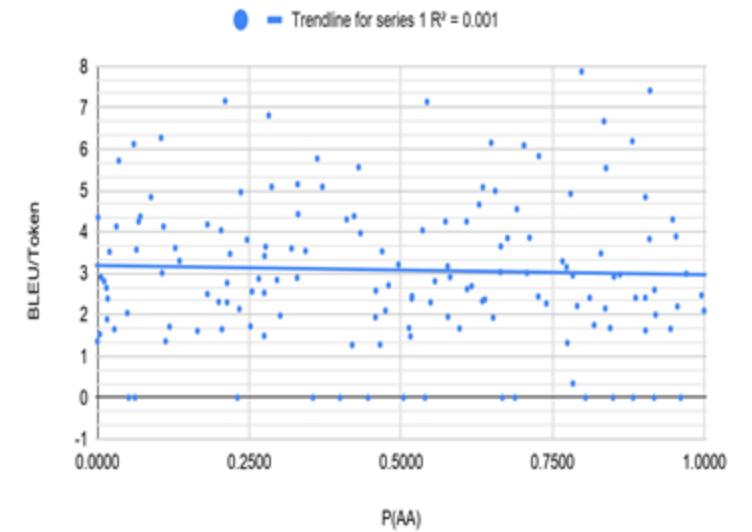
System A: AAE



System B: AAE

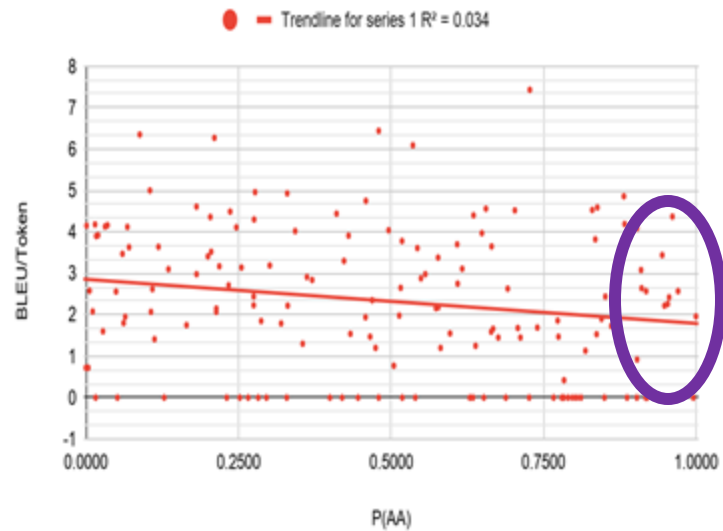


System C: AAE

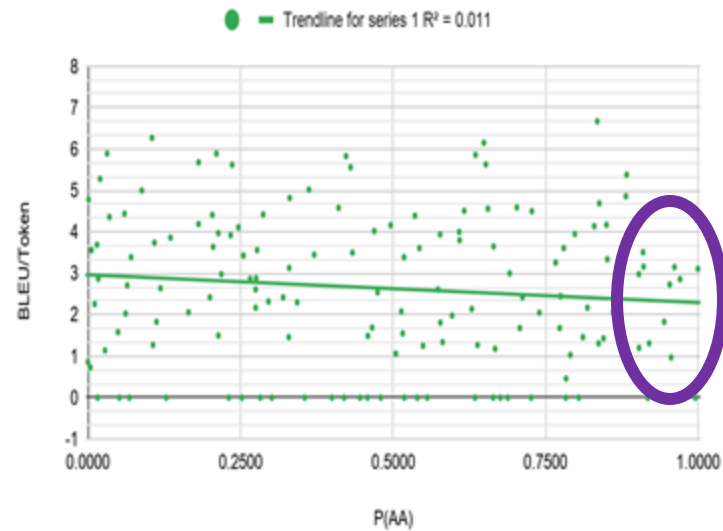


Translations from vernaculars is bad quality

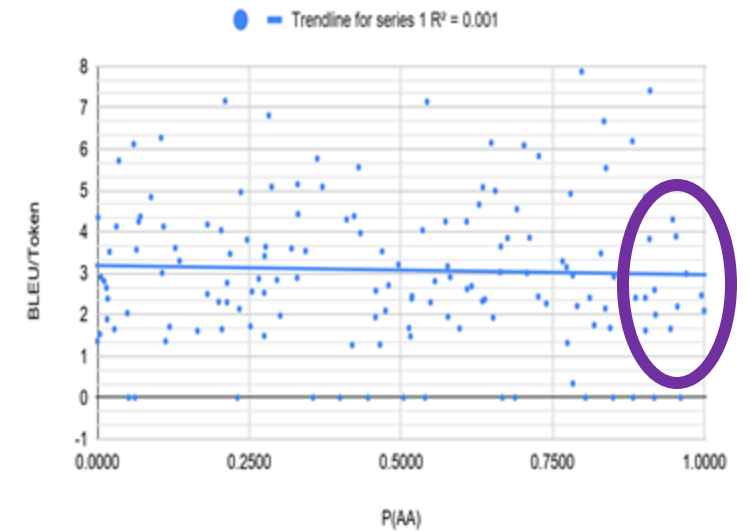
System A: AAE



System B: AAE



System C: AAE

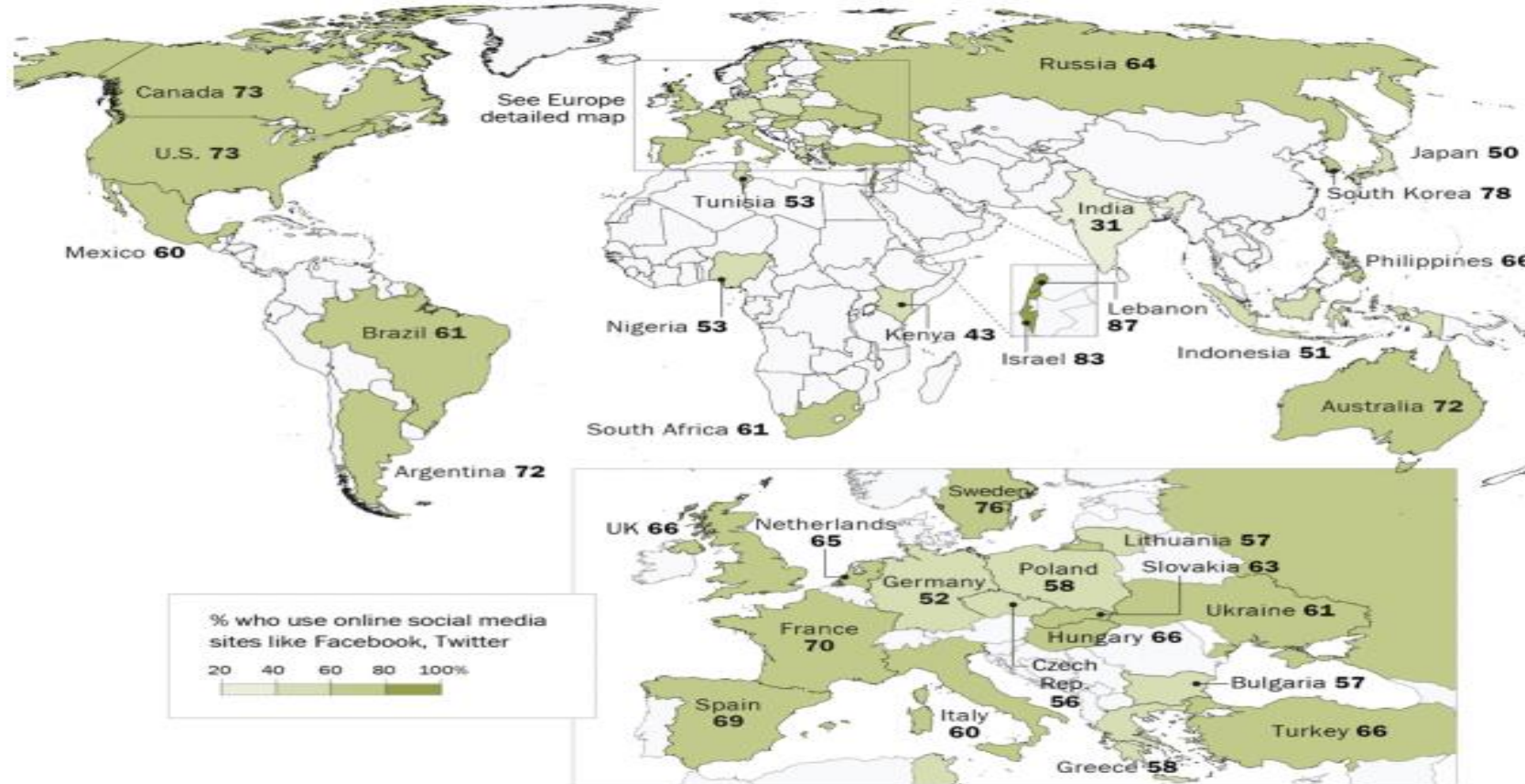


Translation evaluation using top off the shelf MT systems for African American English (AAE) Vernacular to French. BLEU vs. % of AAE present in the source. Downward slope indicates worsening performance

Why should we care

A majority in many countries use some form of social media

% who use online social media sites like Facebook, Twitter or other country-specific forms of social media



Source: Spring 2019 Global Attitudes Survey, Q54.

