

Machine Translation Quality Assessment Lesson 2 – Using Metrics Critically

BlonDe
Doc-COMET
DiscoScore
SLIDE

Context
Document
Direct Assessment
Metrics



Learning Outcomes

LO1 – Develop the awareness of the role of Human Evaluation (HE) in the development of Machine Translation (MT) systems.

LO2 – Develop the awareness of the role of Automatic Evaluation Metrics (AEM) in the development of MT systems.

LO3 – Develop the understanding of key elements in the developments of MT systems, including purpose of system, evaluation metrics, type of evaluators and ethical considerations, to ensure systems are less risky and less biased.

Structure

- 1** - Recap
- 2** - Has Machine Translation achieved Human Parity?
- 3** - Approaches involving Context for Machine Translation
- 4** - Context from Translation Studies
- 5** - What has been done in document-level MT?
- 6** - Context - Human and Semi-Automatic Evaluation
- 7** - Context - Automatic Evaluation Metrics

**Has Machine Translation
achieved Human Parity?**

Has Machine Translation achieved Human Parity?

Wait...what exactly does that mean? What is human parity?

Claims that machine translation systems can match human translators. But... when performing evaluation you have to consider both how it is performed and how it is reported.



Has Machine Translation achieved Human Parity?


Methodology

Shared tasks consist of a competition of systems. They either use automatic evaluation or crowd evaluation.

Reporting

Based on these tasks, there have been claims that a machine matches a human if a segment-level crowd evaluation or automatic evaluation achieves the same score as a human translated text.

Has Machine Translation achieved Human Parity?



WMT is one of the most important shared tasks in the community!

Workshop on Machine Translation

The shared tasks in WMT investigate different aspects of MT quality assessment - such as the capabilities of MT systems, metrics, test suites, correlations with human evaluation.

It is a platform for researchers to compare their models across different tasks.

Has Machine Translation achieved Human Parity?

These contribute to an overestimation of the capability of systems, right?

Correct! Like we spoke last week, unless you perform an evaluation that is comprehensive and replicable, it is very difficult to make a claim that machine translation systems match humans.



Has Machine Translation achieved Human Parity?

But what is a comprehensive evaluation?

Well...that is one big question to answer. There is not an exact consensus on what is “comprehensive”, although it has been reported before...



What is a comprehensive evaluation?

Cavalheiro Camargo, Castilho and Moorkens (2024), for example, asked machine translation teachers what they considered a “comprehensive evaluation”, which included answers, such as:

- Combined human evaluation and automatic evaluation with state of the art metrics.
- Evaluation with platforms that presented good user experience (UX).
- The inclusion of technical elements (training data, speed, pricing, pollution).
- Task-based evaluation, focusing on the intended use of the system.
- Risk assessment.
- Document-level considerations.



Has Machine Translation achieved Human Parity?

So I have to do everything from that list? What should I do?

Again, it is not exactly a consensus. However, remember how methodology is important for evaluation?



Has Machine Translation achieved Human Parity?

So I have to do everything from that list? What should I do?

Claims that machine translation systems match human translators were made in the past when evaluations were performed at sentence level. With current research, we have developed different methods.



