Investigating Active Learning in Interactive Neural Machine Translation

Dr Rejwanul Haque
National College of Ireland, Dublin, Ireland
rejwanul.haque@ncirl.ie

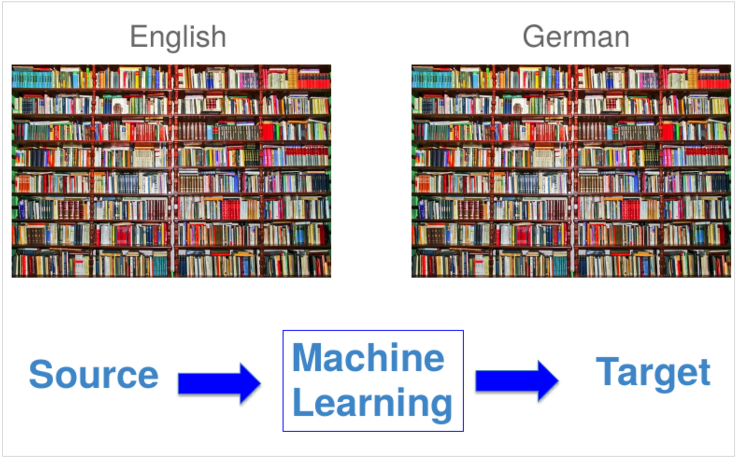
Outline of Presentation

National College of Ireland

Outline of Presentation

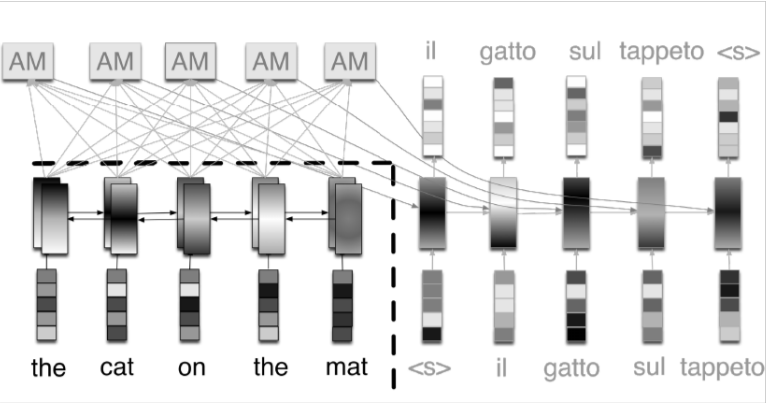
1. Overview on Machine Translation
2. Interactive MT
3. Interactive Adaptive MT
4. Our Approach

Machine Translation



Neural MT

Encoder-Decoder Architecture: Attention Model (Bahdunu et al., 2015)



Outline of Presentation

National College of Ireland

Interactive MT

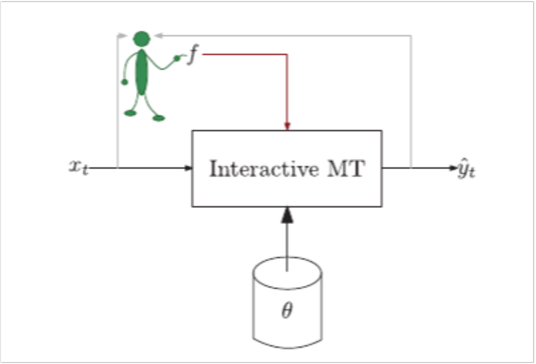


Figure source: Peris et al., 2017

Interactive MT

French (source sentence)

Nous décidons donc, citoyens, de prendre les choses en main

we decide therefore, citizens, to take things in hand.

we decide therefore, citizens, to take control **of things**

Interactive MT

- Why Interactive-predictive MT?
  - An effective way to improve **productivity gain** in translation
  - E.g. Lilt[1]

[1] https://lilt.com/

Outline of Presentation

National College of Ireland

Interactive Adaptive MT

- Interactive MT
  - MT model used in interactive platform: **static**
  - Same mistakes would repeat
  - May **negatively** impact **translation productivity**
- Interactive Adaptive MT
  - Ability of MT models to **change** in response to **customer's data**
  - In other words, online **customisation** of MT models
  - This can **counter** the risk of encountering the same mistakes in future

Interactive Adaptive MT

- Interactive MT
  - MT model used in interactive platform: **static**
  - Same mistakes would repeat
  - May **negatively** impact **translation productivity**
- Interactive Adaptive MT
  - Ability of MT models to **change** in response to **customer's data**
  - In other words, online **customisation** of MT models
  - This can **counter** the risk of encountering the same mistakes in future