PEW RESEARCH CENTER

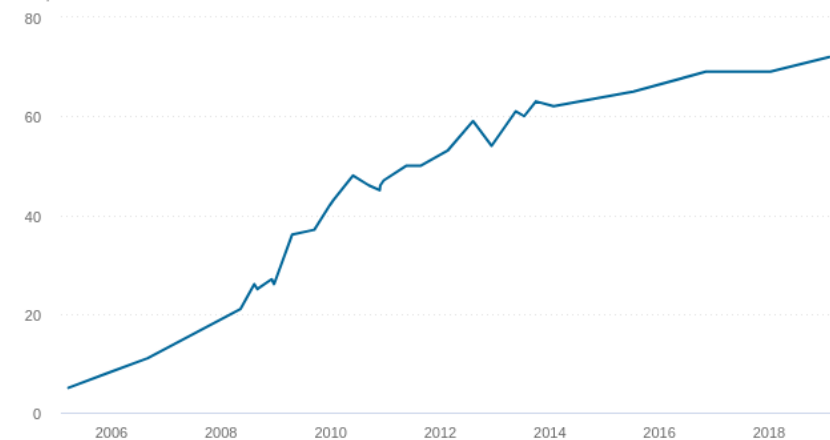
- 62% of adults in USA get their news from social media according to study by Pew Foundation ([Wikipedia](#))
- Number of users who get their news from social media platforms in USA 34% in 2018 up from 28% in 2016 ([Pew Research](#)) but globally 77% of internet users rely on social media for their news

https://www.pewresearch.org/fact-tank/2020/04/02/8-charts-on-internet-use-around-the-world-as-countries-grapple-with-covid-19/ft_2020-04-02_globalinternet_07/

Why should we care

Social media use

% of U.S. adults who use at least one social media site



Source: Surveys conducted 2005-2019.