On context for machine translation

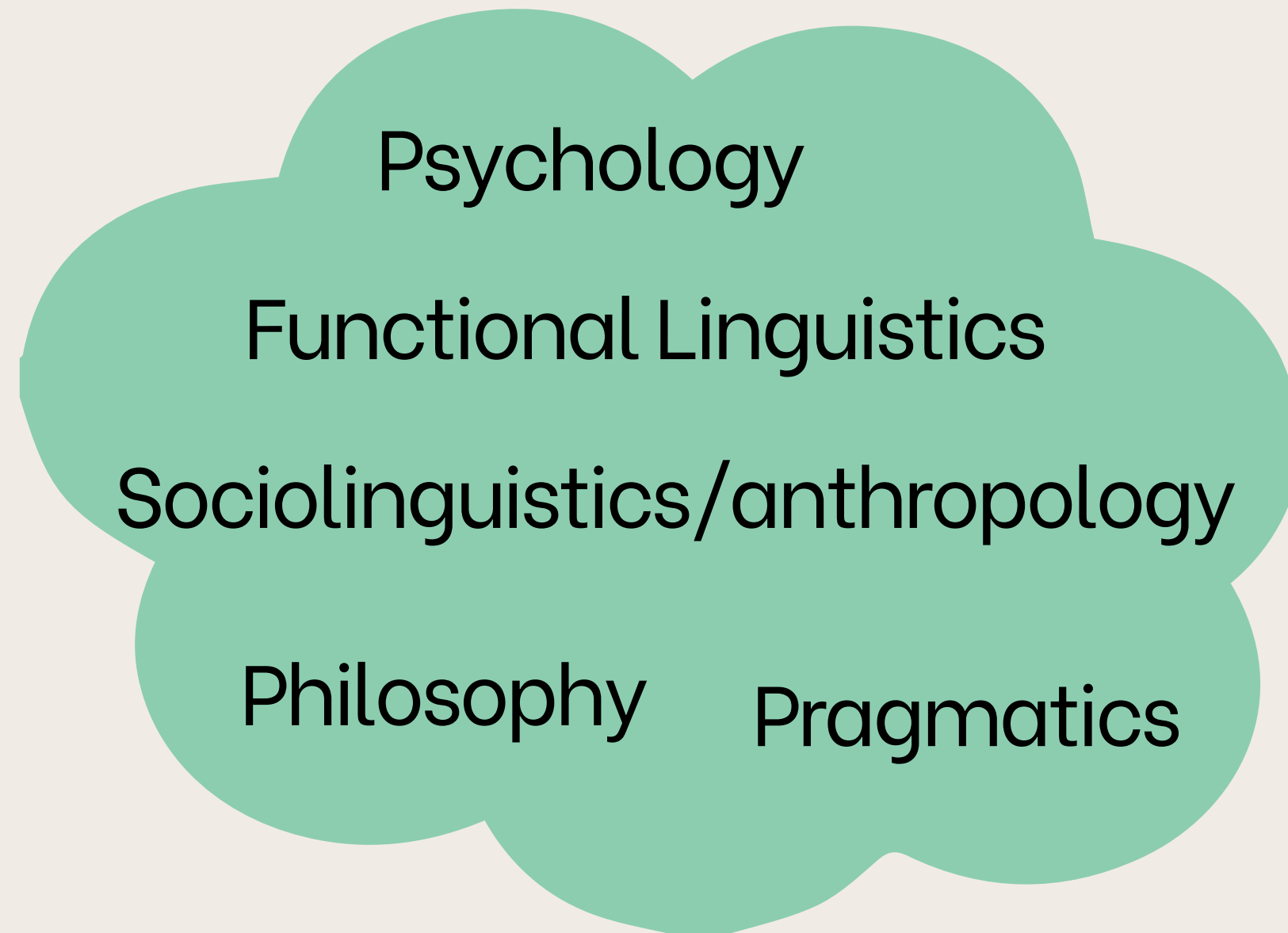
Different approaches have been found in the literature:

- Shared reconstructors and joint learning for a better translation of pronouns in NMT (Wang, 2019)
- Concatenation of segments leading to a context span of sentence pairs (Tiedemann and Scherrer, 2017; Bawden et al., 2019; Müller et al., 2018)
- Use of document substructures (Dobrevva, Zhou, and Bawden, 2020)
- Pushing the limits of the concatenation of segments (Junczys-Dowmunt, 2019; Voita, Senrich, and Titov, 2019b; Lopes et al., 2020).