Interactive Adaptive MT

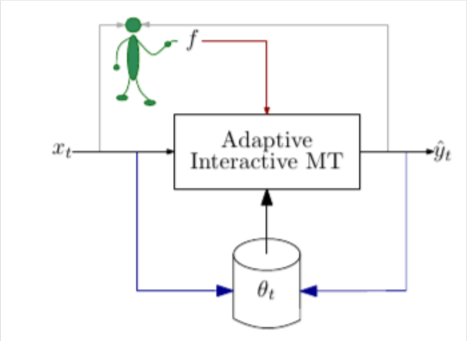


Figure source: Peris et al., 2019

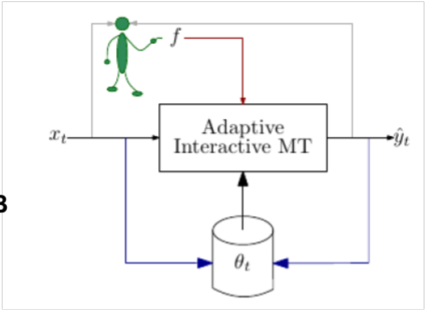
Outline of Presentation

Interactive Adaptive MT

S: input sentences for inference

B: block of sentences from **S**

C: chunk of sentences are sampled from **B**

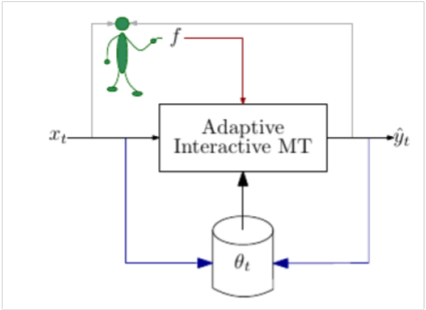


Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]

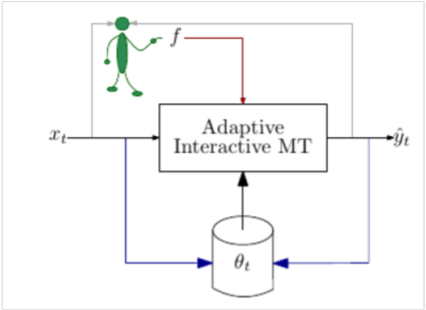


Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



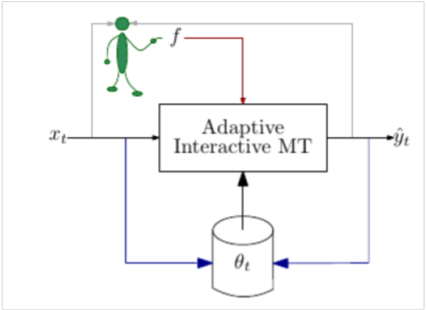
Uncertainty sampling: Labels those instances for which the model is **least certain** about the correct output to be generated

Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



Uncertainty sampling

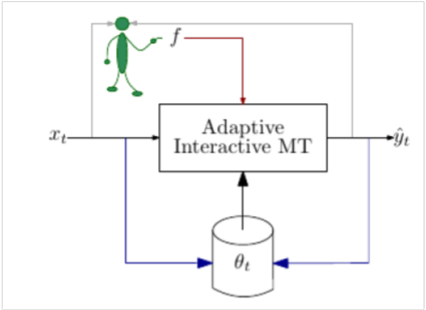
1. **Quality Estimation (QE)**
   - A process of evaluating the MT outputs without using gold-standard references.

Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



Uncertainty sampling

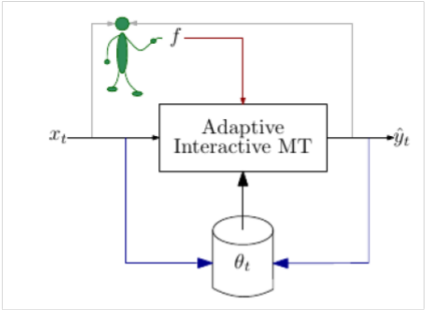
1. **Quality Estimation (QE)**
   - A process of evaluating the MT outputs without using gold-standard references.
   - E.g. translate 10,000 source sentences and using QE tool measure how good / bad the translations are
   - **Openkiwi** toolkit (Kepler et al., 2019)

Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



Uncertainty sampling

2. Round-Trip Translation (RTT)

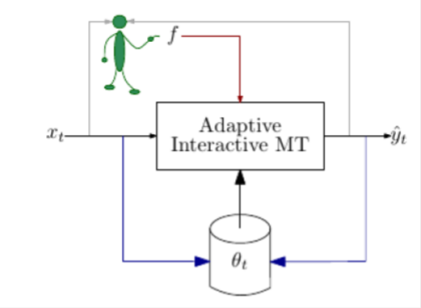
- A back-translation MT engine [target-to-source]
- **Similarity** between the **Source Sentence** and its **RTT**
- Sentences having the **least similarity scores** in **B** are sampled and **supervised** by the user

Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



Uncertainty sampling

2. Round-Trip Translation (RTT)

- **Similarity 1**
  - **Sentence-embedding (Sim$_{EMB}$)**
  - S-BERT (Ramires and Gurevych, 2019)

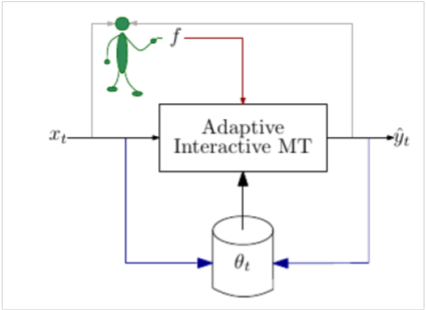
National College of Ireland

Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



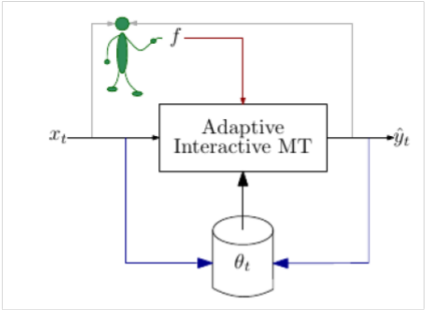
Uncertainty sampling

2. **Round-Trip Translation (RTT)**

- **Similarity 2**
  - **Edit distance (Sim_{fuzzy})**
  - *Fuzzywuzzy tool*

National College of Ireland

Interactive Adaptive MT



S: 50,000

B: 10,000

C: sample B; say 20% [2,000]

Uncertainty sampling

3. Named Entity (NEs) Count

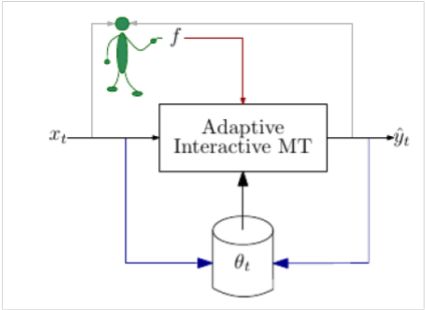
- Sentences having the **most number** of **NE** tokens in the block are considered as "**difficult to translate**" by the NMT model, and hence filtered for supervision

Interactive Adaptive MT

S: 50,000

B: 10,000

C: sample B; say 20% [2,000]



Uncertainty sampling

4. Query-by-committee (QbC)

- A **voted entropy function**: to calculate calculate the **highest disagreement** among the sampling techniques for a sample x
- Samples are then used for **human supervision**

Interactive Adaptive MT

- English-to-German
- German-to-English
- English-to-Hindi
- Spanish-to-English

	English-German	English-Spanish	English-Hindi
Train	1.26m (Europarl)	1.9m (Europarl)	1.6m (IITB corpus)
Dev	1,057 (Europarl)	2000 (Europarl)	599 (IITB corpus)
Testset	59,975 (News-Commentary)	51,613 (News-Commentary)	47,999 (ILCI corpus)

Interactive Adaptive MT

	BLEU
Baseline	23.28
RS	23.88
QE	24.02
Sim_{Fuzzy}	24.55
Sim_{EMB}	24.35
NE Counting	25.22
QbC	25.51

Table: English-to-German (20%)

National College of Ireland

Interactive Adaptive MT

	BLEU
Baseline	24.08
RS	25.19
Sim_{Fuzzy}	25.98
Sim_{EMB}	26.18
NE Counting	25.50
QbC	26.53

Table: German-to-English (20%)

Interactive Adaptive MT

	BLEU
Baseline	25.76
RS	25.84
Sim_{Fuzzy}	25.97
Sim_{EMB}	25.88
NE Counting	25.92
QbC	26.18

Table: English-to-Hindi (20%)

Interactive Adaptive MT

	BLEU
Baseline	38.76
RS	39.16
Sim_{Fuzzy}	39.28
Sim_{EMB}	39.74
NE Counting	39.43
QbC	39.78

Table: Spanish-to-English (20%)

Conclusions

- Explored the applicability of various **sampling techniques** in **active learning** to update NMT models in interactive-predictive translation platform
- Novel sampling methods
  - QE
  - RTT
  - NE Counting
- QbC with proposed sampling methods provided us best results across all LPs

National College of Ireland

Interactive Adaptive MT

- ***Investigating Active Learning in Interactive Neural Machine Translation***
- **Kamal Gupta** and **Asif Ekbal** from IIT Patna (India)
- **Pushpak Bhattacharyay** from IIT Bombay (India)
- **MT-Summit 2021**
- 16–20 August, Orlando, Florida, USA

National College of Ireland

Thanks for listening!

National College of Ireland