<https://www.pewresearch.org/internet/fact-sheet/social-media/>

July 2020 Worldwide Users

Total population: **7.79B**

Unique Mobile Phone Users: **5.15B**

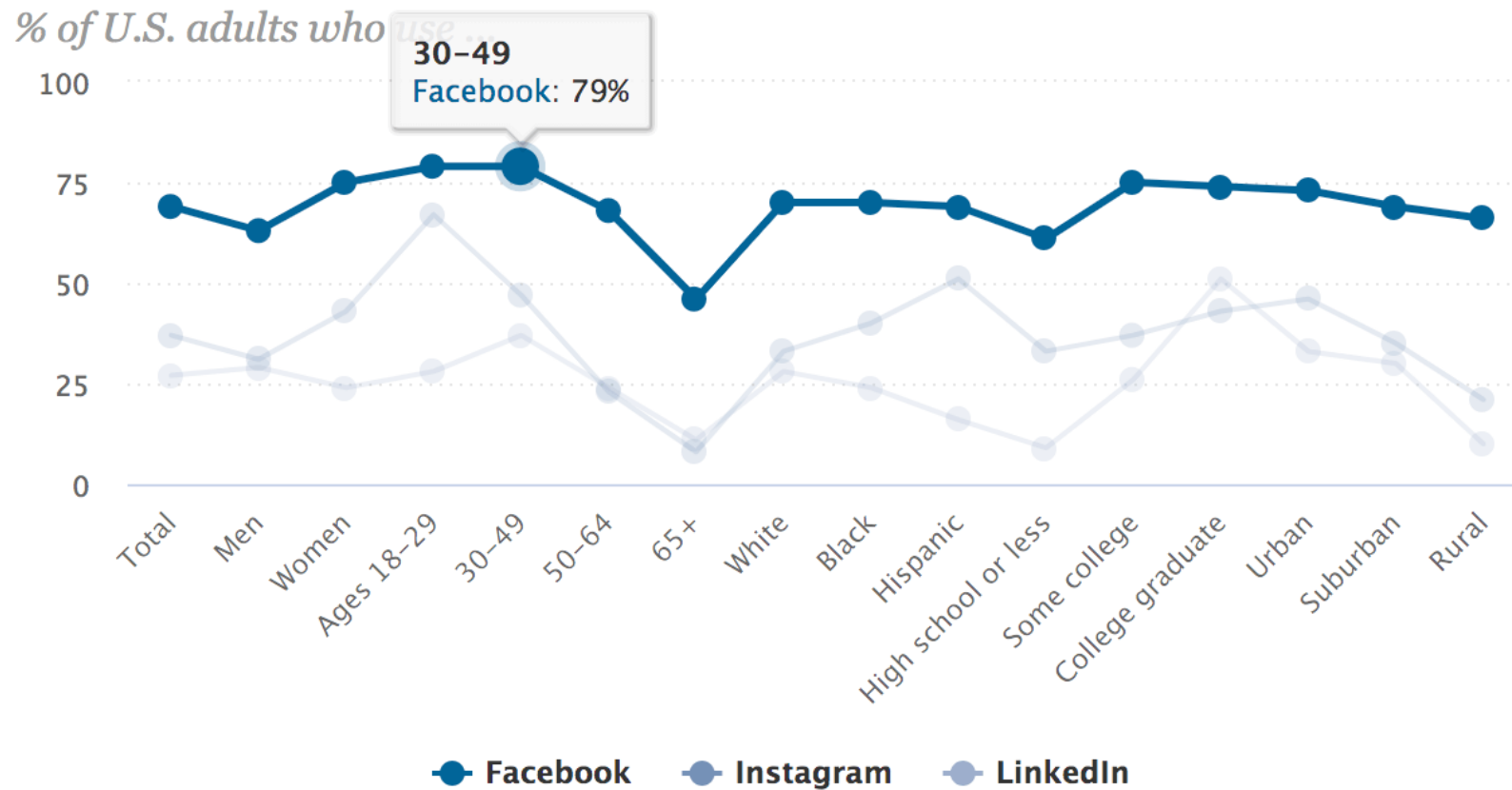
internet users worldwide: **4.57B**

Active social media users: **3.96B**

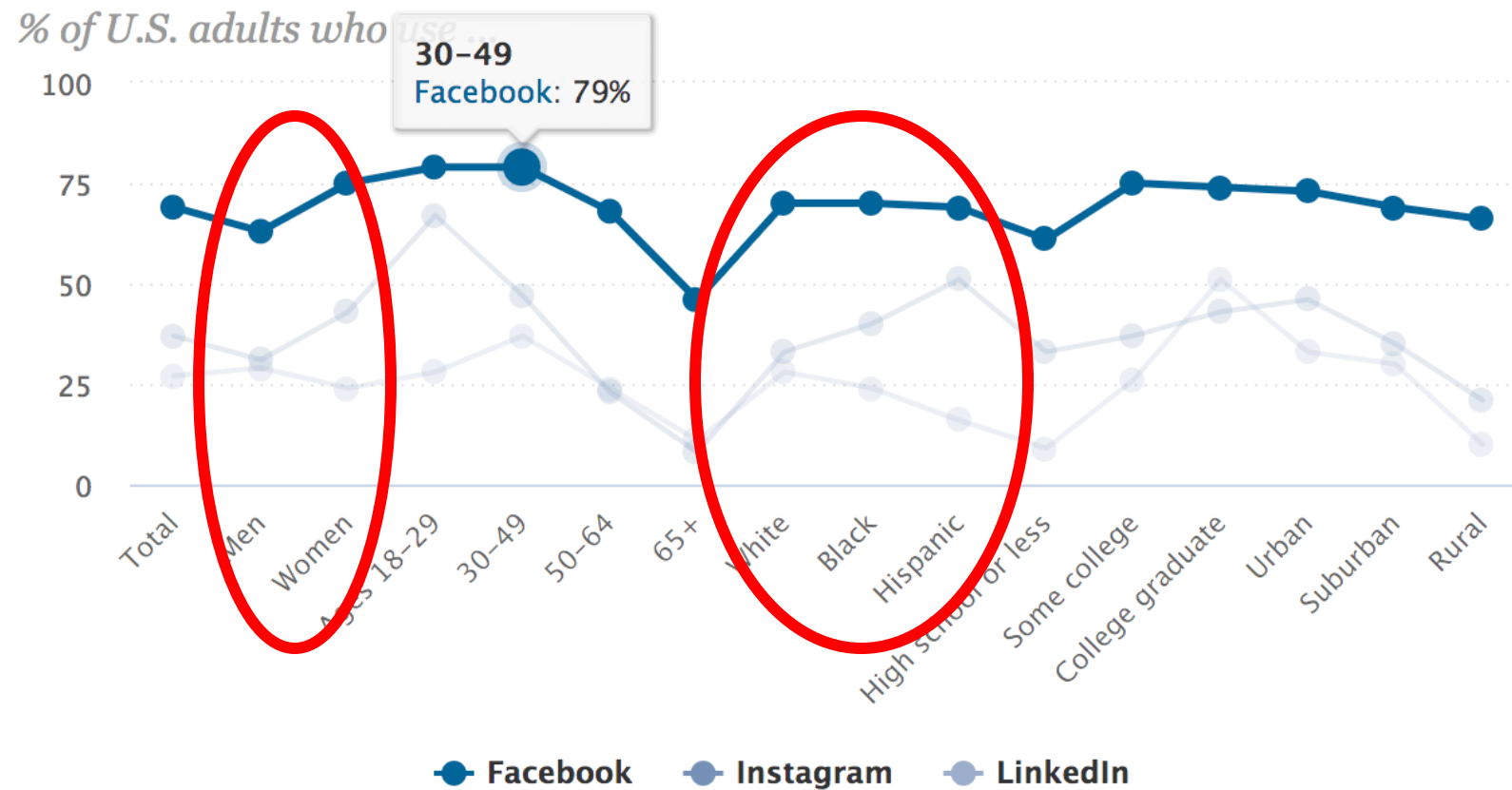
(English/Non English)	%
East Asia	63%
North America	69%
South America	68%
Northern Europe	66%
Western Asia	56%
North Africa	40%
Middle Africa	7%

<https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>

Why should we care



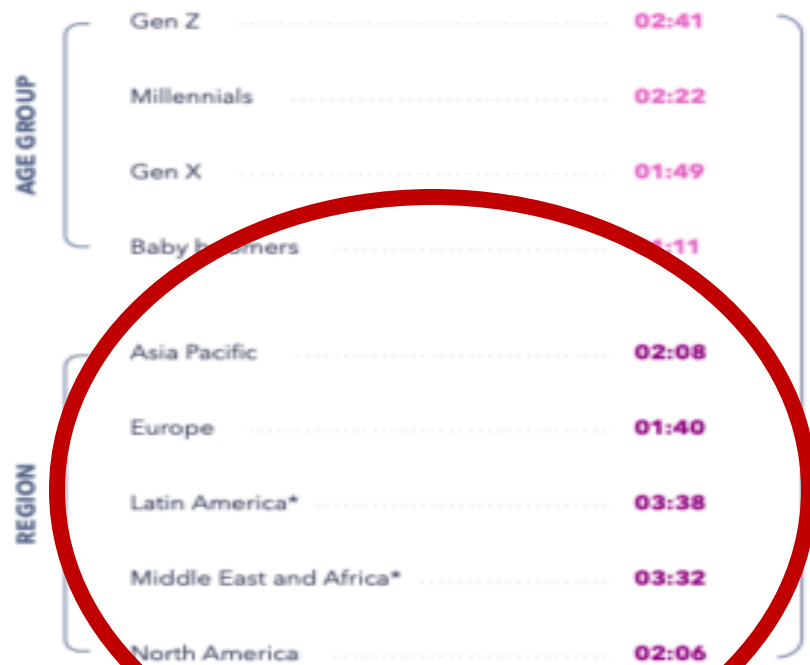
Why should we care



Why should we care

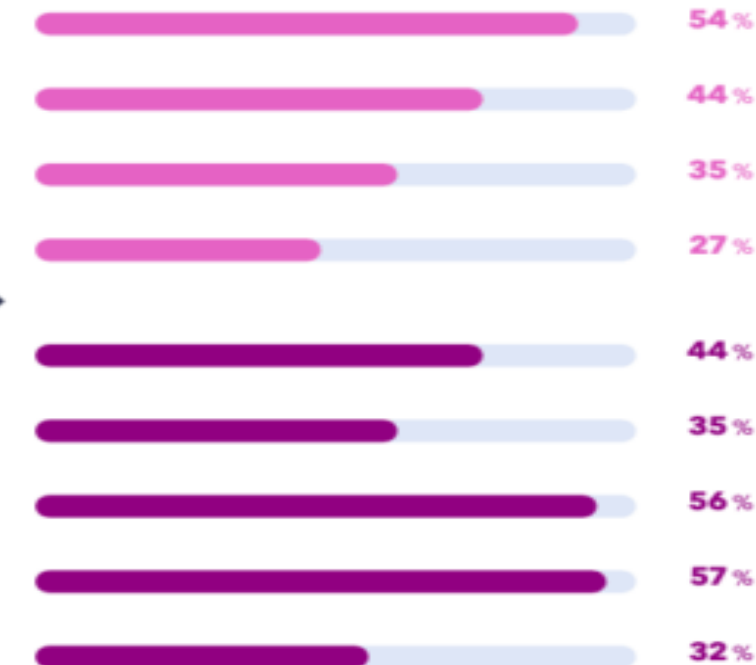
DAILY TIME SPENT ON SOCIAL MEDIA (JANUARY-MARCH)

Average hh:mm spent using social networks on a typical day



SPENDING LONGER ON SOCIAL MEDIA (MAY)

% in each demographic who have been spending longer on social media because of the outbreak



In May 2019, Gen Z and millennials, together with digital consumers in the MEA and Latin America, have been the driving force behind recent increases in social media consumption.

Why should we care

20B+ daily translations

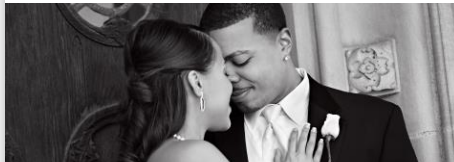



Slide Courtesy Paco Guzman

Native

 **Jole Simmons**
33 mins · 🌐

Congratulations to the newlyweds! I love you!
#maxiejohnsonwedding2017



 **National Cherry Blossom Festival**
2 hr · 🌐

今週末に来てください。花は満開ですし、絶対に素晴らしいです!



 **Le Petit Chat**
1 hr · 🌐

Regardez comment nous fabriquons nos délicieux croissants. #beurre!



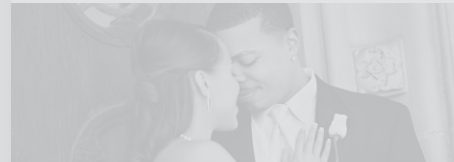
👍👍👍 Bill Russell, Joe Tony and 80 others


👍 Like 💬 Comment ➦ Share

How English Speaker Sees

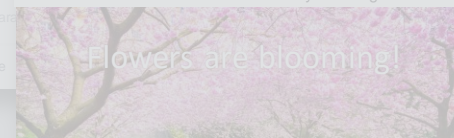
 **Jole Simmons**
33 mins · 🌐

Congratulations to the newlyweds! I love you!
#maxiejohnsonwedding2017



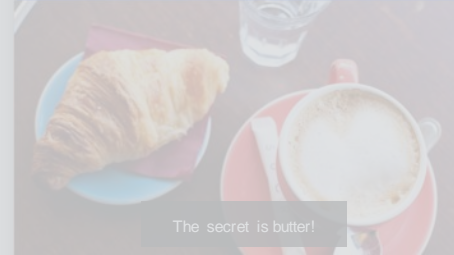
 **National Cherry Blossom Festival**
2 hr · 🌐

Please come down and join us this weekend. The flowers are in full bloom and absolutely stunning!



 **Le Petit Chat**
1 hr · 🌐


Take a look how we make our delicious croissants. #butter!