Context for Translation Studies

Context for Translation Studies is also quite important, stemming from different traditions:



Reviewed by
House (2006)

Context for Translation Studies

A practical five-part definition by Melby and Foster (2010)

Co-text

- Pertains to the word or phrase limited to the surrounding text but not limited to the current sentence.
- **Example:** Definition of a component of an object in an instructions manual

Chron-text

- Chronological changes in a source text
- **Example:** Different versions of a Brazilian Literature book translated into Spanish, then English.

Rel-text

- It is defined by related documents and other resources of the document being translated.
- Example: A monolingual dictionary or a terminology glossary.

Bi-text

- A text aligned with its translation.
- **Example:** Translation memories used by translators can be considered a bi-text, as they are a bilingual database.

Non-text

- Aspects of context not accessed through written texts.
- **Example:** Technical knowledge or cultural knowledge not immediately available in the text.

Context for MT

Intersentential Context

- Incorporating context between sentences is beneficial.
- Potential to provide information on pronouns, deixis, ellipsis, cohesion.
- Translation of paragraphs or entire documents at one time.

World knowledge and external information

- Incorporating levels of formality, such as informal or formal use of pronouns, honorifics, etc.
- Control of domain specific features.
- Specific use case systems that incorporate world knowledge and external information.

Terminology

- Similar to world knowledge, but no clear distinction.
- Emphasis on introducing additional context from a lexicon or terminology resource into MT.
- Consistency in the whole document.

Castilho and Knowles (2024)

What has been done in document-level MT?

Context in Neural Machine Translation: A review of models and evaluations

Popescu-Belis (2019)

- ✓ Summarised studies from 2017–2018
- ✓ Studies in three categories: lexical choices, reference phenomena and discourse structure
- ✓ They were only able to determine the usefulness of systems because a combination of human evaluation and automatic metrics was presented.
- ✓ Test sets were vital to test the systems.

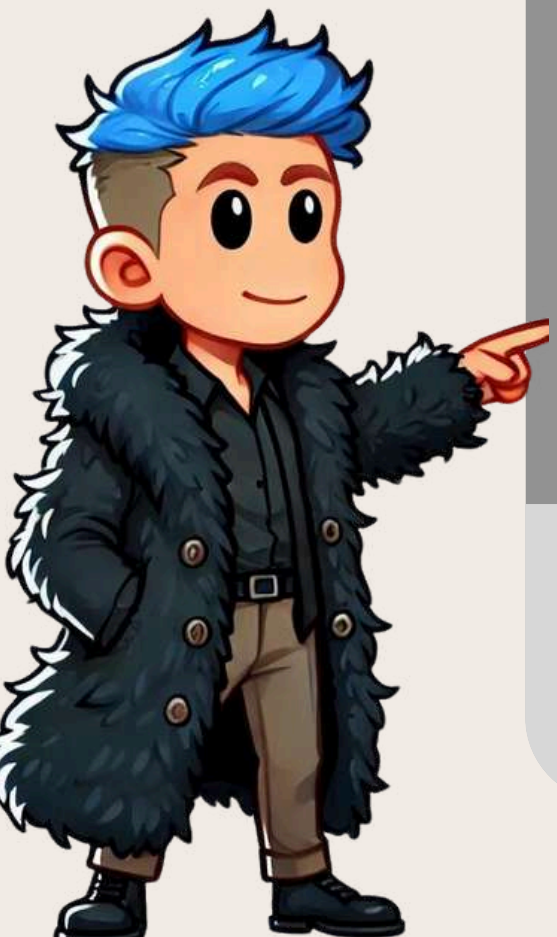


What has been done in document-level MT?

When and Why is Document-level Context Useful in Neural Machine Translation?

Kim, Tran, Ney (2019)

- ✓ Looked at different approaches to include context in NMT
- ✓ Approaches tested: Single-Encoder approach, Multi-Encoder Approach (Inside or Outside it) and Word filtering.
- ✓ Long-range context raises quality only slightly in comparison to regular models.
- ✓ Make a strong sentence-level NMT baseline first before applying these approaches.

