A Survey of Qualitative Error Analysis for Neural Machine Translation Systems

Denise Diaz
Joint work with Vishrav Chaudhary, James Cross, Ahmed El-Kishky, Philipp Koehn

What prompts this study?

- Internet and social media are proliferating rapidly
- Communication and information need to be available to a wide audience in many different languages
- MT has become widely adopted

End-user trust is the goal

With this wide adoption, it has become important to understand where MT models excel and where they struggle in order to improve MT models and ensure end-user trust (Lommel, 2018).

2020 MT Challenges - **Problematic translations**

Problematic translations are those that are **misleading** and may:

- Carry health, safety, political, legal or financial implications

or

- Introduce toxic language not present in source

Qualitative analytic evaluation

- Specific common errors found in neural machine translations (NMT) on the FB platform

- Problematic errors since these are the riskiest of the bunch

Why a qualitative analysis is important

While automatic metrics such as BLEU capture the average case for how well a MT model translates sentences, they don't give insight into *which linguistic aspects* MT models struggle with.

In this qualitative analysis, we investigated MT samples with native speakers so we could review the *linguistic aspects* of MT errors.

Categorizing errors and making a challenging test set is the first step in benchmarking and improving MT performance in linguistic aspects.

10 Language families, 33 languages

Altaic
Turkish

Afro-Asiatic
Semitic
Amharic
Arabic
Hebrew
Cushitic
Somali
Chadic
Hausa

Niger Congo
Zulu

Sino-Tibetan
Chinese

Japanese
Japanese

Austronesian
Tagalog

Austro-Asiatic
Vietnamese

Kra-Dai
Lao

Dravidian
Kannada
Malayalam
Tamil

Indo-European
Balto Slavic
Belarusian
Russian
Bulgarian
Germanic
Swedish
German
Norwegian
Romance
Catalan
French
Italian
Portuguese
Spanish
Indo-Iranian
Farsi
Pashto
Indo-Aryan
Hindi
Marathi
Sinhalese
Urdu

Why we chose these languages

Languages our model supports

Native language informant availability

Diverse language families

High, mid and low resource languages

Error categories

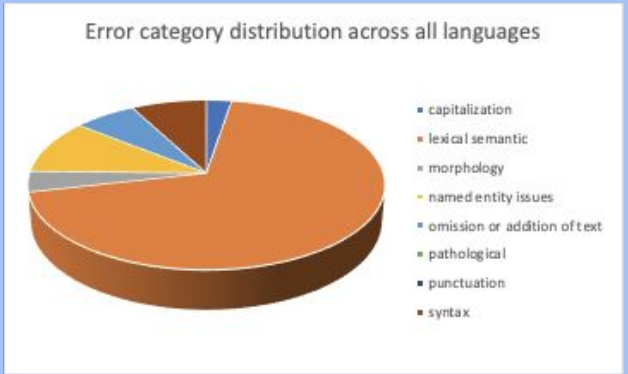
1. Lexical-semantic
2. Named entity issues
3. Morphology
4. Syntax
5. Omission or addition of text
6. Punctuation
7. Capitalization
8. Pathological

★ synthetic samples for illustration

★ no user data is displayed for privacy reasons

Error category average percentages - all languages

Lexical semantic Word ambiguity Noisy source Unknown words Code-switching Dialectal variants	30.00%
Named entity issues	4.00%
Omission or addition of text	7.00%
Pathological translations	3.00%
Syntax	3.00%
Morphology	2.00%
Capitalization	1.00%
Punctuation	0.01%



Lexical semantic

Broad triggers for inappropriate lexical choices in MT include:

- **Word ambiguity**
- **Idiomatic expressions**
- **Phrasal verbs**
- **Noisy source**
  - Misspellings / typos
  - Reduplicated letters
  - Typographical substitution
- **Unknown words**
  - Abbreviations
  - Neologisms or archaic words
  - Vernacular
- **Code switching**
- **Dialectal variants of lexical items**

This was the most prevalent error category across all 33 languages with an average of 30%. In these instances the model was unable to output an appropriate lexical choice to match the source, thus derailing the meaning of translations.

Word ambiguity

"Learning how to disambiguate ambiguous words is one of the most difficult and most important challenges in MT." (Popovic, 2018)

Not much context, just a named entity!

Source Portuguese	Target English	Desired English output
Morro de São Paulo	*I die of São Paulo*	*Morro de São Paulo*

More context helped the model to disambiguate from the verb form to the noun

Source Portuguese	Target English
Vou para o Morro de São Paulo	*I'm going to São Paulo hill*

Idiomatic expressions

Source English	Target Italian	Desired Italian output
Twist my arm!	*Girami il braccio!*	*Non devi convincermi!*

Phrasal verbs

The model sometimes does not recognize phrasal verbs, verbs that are accompanied by a particle or more.

The particles flanking the verb tend to nuance or even change the original meaning of the verb within the phrase, confusing the model.