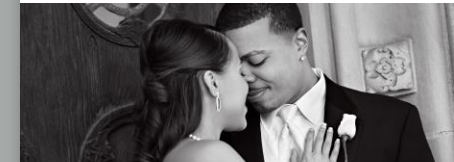
👍👍👍 Bill Russell, Joe Tony and 80 others

👍 Like 💬 Comment ➦ Share

How Turkish Speaker Sees

 **Jole Simmons**
33 dk · 🌐

Yeni evlere tebrikler! Sizi seviyorum! #maxiejohnsonwedding2017



 **National Cherry Blossom Festival**
2 saat · 🌐

Lütfen aşağı gelin ve bu hafta sonu bize katılın. Çiçekler tamamen açtılar ve çok etkileyiciler!



 **Le Petit Chat**
1 saat · 🌐

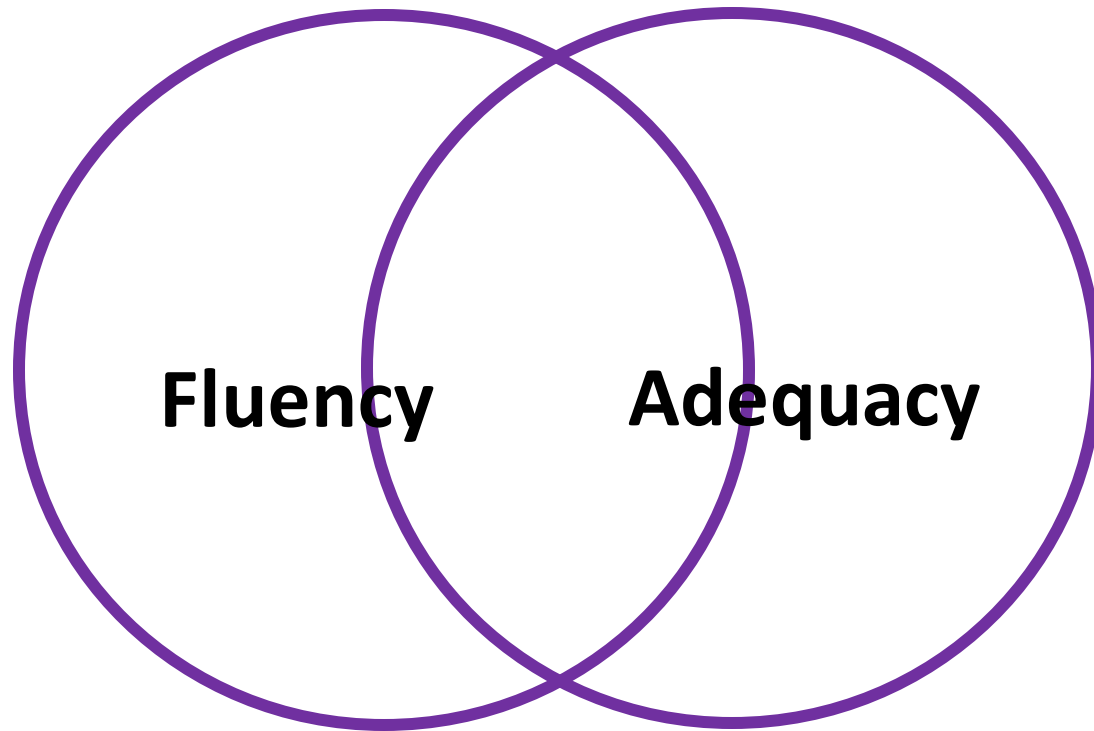
Gelin nasıl lezzetli kruvasanlar yaptığımıza bakın. #tereyağ!



👍👍👍 Bill Russell, Joe Tony ve 80 diğer kişi

👍 Beğen 💬 Yorum Yap ➦ Paylaş

Current Mindset



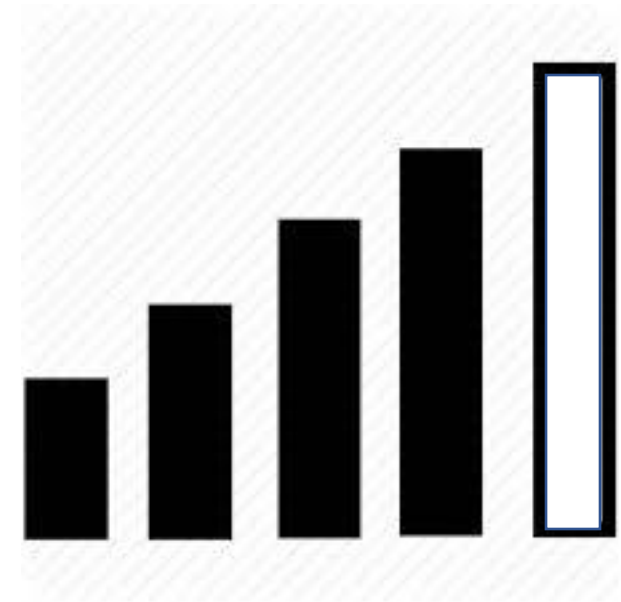
Current Mindset across all data genres



Source

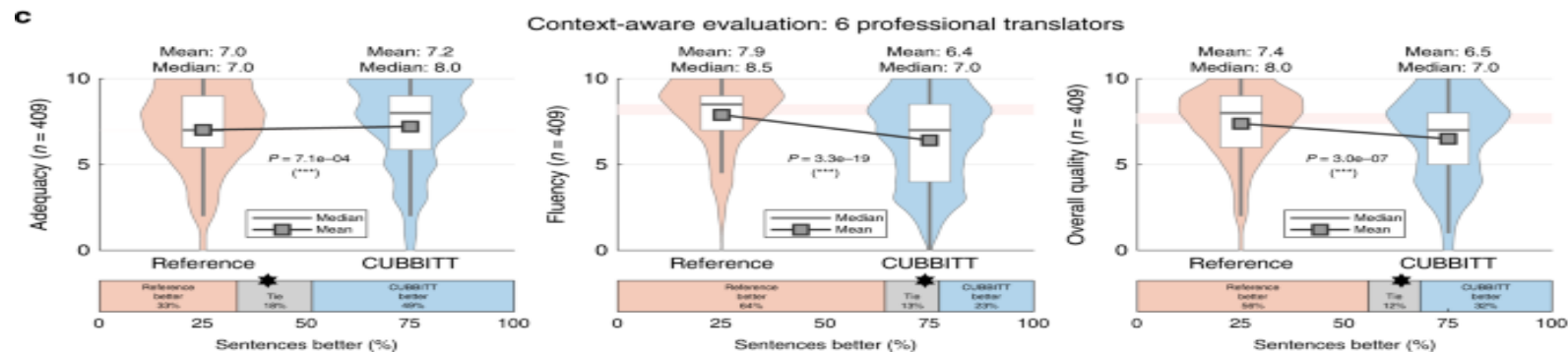
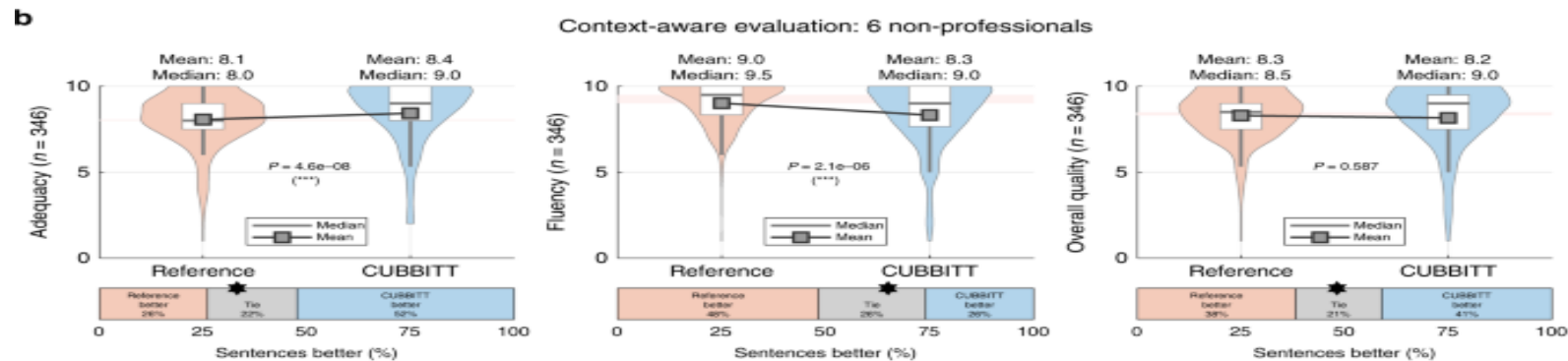
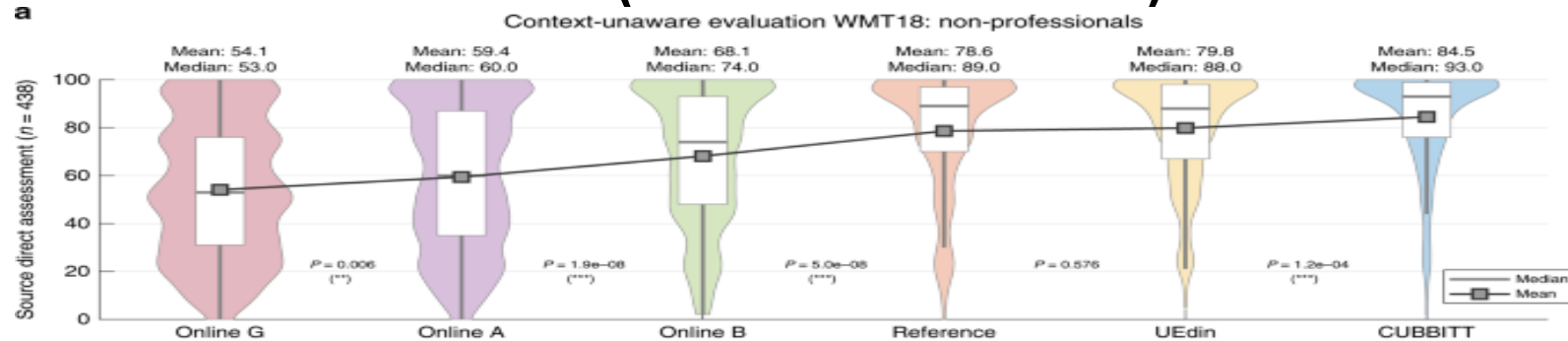