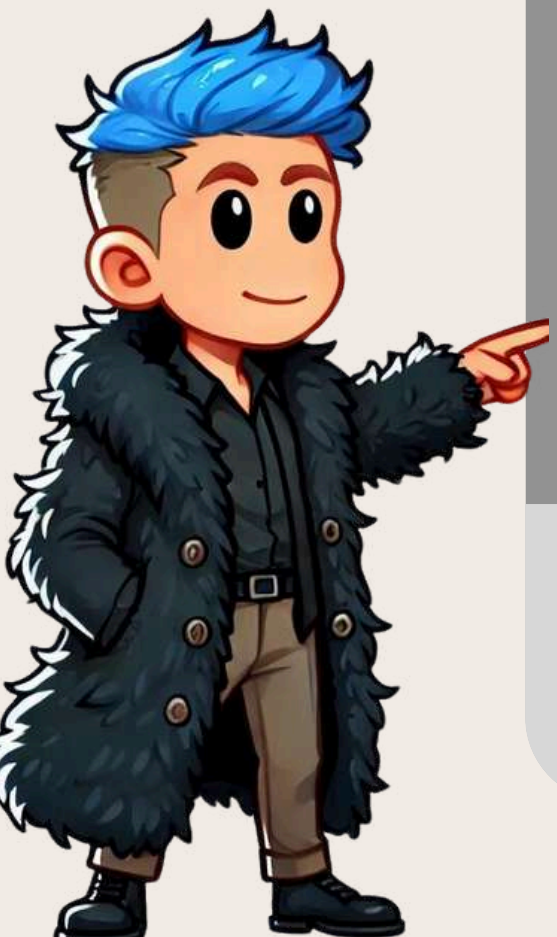


What has been done in document-level MT?

Document-level neural MT: A Systematic Comparison

Lopes et al. (2020)

- ✓ Analysed context-aware NMT methods using large datasets.
- ✓ Combined human evaluation (Error annotation) with automatic evaluation (BLEU).
- ✓ Test sets used to test anaphora, lexical choice, pronouns.
- ✓ Conclude that the approaches at the time are less advantageous in scenarios with larger datasets.



What has been done in document-level MT?

A survey on Document-level NMT: Methods and Evaluation

Maruf, Saleh, Haffari (2021)

- ✓ Analysed implementations up until 2020.
- ✓ Described systems that were used in WMT2019 and WNGT2019.
- ✓ The automatic metrics used were mostly BLEU and METEOR, but TER was found in some evaluations.
- ✓ The authors recommend a middle ground for automatic and human evaluation.



What has been done in document-level MT?

Challenges in Context-Aware NMT

Jin et al. (2023)

- ✓ Focus on obstacles of document-level evaluation through an empirical analysis.
- ✓ Investigate using additional context to the source by concatenating previous and subsequent sentences.
- ✓ Used BLEU, COMET and BlonDe on XFMR and MEGA architectures.
- ✓ No human evaluation, so results are limited!