Source English	Target Spanish	Expected Spanish output
Could you break down those dance moves?	*Podrías romper esos movimientos de baile?*	*Podrías mostrar esos movimientos de baile?*

Noisy source: typos

Source French	Target English	Desired English output
Occupez vous de vis enfants	*English: Take care of kids screws*	*Take care of your kids*

Unknown words: vernacular, neologisms, abbreviations

Vernacular, also current neologism

Abbreviation

Source English	*Decoded*	*Target Spanish*
steezy	*Style with ease*	*Steezy*
TMI	*Too much information*	*tmi tmi*

Dialectal differences

- Phonetic:

English term	IPA transcription with stressed back vowel /ɑ/	IPA transcription with stressed front vowel, /æ/
pajamas	*pə ˈdʒɑː ˌməz*	*pə ˈdʒæː ˌməz*

- Semantic:

Source: British English vernacular	(equivalent Standard American English)	French output:
Dying for a fag!	*Dying for a smoke!*	*Je meurs d'envie d'une tapette*

2. Named entity issues

"Named entities have proven to be some of the most difficult lexical items for the model to tackle." (Ugawa et al., 2018)

Arabic: أم كلثوم

English: *The mother of Kalthoum*

Desired output: *Oum Kalthoum*

NE issues occurred on average 4% across all languages

3. Morphology

English: *Cool down the brake system, cool it!*

Portuguese: *Esfrie o sistema de freio, esfrie!*

THE PRONOUN FOR "IT" IS ABSENT

Desired output: *Esfrie o sistema de freio, esfrie-o!*

Morphological errors occurred 2% on average across languages

4. Syntax

Source Spanish	Target English	Desired English output
disponibles relojes originales en cali	*Original Cali watches available*	*Original watches available in Cali*

3% average across all languages

5. Omission or addition of text

Source Spanish	English output	Desired English output
Dr. Núñez 👨🏽‍⚕️	*Dr.* 👨🏽 ⚕️	*Dr. Núñez* 👨🏽‍⚕️

7% average across all languages

6. Punctuation

English source	Target Arabic	Desired Arabic output
Wow!	!واو	واو!

.09% across all languages

7. Capitalization

Source English	Target Italian	Desired Italian output
Vivaldi's Four Seasons!	*Le quattro stagioni di Vivaldi!*	*Le Quattro Stagioni di Vivaldi!*

2% incidence across all languages

8. Pathological errors

- ➢ Nonsensical or ludicrous
- ➢ Problematic, introducing language that is confusing or even potentially dangerous

- Stuttering
- Toxic language not present in source
- A reversal in polarity or sentiment
- Health or safety risks due to misinformation
- Mistranslated named entities
- Changed units/time/date/numbers

"*With pathological errors the model renders an aberrant output, untethered from source, displaying what are known in industry as hallucinating errors.*" (Koehn and Knowles, 2017; Stahlberg, 2020).

Pathological translation samples

Nonsensical but not toxic

Toxic language is introduced

Source Italian	Target English	Desired English output
Congratulazioni! 🥂	*I'm sorry!* 🥂	*Congratulations* 🥂
È deceduto Antonio	*F..k Antonio*	*Antonio passed away*

Stuttering of additional text

Source English	Target Italian	Backtranslation	Desired output Italian
J. Hill I think	Ciao. Ciao. Hill, credo	Hi. Hi. Hill, think	J. Hill credo

Machine translation is continuously improving!

- Source phrases sampled last year no longer display many of the original errors from 2018-2019!

WITH CONTINUOUS TRAINING

- MT models continue to improve with more training data

  but

- They need to keep improving in order to ensure optimal end-user trust!

What is next?

1. Developing techniques to improve translations for named entities
2. Developing techniques for profanity aware translation (false positives)
3. Developing techniques for translating into morphologically-rich languages.
    a. Small changes in morphology can mean important changes in meaning
4. Curating a new dataset that includes a variety of errors described today
    a. In addition to BLEU, evaluate MT performance on these error types


thank you
danke
謝謝
gracias
merci
obrigado
شكراً جزيلاً
ngiyabonga
teşekkür ederim
go raibh maith agat
mochchakkeram
dziękuję
bedankt
sukriya
kop khun krap
terima kasih
спасибо
Баярлалаа
рахмат
dank je
arigatô
dakujem
grazie
takk
sagolun
감사합니다
ありがとう
ευχαριστώ
мерси
tapadh leat
hvala
maururu
chnorakaloutioun
xièxie
rahmet
tanemirt
diolch
dhanyavadagalu
shukriya
trugarez
kam sah hamnida
najis tuke
didi madloba
mèsi
sobodi
dêkuji
enkosi
bayarlalaa
gràcie
dhanyavad
kiitos
dankie
nandri
nanni
köszönöm
faafetai lava
vinaka
спасибі
blagodaram
mersi
kia ora
barka
welalin
tack
spas
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
djiere dieuf
tau
дякую
mamnun
chokrane
gratias ago
gràcies
sulpáy
taiku
akun
dankon
aciú
merce
rahmat

Q & A

Contact information:

denisediaz@fb.com

References

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation.

Lommel, A. (2018). Metrics for translation quality assessment: a case for standardising error typologies. In Translation Quality Assessment, pages 109–127. Springer.

Popovic, M. (2018). Error classification and analysis for machine translation quality assessment. In Translation Quality Assessment, pages 129–158. Springer.

Stahlberg, F. (2020). The Roles of Language Models and Hierarchical Models in Neural Sequence-to-Sequence Prediction. PhD thesis, University of Cambridge.

Ugawa, A., Tamura, A., Ninomiya, T., Takamura, H., and Okumura, M. (2018). Neural MT incorporating named entity. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3240–3250, S