Opaque Box Model



Translation

Fluency & Adequacy are possibly sufficient for Newswire (edited text)



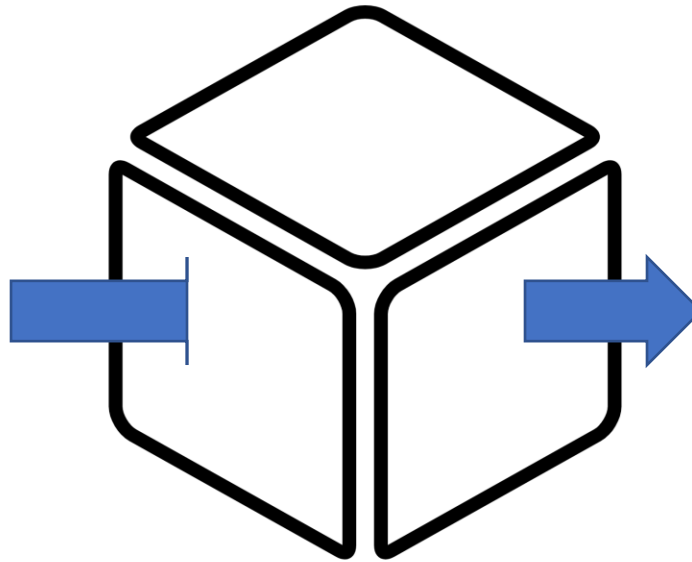
- CUBBITT shows that MT performance passes the Human Turing test EN->CZ newswire (Popel et al, Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals, Nature communications, 2020)

- Adequacy is higher than fluency

Ideally shoot for the North Star



Source



Glass Box Model

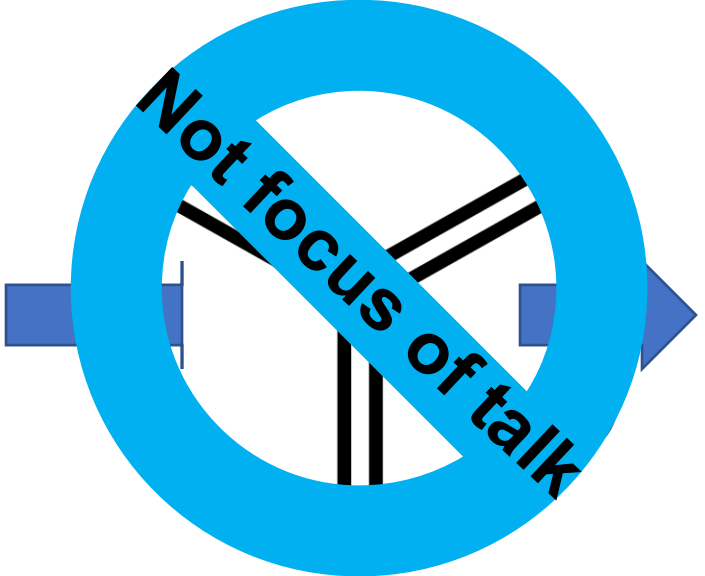


Translation

The North Star: Landing on the moon



Source



Glass Box Model



Translation

What is Faithful MT

- No .. Not about prayers or supplications 😊
- Producing MT that is faithful to the source reflecting the “exact” meaning of the source with no additions (aka hallucinations), deletions, nor “egregious” substitutions
- Beyond translationese
- What it is not
 - Not a judgment on the veridicality of the content
 - Not a judgment on the provenance of the source

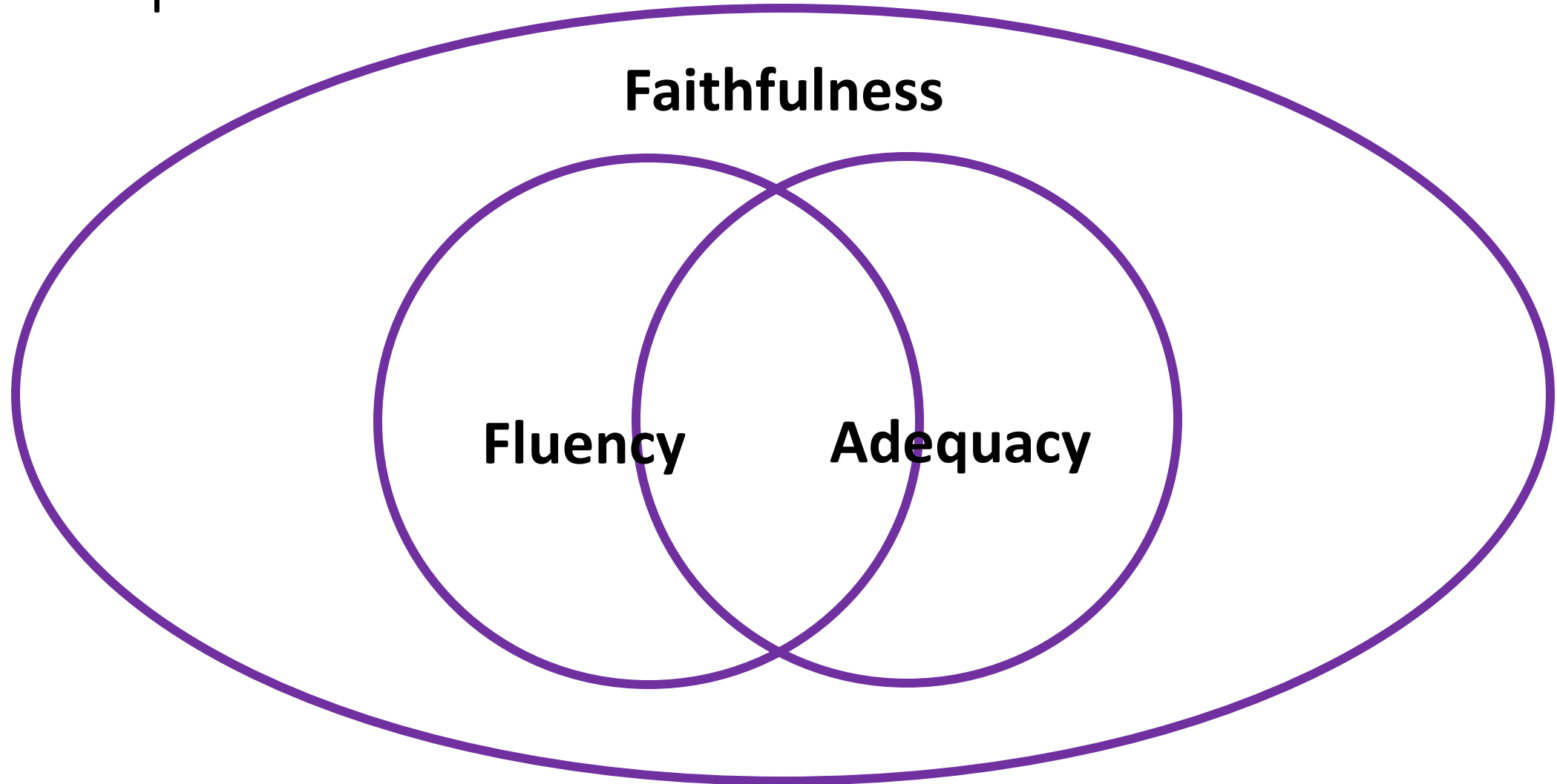
In a nutshell

Faithful translation is about achieving meaning, expression, and “usage” equivalence with the source while maintaining minimal distance

