

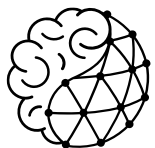
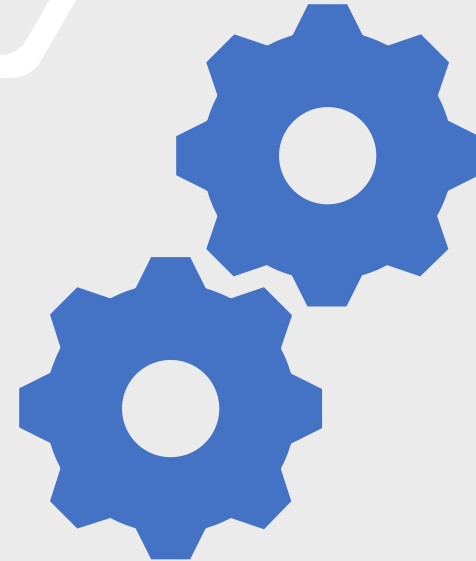
# Quality Estimation for Machine Translation

(QE4MT)

Diptesh Kanojia



CENTRE FOR  
TRANSLATION  
STUDIES  
UNIVERSITY OF SURREY

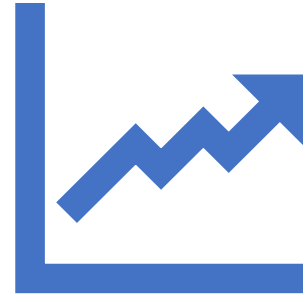


**People-Centred AI**  
UNIVERSITY OF SURREY

# Research Domains



Evaluation



Improvements

# Research Domains



Evaluation



Improvements

# Why Estimate?



**BLEU scores have become a de-facto standard** when it comes to evaluation of machine generated text (translation/summarization/...)

**Criticized for low co-relation** with human evaluation of machine translated output (Reiter, 2018)

Other statistical measures, like BLEU, do not take 'semantics' into account.



Need for a **measure which takes a more 'meaningful' comparison** into account.



Distributional Semantics (word embeddings) provide a viable method to compare source input with target side output.

# Quality Estimation

## Input

- Source side text
- Target side text

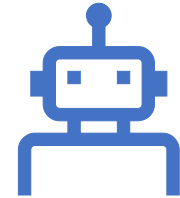
## Output

- Score on a scale of 0-100\*

No more reliance on parallel data for evaluation

# Building Upon The \*QueST\*

- Early adoption of QE research was based on models produced using QuEst/QuEst++.
  - TransQuEst (Ranasinghe et. al., 2020) provides a reliable framework for building Quality Estimation (QE) models for many language pairs.
  - Research on QE is growing as new language pair data is introduced.
- 
- However, our research begets questions on the **reliability** and **robustness** of these systems in evaluating MT output.



# Errors in Machine Translation

## Hallucination

Text with a significant word count.

Text which contains special characters.

## Negation

Ignore the present of 'not', 'neither', 'none' in some cases.

## Incorrect use of upper/lower case letters

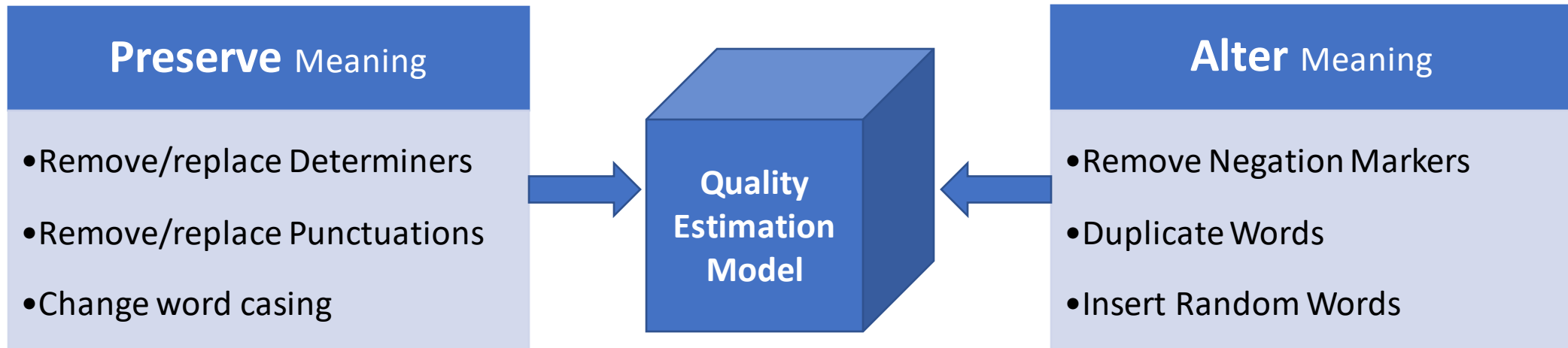
Terms are written in upper case, depending on the context (King/king), but an MT system is unable to detect when to do this