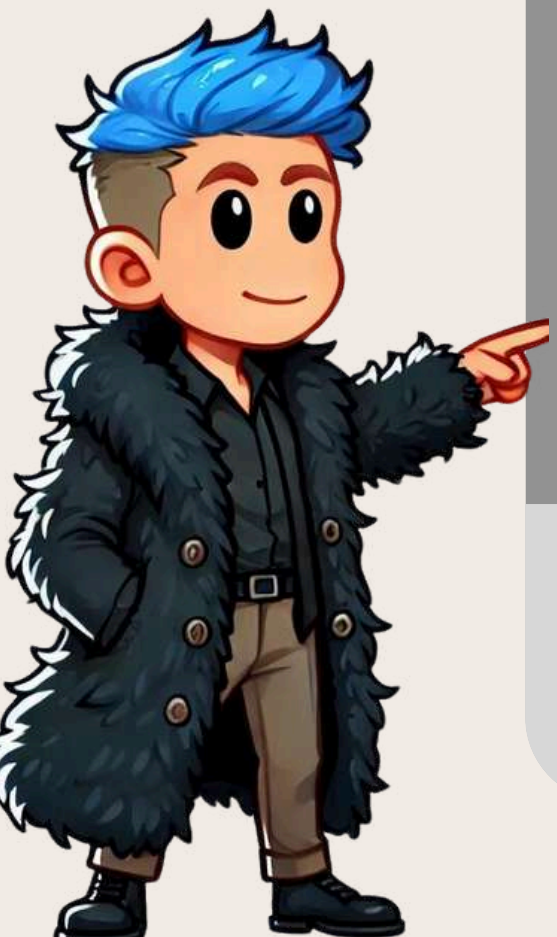


What has been done in document-level MT?

A Survey of context in NMT and its evaluation


Castilho and Knowles (2024)

- ✓ Investigates studies that evaluate approaches for NMT and Large Language Models (LLM) involving context.
- ✓ Report the studies from the previous slides, in addition to evaluation metrics.
- ✓ Identify the challenges in the field – greater emphasis for ethical considerations and for automatic and human evaluation regarding context.



Human Evaluation, Semi-Automatic and Context

The field received many recommendations on best practices in evaluation.



We will see these
in our evaluation
planning lecture!

Some recommendations

Toral et al. (2018)
Läubli et al. (2018)
Graham et al. (2019)
Läubli et al. (2020)
Gilbert (2023)

Human Evaluation, Semi-Automatic and Context

Research was made to investigate how much context was needed to be shown to translators.



We often call this
“context span”


On Context Span Needed for Machine Translation Evaluation

Castilho, Popovic, Way (2020)

- Across three different domains and 18 target languages, over 33% sentences' ambiguity could be resolved with two preceding sentences.
- Other sentences required further preceding sentences, following sentences and global context.

Human Evaluation, Semi-Automatic and Context

Research was made to investigate how much context was needed to be shown to translators.



The document-level setup mitigated cases of misevaluation.

On the Same Page? Comparing Inter-Annotator Agreement in Sentence and Document Level Human Machine Translation Evaluation

Castilho (2020)

- Tasked translators to evaluate MT output on fluency, adequacy, ranking and error annotation.
- Tested two designs: i) a single score per isolated sentence; ii) a single score per entire document.

Human Evaluation, Semi-Automatic and Context

Research was made to investigate how much context was needed to be shown to translators.



The Inter-Annotator Agreement of this new design was satisfactory!

Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation

Castilho (2021)

- Study performed with larger group of translators.
- One additional design – individual sentences evaluated with access to the full document in addition to the random-sentence methodology and full documents.

Human Evaluation, Semi-Automatic and Context

- The Conference for Machine Translation (WMT), in 2019, employed document-level human assessment
- Direct Assessment (DA) (Graham et al., 2016) was employed to the task given to crowdworkers.
- Raters would evaluate randomly selected segments, consecutive segments in their original order, and entire texts.



Human Evaluation, Semi-Automatic and Context

- While in 2020, WMT modified the approach by expanding the contextual span to encompass full documents. sessment
- Direct Assessment (DA) (Graham et al., 2016) was employed to tasks given to crowdworkers.
- Raters would evaluate randomly selected segments, consecutive segments in their original order, and entire texts.



Human Evaluation, Semi-Automatic and Context

- Since 2022, WMT employs a source-based (called “bilingual”) DA + Scalar Quality Metrics (SQM). (Castilho and Knowles, 2024)
- SQM collects ratings of sentence level with the document provided as a context. (Kocmi et al., 2022)



Human Evaluation, Semi-Automatic and Context

Consistent Human Evaluation of Machine Translation across Language Pairs

Licht et al. (2022)

- Based on the metric Semantic Text Similarity (STS) (Agirre et al., 2012)
- Bigger emphasis on meaning (adequacy) rather than fluency.
- Bases itself on a scale from 1 to 5.
- Can be used with post-editing with critical errors.
- Could be useful with focus on context.