Proposal: A mindset shift

- Moving beyond the *MVP* mindset to adopting a *faithfulness* mindset
 - Go beyond adequate meaning and fluency to include usage considerations, i.e. pragmatics, e.g.:
 - Emotional intensity reflect trends as they have downstream implications
 - Level of formality/coherence in a post is part of the message
 - Use of idiomatic expressions convey cultural nuance
 - Hedging in a message has implications (eg. deception detection)
 - Sarcasm/irony, eg humor vs. hate speech detection
- ***Especially*** pertinent in social media and opinion data (subjective data)
- Potential implications of shifting our mindset
 - adopting faithfulness mindset allows for us to think of other approaches to evaluation (task based)
 - Forcing function toward building more interpretable models

Proposed Mindset

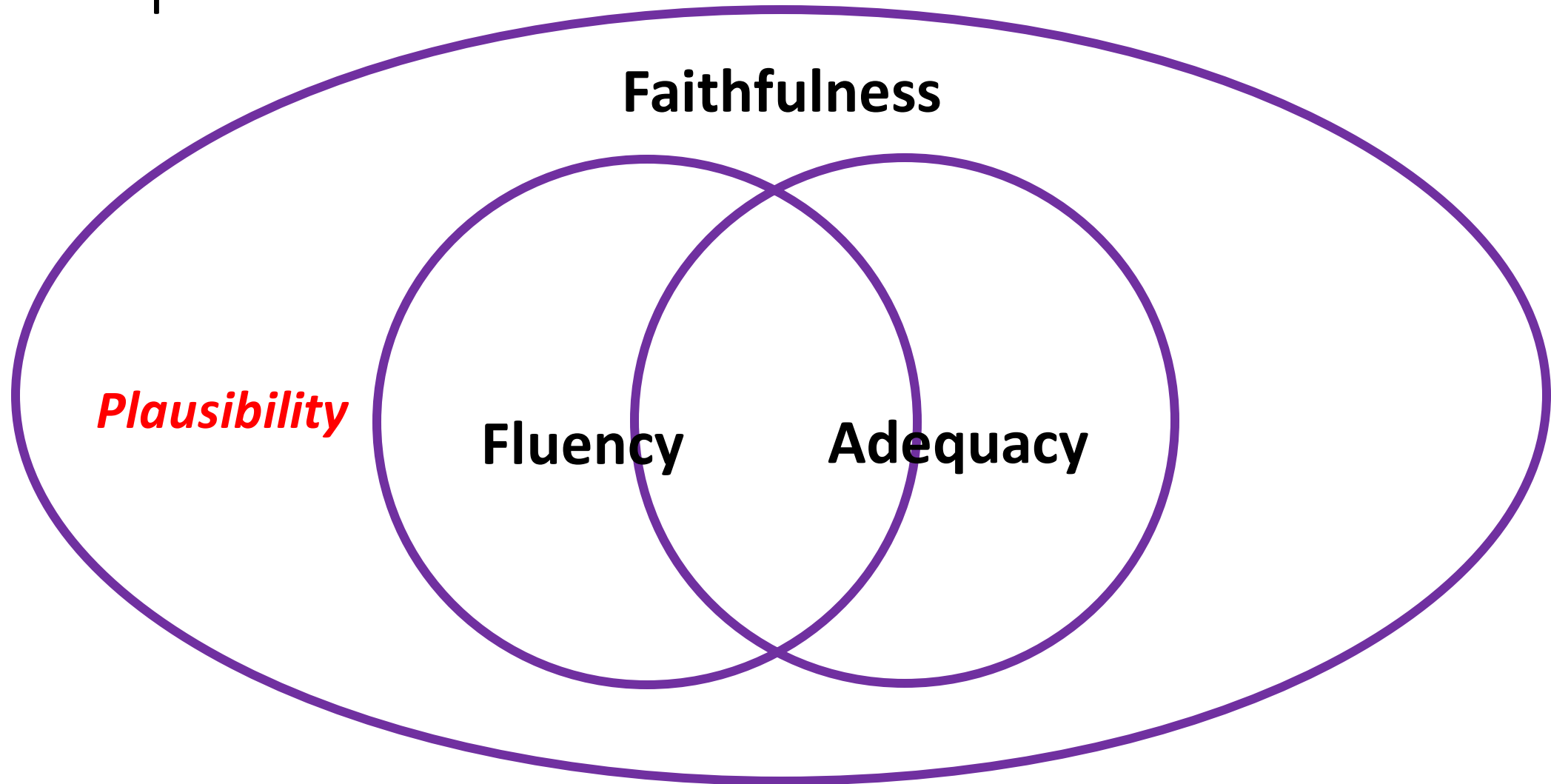


Faithfulness

Fluency

Adequacy

Proposed Mindset



Plausibility beyond fluency

- “Plausibility” refers to how likely is a translation to appear in a target language, e.g.:
 - You made my day -> انت صنعت يومي
 - (verbatim: you built/manufactured my day)
 - **Fluent Arabic output but quite implausible**
 - Correct faithful translation: فرحتيني (Egyptian Arabic)

Faithfulness beyond Fluency

Typically we aim for fluency in the target almost ignoring source fluency, however in social media speech effects play a pragmatic role

- Source

بحب الكوره أووووي

- Fluent Translation

I love football

- Faithful Translation

I loooooove football

Faithfulness beyond Adequacy

- Source

الفسستان يجنن، هاياكل منك حته

- Adequate Translation

The dress is nice, it will eat a piece from you

- Faithful Translation

The dress is stunning, it really fits you nicely

- Source

The president is close his people

- Adequate Translation

الرئيس **قريب** من شعبه

- Faithful Translation

الرئيس **محبوب** من شعبه

BT: The president is admired by his people

What throws users off? Catastrophic translations

- **Bad Named Entity translation** (SimSim [PER] → sesame)
- **Wrong pronouns** (*it, he* instead of *she*)
- **Deleting critical information**
- **False profanities** (*Dog* → *Bitch*)
- **Implausible translations** (made my day → جعلت يومي [realized my day])
- **Introduction of violent terms** (oppose them → attack them)
- **Reverse polarity** (vote him out! → votez pour lui [vote for him])
- **Hallucinating unwarranted target text**

What throws users off? Catastrophic translations

- **Bad Named Entity translation** (SimSim [PER] → sesame)
- **Wrong pronouns** (*it, he* instead of *she*)
- **Deleting critical information**
- **False profanities** (*Dog* → *Bitch*)
- **Implausible translations** (made my day → جعلت يومي [realized my day])
- **Introduction of violent terms** (oppose them → attack them)
- **Reverse polarity** (vote him out! → vote pour lui [vote for him])
- **Hallucinating unwarranted target text**

Adequacy

Faithfulness

MT State of the art on social media

Direction	Source	MT	HT
AR→EN	قدموا هون من خلال الرابط	They provided here through the link	Apply here through the link
EN→FR	Don't forget to hit me up.	N'oublie pas à me frapper . BT: Don't hesitate to hit me.	N'oublie pas à me contacter .
TR→EN	30 derecede sıkmadan ve bastırmadan yıkanabilir	30 degrees can be washed without pressure and fucking	washable at 30 degrees without wringing or pressing
EN→AR	Super relate. Silent treatment to the max	سوبر تتصل . معاملة صامتة إلى أقصى الحدود BT: Super is calling. Silent treatment to the extreme	أنا معاك بالظبط . أتجاهل لأقصى الحدود
AR→EN	سوفجارديت الفيشي	Souvardite Vichy	Saved the file
EN→AR	Vote him out!	صوت لصالحه BT: Vote for him	خرجوه بالانتخابات
AR→EN	مشيت الكلبة بتاعتي في الشارع	My bitch walked down the street	Walked my dog in the street