## Untranslated acronyms

WHO (World Health Organization), which, in Spanish is OMS (Organización Mundial de la Salud)

# Linguistic Perturbations

**An initial study which posed questions over the adequacy of the machine translated text and evaluated the performance of QE models using linguistic perturbations.**





# Outcomes



**State-of-the-art (SoTA) QE models are able to capture errors many errors and penalize accordingly.**



Robust to meaning preserving changes  
**Unreliable performance when meaning is altered by machine translation output**



There are many issues to consider:

- Removal of negation does not render a very different score.
- Replacing words with their antonyms had practically no effect on many examples.
- and then some more...



Unlike other multilingual natural language processing applications; multilingual QE models do not perform as well as models trained on a single language pair.

# Limitations and Future Directions

- Resource restricted scenario
  - Use of automated methods to generate perturbed examples restricted us to use of data from language pairs where English was used.
  - Observations on five language pairs.
- A further limitation in exploring/creating tools for other languages in this space is non-availability of datasets for which these tools would possibly be designed.
- Therefore, **let us first create the QE datasets.**



# Indian Languages

## – Low-resource?

- **No datasets for Quality Estimation** for Indian languages.
- Training a quality estimation **model requires ‘direct assessment’ scores from human annotators.**
  - Native speakers of Indian languages like Marathi, Hindi, Tamil, *etc.* who understood English.
- **No leaderboard** for the larger language understanding paradigm, let alone Quality Estimation.
- However, **embeddings models and language models including multilingual variants have been produced** from recent research.
- **Large monolingual and parallel corpora** for many Indian languages/pairs (including English).

# Recent and Upcoming Datasets

## Indo-Aryan Language Family

- English – Marathi Quality Estimation
  - Released at Conference for Machine Translation (WMT) 2022 Quality Estimation Shared Task.
- English – Hindi Quality Estimation
  - Data Collection ongoing
- English – Gujarati
- English – Bengali
- English – Assamese

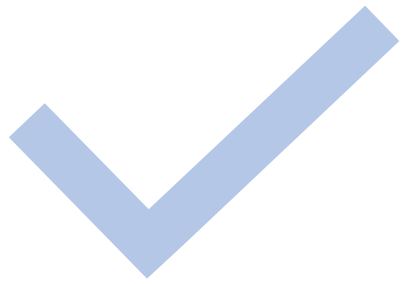
## Dravidian Language Family

- English – Tamil
- English – Telugu
- English – Kannada

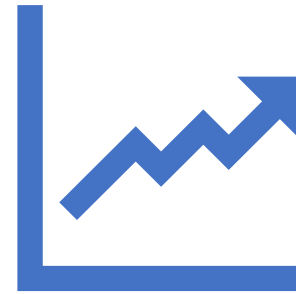
## Further Research

- **Multi-tasking QE models** which perform sentence and word-level QE at the same time.
- **QE models which are robust to linguistic perturbations** generated synthetically.
- **Multilingual QE model** for Indian Languages
- **Document-level Quality Estimation** for English-Tamil

# Research Domains



Evaluation



Improvements

# Automatic Post-Editing

## Indo-Aryan Language Family

- English – Marathi Post-edits
  - Released at Conference for Machine Translation (WMT) 2022 APE Shared Task.
- English – Hindi Post-edits

## Dravidian Language Family

- English – Tamil

## Other Collaborations



JUST: Access





# Thank you!

**Questions at the end of the panel.**



**d.kanojia@surrey.ac.uk**

# References

- Ehud Reiter; A Structured Review of the Validity of BLEU. *Computational Linguistics* 2018; 44 (3): 393–401.  
doi: [https://doi.org/10.1162/coli\\_a\\_00322](https://doi.org/10.1162/coli_a_00322)
- Ranasinghe, T., Orasan, C., & Mitkov, R. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*.