Human Evaluation, Semi-Automatic and Context

Test suites were also important. Several test suites were developed to test how MT systems would approach context.

The availability of test suites for document-level MT is still limited. (Castilho and Knowles, 2024)

Vojtechová et al. (2019)

Rysová et al. (2019)

Castilho et al. (2021)

And we will see test suites with more detail next lesson!



Automatic Evaluation and Context

Let us remember the widely used metrics in MT evaluation...

- BLEU (Papineni, 2002)
- METEOR (Banerjee, 2005)
- chrF (Popovic, 2015)
- TER (Snover, 2006)
- BERTScore (Zhang, 2019)



Automatic Evaluation and Context

Let us remember the widely used metrics in MT evaluation...

However, these metrics were designed for a sentence-level examination of MT.

- chrF (Popovic, 2015)
- TER (Snover, 2006)
- BERTScore (Zhang, 2019)



Automatic Evaluation and Context

One of the approaches involves merging sentences within a document with concatenation and applying the traditional metrics to evaluate it (Castilho and Knowles, 2024, p. 14)

These are the studies that attempted it.

Wong and Kit (2012)
Gong, Zhang and Zhou (2015)
Xiong et al. (2019)
Liu et al. (2020)
Saunders, Stahlberg and Byrne (2020)



Automatic Evaluation and Context

BlonDe

Jiang et al. (2022)

- Measures similarity in terms of **entity**, **tense**, **pronoun** and **discourse markers**.
- Can be combined with human annotation for BlonD+ – it compares the annotation with the inference calculated by the metric.



Automatic Evaluation and Context

Converting pretrained metrics into a document-level metric

Vernikos et al. (2022)

- An approach to extend BERTScore, Prism, COMET and COMET-QE to document-level – DOC-COMET
- It uses contextual embeddings from the MT output and human reference sentences. After generating the embeddings, the extra context is discarded before the metric is computed.



Automatic Evaluation and Context

DiscoScore

Zhao et al. (2023)

- Uses BERT to create DiscoScore, which is reference based. Has two variants: FocusDiff (DS-Focus) and SentGraph (DS-Sent)
- DS-Focus emphasises coherence. It tracks the focus across sentences of entities or nouns.



Automatic Evaluation and Context

DiscoScore

Zhao et al. (2023)

- Uses BERT to create DiscoScore, which is reference based. Has two variants: FocusDiff (DS-Focus) and SentGraph (DS-Sent)
- DS-Sent emphasises the interdependence of sentences. It is graph based, it checks how the MT output and reference match (or not).



Automatic Evaluation and Context

DiscoScore

Zhao et al. (2023)