SoTA MT Analysis: Adequacy vs. Faithfulness

Direction	Source	MT	HT
AR→EN	قدموا هون من خلال الرابط	They provided here through the link	Apply here through the link
EN→FR	Don't forget to hit me up.	N'oublie pas à me frapper. BT: Don't hesitate to hit me.	N'oublie pas à me contacter .
TR→EN	30 derecede sıkmadan ve bastırmadan yıkanabilir	30 degrees can be washed without pressure and fucking	washable at 30 degrees without wringing or pressing
EN→AR	Super relate. Silent treatment to the max	سوبر تتصل. معاملة صامتة إلى أقصى الحدود BT: Super is calling. Silent treatment to the extreme	أنا معاك بالظبط. أتجاهل لأقصى الحدود
AR→EN	سوفجارديت الفيشي	Souvardite Vichy	Saved the file
EN→AR	Vote him out!	صوت لصالحه BT: Vote for him	خرجوه بالانتخابات
AR→EN	مشيت الكلبة بتاعتي في الشارع	My bitch walked down the street	Walked my dog in the street

If our MVP is not there for social media why even think of Faithfulness

- Shouldn't we just aim for MVP (fluency/Adequacy) and *THEN* think of faithfulness
 - Isn't faithfulness a luxury, a "nice to have"!
- I argue not, especially in subjective data (social media and opinion data) where pragmatics are crucial
 - Examples
 - Super relate. Silent treatment to the max → (MT) "Super is calling"
 - (MT) My bitch walked down the street vs. (HT) Walked **my dog** in the street

Challenge of Pragmatic Phenomena

- Many of these pragmatic phenomena are high type/low token frequency (i.e. expressed with very high variability), eg. Named entities, Idiomatic expressions, Neologisms, hashtags
- Social media seems to expedite new expressions being coined in the language (dynamic) tapping into the Gen Z's creativity
- Problem exacerbated for low resource languages such as dialectal variants
- CL Technology Challenges
 - Negation detection
 - Sarcasm detection
 - Humor detection and translation
 - Quantifier scope

What would it take to get there?

- Addressing Style Correspondence, eg.

Source: Hey yu'all what's up?

Faithful Translation: (informal EGY Dialect) سلام يا جماعه، أخباركم إيه؟

VS.

Adequate Translation: (MSA) السلام يا رفاق ما اخباركم؟

- Addressing Emotion Intensity, eg.

• **Source:** ساعات بحب أكل سوشي في العشا

• **Faithful Translation:** *I sometimes like to eat sushi for dinner*

VS.

• **Adequate Translation:** *I love to eat sushi for dinner for hours*

What would it take to get there?

- Addressing Belief Modality

- **Source:** GM may lay off workers
- **Faithful Translation:** جي أم قد تسرح العمال

VS.

- **Adequate Translation:** جي أم ستسرح العمال
BT: GM will lay off workers

- Addressing hallucinations especially around profanities

- **Source:** الفقراء بيطالبو بحقهم في الحياه في أمريكا
- **Faithful Translation:** The poor are demanding their rights in America

VS.

- **Adequate Translation:** The deplorable people are demanding their right to life in America

What would it take to get there?

- Addressing Idiomatic Expressions (Multi Word Expressions [MWE])

- **Source:** !! شوفتم ياناس فلوس العرب بتروح فين ؟ فى الهواء الطلق !!

- **Faithful Translation:** People, Did you see where Arab wealth goes? Into thin air!!

vs.

- **Adequate Translation:** Did you see where Arab money goes? outdoors!!

- Addressing Sarcasm

- **Source:** يا سلام على بجاحتكم يا أخي!

- **Faithful Translation:** Wow what audacity dude!

vs.

- **Adequate Translation:** Oh peace be on your daring attitude bro!

What would it take to get there?

- Addressing Code Switching (MSA and dialect)
 - **Source:** (Dialect and MSA) قدموا هون من خلال الرابط
 - **Faithful Translation:** [Apply through this link](#)
- **vs.**
- **Adequate Translation:** [They arrived here through the link](#)

- Addressing Code Switching (French and MSA)
 - **Source:** سوف جارديت الملف
 - **Faithful Translation:** [Saved the file](#)
- **vs.**
- **Adequate Translation:** [I saved the file](#)

What would it take to get there?

- Addressing other phenomena
 - Shorthand (LOL, TTYL), Neologisms (frenemy, noob), Hashtags
 - Humor and Irony
 - Negation
 - Quantifier scoping
 - Etc.

Modeling Considerations

Light at the end of the tunnel?

Addressing Pragmatics in MT

- Data augmentation for a specific phenomenon, Eg.
- Zaninello & Birch (2020) Multiword Expression aware Neural Machine Translation. LREC
 - EN → IT
 - 4 Approaches
 - Augmenting training data with MWE parallel dictionary lists
 - Leveraging MWE detection tools for source language, then preprocessing as a tokenization step grouping words in an MWE as a single token (Similar to Carpuat & Diab, NAACL 2010 using SMT)
 - Applying factored NMT on the word level by concatenating feature embeddings
 - Backtranslation of target leveraging usage examples thereby augmenting training data
 - Results
 - Backtranslation yields highest results (+1.3 BLEU on general test set, +5.09 on MWE test set)

Light at the end of the tunnel?

Addressing Pragmatics in MT

- Style Transfer in NMT (Controlled Formality Generation), Eg.
- Niu, Rao, Carpuat (2018) **Multi-Task Neural Models for Translating Between Styles Within and Across Languages**. COLING
 - FR → EN, VI → EN
 - **Approach**
 - multi-task learning model performing both bi-directional English formality transfer and translate XX to English with desired formality.
 - It is trained jointly on monolingual formality transfer data and bilingual translation data.
 - **Results**
 - Achieves comparable results to SoTA supervised Side Constraints integration approach despite not needing formality tagged data for the bilingual data used in translation

Light at the end of the tunnel?

Addressing Pragmatics in MT

- Hallucination detection in Neural Sequence Generation (NMT), Eg.
- Zhou et al (Under Review ICLR) **Towards Safe Generation: Detect Hallucination in Neural Sequence Generation.**
 - CH → EN
 - Hallucination: fluent generation but unfaithful to the source input
 - Hallucination is common in **neural machine translation** in **out-of-domain** and **low-resource** test sets (Muller et al., 2019)
 - Task: Given (source, generation), predict if each token in the generation is a hallucination.
 - **Approach**
 - Unsupervised learning of hallucination prediction: 1. create synthetic supervised data; 2. Finetune pretrained LM on the synthetic data
 - **Results**
 - Sentence level hallucination detection using transformer based standard NMT is 78.5% F1

Light at the end of the tunnel?

Enabling Technologies for Pragmatics

- Style Transfer using supervised, unsupervised, semi supervised methods, eg.
 - Shen et al. (2017) Style Transfer from Non-Parallel Text by Cross-Alignment. NIPS
 - Prabhume et al. (2018) Style Transfer Through Back-Translation. ACL
- Emotion (intensity) Classification, eg.
 - *Tafreshi & Diab. (2020) Leveraging Label Projection and Direct Word Mappings for CrossLingual Emotion Detection in Low Resource Languages. [submitted]*
- Sarcasm Detection, eg
 - Ghosh & Veale. (2017) Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. EMNLP
- Level of committed belief, eg.
 - Ulinski, Benjamin, Hirschberg. (2017) Using Hedge Detection to Improve Committed Belief Tagging. ACL Workshop on Computational Semantics beyond Events and Roles

Light at the end of the tunnel?

Potential solutions from enabling to MT modeling

- Similarly for Negation detection, MWE classification, Intra sentential Code Switching
- Various integration modeling techniques
 - Side Constraints
 - Projections through parallel corpora
 - Leveraging cross lingual embeddings for comparable and unrelated corpora
 - MultiTask Learning

Light at the end of the tunnel?

Potential solutions from enabling to MT modeling

- Similarly for Negation detection, MWE classification, Intra-sentential Code Switching
- Various integration techniques
 - side constraints
 - projections through lexical
 - Leveraging cross-lingual
 - MultiTask

**Majority of these systems exist for English
or at best high resource languages**

Light at the end of the tunnel?