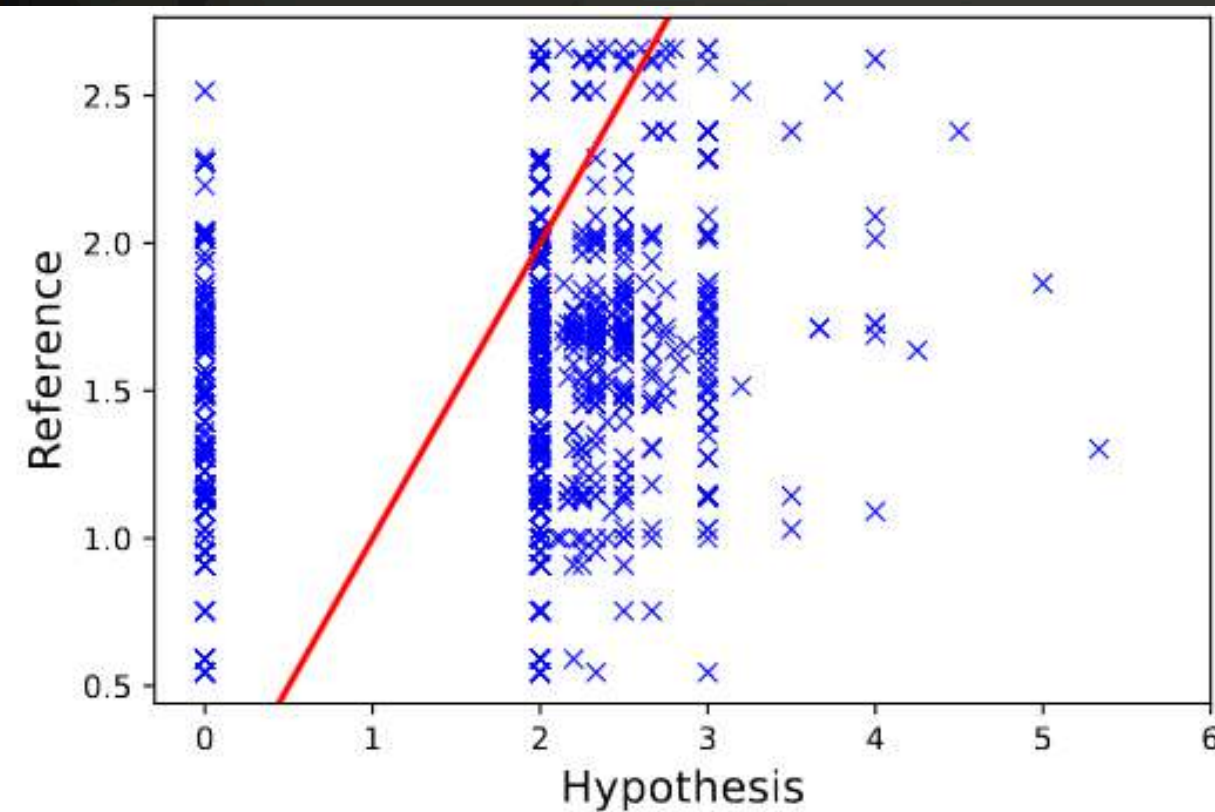


Figure 2: Scatter plot to display FREQ(hyp) (based on NN) on x-axis and FREQ(ref) on y-axis on SUMMEval. Each point contains two frequencies from a pair of hypothesis and reference. The points below the auxiliary line are the ones for which $\text{FREQ}(\text{hyp}) > \text{FREQ}(\text{ref})$.



Automatic Evaluation and Context

SLIDE

Raunak et al. (2023)

- Called Sliding Document Evaluator (SLIDE), it defines a window size to be calculated to COMET as a single input.



Automatic Evaluation and Context

SLIDE

Raunak et al. (2023)

#	docid	sentence
1	7759	There is no city in the high-risk...
2	7759	Iran's First Vice President...
3	7759	"Today there is neither concern...
4	doc0	I see, may I have your eReader...
5	doc0	To find your eReader's software...
6	doc0	1)Go to your Home screen.
7	doc0	2)Tap the More icon at the bottom...
8	doc0	3)Tap Settings.
9	doc0	4)Tap Device information.
10	doc0	5)Beside 'Software version', you'll...

Figure 1: SLIDE extraction for ($w = 4, s = 2$). The solid green boxes denote extracted chunks, which are then joined with a space and sent to COMET as a single unit. The dashed red boxes denote partial documents: a document that is too short (top), and a document remainder (bottom).



In this lecture you were able to...

Understand how MT development has adapted to include context in its evaluation

Understand how you can include context for human evaluation and automatic evaluation



Thank you!
See you next
class.

Questions?

Send an e-mail to
[joo.cavalheirocamargo2@
mail.dcu.ie](mailto:joo.cavalheirocamargo2@mail.dcu.ie)