Potential solutions from enabling to MT modeling

- Similarly for Negation detection, MWE classification, Intra-sentential Code Switching
- Various integration techniques
 - side constraints
 - projections through lexical
 - Leveraging cross-lingual
 - MultiTask

MT is already English centric since most translation directions lack direct L1-L2 parallel data, hence MT systems pivot thru English, i.e $L1 \rightarrow EN, EN \rightarrow L2$

Mitigating English Centric Bias: Possible solutions

- Increase the pool of pivot languages with various language family representatives (13/14 language families) heeding the following characteristics:
 - Varying typologies
 - High resource (as much as possible)
- Exploiting language relatedness in more creative ways (eg. Aminian & Diab, 2015)

Evaluation Considerations

Evaluation considerations

- Shortcoming of BLEU
 - HT: I did not go to the office
 - (MT1) I have not been to the office << (MT2) I did go to the office
- Desiderata
 - Create targeted data sets that reflect relevant faithfulness phenomena
 - Develop metrics that focus on equivalence (optimizing for faithfulness) rather than translationese, eg. Semantic textual similarity (STS)
 - Minimizing reliance on references
 - Grounding evaluation in user studies on a continuous basis
 - Invest in automated relevant enabling technologies (eg. hallucination detection)
 - Adopt task-based evaluation paradigms such as question answering to address critical errors (e.g. FEQA: Durmus et al. ACL, 2020)

Evaluation Data Sets Mitigating Gender Bias

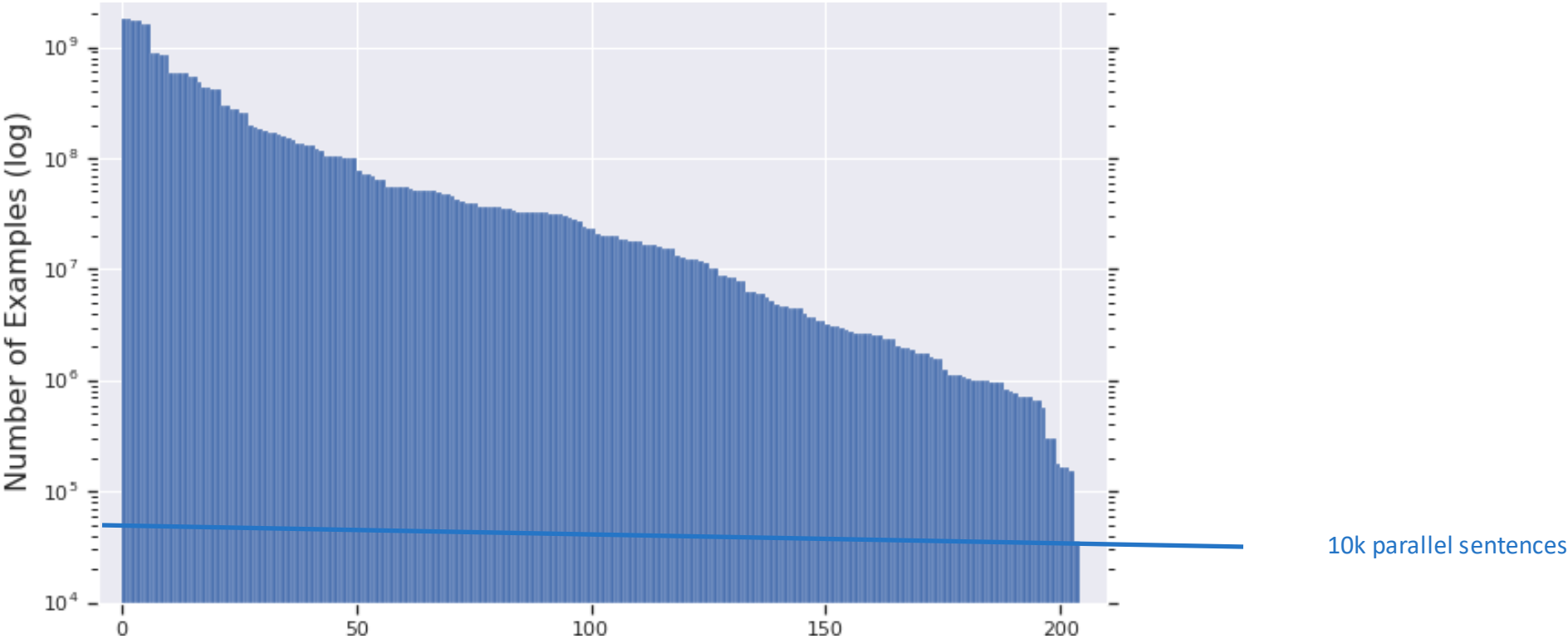
- Gender balanced generation, Eg.
- Habash, Bouamor, Chung. (2019) Automatic Gender Identification and Reinflection in Arabic. ACL 1st Workshop on Gender Bias in NLP
 - EN → AR
 - **Approach**
 - Joint neural model of Gender identification and gender re-inflection to create balanced MT output in Arabic
 - Focus on 1st person singular
 - Apply as post processing step wrapper around NMT
 - Create a gender balanced data set
 - **Results**
 - Achieves 8% relative BLEU improvement



Are all these solutions and considerations for high resource languages

Where are our ethical considerations of Equity and Equality

Low Resource Machine Translation



facebook Artificial Intelligence

Chart credit: Siddhant et al., 2019
Evaluating the Cross-Lingual Effectiveness of Massively Multilingual Neural Machine Translation

Beyond data scarcity



Language
similarity



Domain



Evaluation



FLORES dataset:
Nepali-English, Sinhala-English

Multilingual Training: Efforts to address low resource languages

Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

Table 3: Languages in ML50 Benchmark. We display the languages included in the ML50 Benchmark and the quantity of training data in bitext pairs. Full breakdown is provided in Appendix Table 6.

Data	Translation to English					Translation from English				
	BL-FT	ML-SC		ML-FT		BL-FT	ML-SC		ML-FT	
	→en	N→1	N↔N	N→1	N↔N	en→	1→N	N↔N	1→N	N↔N
>10M	2.7	2.8	1.9	3.8	1.4	1.9	-0.6	-1.7	-0.3	-1.7
1M-10M	3.9	4.8	4.1	6.2	4.4	3.3	1.5	1.0	1.7	0.6
100k-1M	5.7	6.9	7.0	8.2	7.4	4.4	4.0	3.4	4.0	3.2
10K-100K	16.8	17.9	18.3	22.3	20.6	13.4	13.6	13.9	13.5	13.6
4k-10k	11.6	13.1	14.1	18.9	15.0	8.7	10.6	10.9	9.9	9.7
All	8.7	9.7	9.8	12.3	10.6	6.8	6.4	6.0	6.3	5.7

Table 4: Multilingual Finetuning on 50 languages comparing to bilingual models. Improvement in BLEU compared to bilingual training from scratch is shown.

Multilingual Training: Efforts to address low resource languages

Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

Table 3: Languages in ML50 Benchmark. We display the languages included in the ML50 Benchmark and the quantity of training data in bitext pairs. Full breakdown is provided in Appendix Table 6.

Data	Translation to English					Translation from English				
	BL-FT	ML-SC		ML-FT		BL-FT	ML-SC		ML-FT	
	→en	N→1	N↔N	N→1	N↔N	en→	1→N	N↔N	1→N	N↔N
>10M	2.7	2.8	1.9	3.8	1.4	1.9	-0.6	-1.7	-0.3	-1.7
1M-10M	3.9	4.8	4.1	6.2	4.4	3.3	1.5	1.0	1.7	0.6
100k-1M	5.7	6.9	7.0	8.2	7.4	4.4	4.0	3.4	4.0	3.2
10K-100K	16.8	17.9	18.3	22.3	20.6	13.4	13.6	13.9	13.5	13.6
4k-10k	11.6	13.1	14.1	18.9	15.0	8.7	10.6	10.9	9.9	9.7
All	8.7	9.7	9.8	12.3	10.6	6.8	6.4	6.0	6.3	5.7

Table 4: Multilingual Finetuning on 50 languages comparing to bilingual models. Improvement in BLEU compared to bilingual training from scratch is shown.

Open challenges/opportunities to the community

- What does it mean to really model fluency
- What does it really mean to model humor that doesn't transfer over
- How do we deal with neologisms (noob, selfie, frenemy), typos in source, shorthand (ttyl, lol, etc), hashtags
- How do we handle code switching on the input
- How do we generate code switching where appropriate
- How to handle source incoherence
- Reference translations are complicated to generate (especially for low resource scenarios – languages and dialects), resulting in evaluation implications

Take home message

Let's rise above the inherent mediocrity of an MVP

- A lot more questions than answers

BUT

- Shifting our mindset to consider faithfulness as necessary (not optional)
- Keep our eye on the ethical dimensions of our solutions
- Potentially a mindset shift (from only optimizing for fluency and adequacy to optimizing for faithfulness) will result in hybridization of (minimal) feature engineering and representation lending a hand to interpretability (remember that glass box)
- Be creative about our evaluation
- Be grounded in user experience
- Concerted community wide effort is needed to get there

Thank you!
Let the conversation begin