Empowering Human Translators via Interpretable Interactive Neural Machine Translation

Gabriele Sarti

XAI4Debugging Workshop @ NeurIPS ’21
A glimpse of the future track
My Background

MSc. thesis “Interpreting Neural Language Models for Linguistic Complexity Assessment”.

Focus: how NLMs encode linguistic complexity (gaze metrics, reader & writer-POV annotations), connection to linguistic properties (syntax, lexicon, etc.)

Intrinsic analysis of learned representations:

❖ Probing the presence of linguistic features before/after complexity fine-tuning, representational similarity (RSA, PWCCA) across tasks and layers.

→ Language properties encoded by NLMs are very task-dependent.
The Project and the InDeep Consortium

My project:

w/ Arianna Bisazza, Malvina Nissim & Grzegorz Chrupała

Improving interpretability for Neural MT from a user-centric perspective → practical benefits over theoretical insights.

Part of the NWO project “InDeep: Interpreting Deep Learning Models for Text and Sound”

→ XAI on NNs for text, speech & music applications.

❖ Develop new theoretical insights and methods.
❖ Apply shared practices to different domains (MT, fraud detection, disaster relief, speech classification, etc.)
Human Translation in 2021

Text sources → Translation Environment → Human Work

Text extraction
Human Translation in 2021

Text sources

Automated Work

Human Work

Editing Environment

Text extraction

Fuzzy matching

Pre-translations

Human Post-editor

Internet

Dictionaries

DOC

TXT

PDF

TMs & MT

G. Sarti, XAI4Debugging @ NeurIPS’21
Human Translation in 2021

Issues:

❖ Uncreative and uniform translations, lower reader engagement.
❖ Unpleasantness of post-editing.
❖ Better MT doesn’t always mean faster & better post-editing.

Text sources

DOC PDF

TXT

Text extraction

DOC

TXT

PDF

Pre-translations

Post-editor

Fuzzy matching

Internet

Dictionaries

Editing Environment

Human Work

Automated Work

Guerberof-Arenas & Toral 2021; Zouhar et al. 2021
Human Translation in 2021

Text sources

DOC PDF TXT

Text extraction

Fuzzy matching

Pre-translations

Automated Work

Internet

Dictionaries

TM & MT

Editing Environment

Human Post-editor

Human Work

Issues:

❖ Uncreative and uniform translations, lower reader engagement.
❖ Unpleasantness of post-editing.
❖ Better MT doesn’t always mean faster & better post-editing.

Interpretable and interactive MT systems can help!

Guerberof-Arenas & Toral 2021; Zouhar et al. 2021
From SMT to NMT

- **SMT** → specialized hand-crafted modules for lexical translation, reordering and target-side LM.
- **NMT** → end-to-end generation, + performance, + opaqueness.
From SMT to NMT

- **SMT** → specialized hand-crafted modules for lexical translation, reordering and target-side LM.
- **NMT** → end-to-end generation, + performance, + opaqueness.
- After 2017, transition from RNN-based to **transformer-based systems**, attention beneficial for long-range dependencies.
- **Encoder-decoder** models, mostly autoregressive generation conditioned on the prefix.

**Fig. 2:** The Transformer model architecture.
Source: Vaswani et al. 2017
From SMT to NMT

- **SMT** → specialized hand-crafted modules for lexical translation, reordering and target-side LM.

- **NMT** → end-to-end generation, + performance, + opaqueness.

- After 2017, transition from RNN-based to **transformer-based systems**, attention beneficial for long-range dependencies.

- **Encoder-decoder** models, mostly autoregressive generation conditioned on the prefix.

- **Issues:** Hallucinations and degeneration (Lee et al. 2018, Raunak et al. 2021, Chiang et al. 2021), failures on edge cases (terminology and inflections) for challenging language pairs. (Bisazza et al. 2021)
Towards Explaining NMT Models

**XNMT driven by end-user needs** → Mostly **post-hoc** methods (both **local** and **global**)

Goals:
Towards Explaining NMT Models

XNMT driven by end-user needs → Mostly post-hoc methods (both local and global)

Goals:

🔍 Pinpoint model weaknesses and challenging settings
Towards Explaining NMT Models

XNMT driven by end-user needs → Mostly post-hoc methods (both local and global)

Goals:
- Pinpoint model weaknesses and challenging settings
- Identify common patterns of misbehavior
Towards Explaining NMT Models

XNMT driven by end-user needs → Mostly post-hoc methods (both local and global)

Goals:
- Pinpoint model weaknesses and challenging settings
- Identify common patterns of misbehavior
- Trace problematic behaviors to training data
Towards Explaining NMT Models

**XNMT driven by end-user needs** → Mostly **post-hoc** methods (both **local** and **global**)

**Goals:**
- Pinpoint **model weaknesses and challenging settings**
- Identify **common patterns of misbehavior**
- Trace problematic behaviors to **training data**

→ **Useful accessible insights** for MT developers and translators.

💡 Especially useful for challenging and less-represented language pairs!
Open Questions in Seq2seq Explainability

What is a good explanation for NMT?

Faithfulness with model predictions vs. consistency with human intuition. (Ju et al. 2021, Madsen et al. 2021)

(Vafa et al. 2021)
Open Questions in Seq2seq Explainability

❓ What is a good explanation for NMT?

Faithfulness with model predictions vs. consistency with human intuition. (Ju et al. 2021, Madsen et al. 2021)

❓ What does “explaining a translation” mean?

XNLP approaches are classification-centric - how to extend them to autoregressive generation?

[CLS] amazing movie . some of the script writing could have been better

' s " the dead " is alluded to throughout the movie . beautiful scenery

highly recommend . [SEP]

POSITIVE
Open Questions in Seq2seq Explainability

❓ What is a good explanation for NMT?
   Faithfulness with model predictions vs. consistency with human intuition. (Ju et al. 2021, Madsen et al. 2021)

❓ What does “explaining a translation” means?
   XNLP approaches are classification-centric - how to extend them to autoregressive generation?

❓ How to faithfully attribute cross-attention? (Voita et al. 2021a, Voita et al. 2021b)
Open Questions in Seq2seq Explainability

❓ What is a good explanation for NMT?
   Faithfulness with model predictions vs. consistency with human intuition. (Ju et al. 2021, Madsen et al. 2021)

❓ What does “explaining a translation” means?
   XNLP approaches are classification-centric - how to extend them to autoregressive generation?

❓ How to faithfully attribute cross-attention? (Voita et al. 2021a, Voita et al. 2021b)

❓ How can we reduce the exponentially high cost of explaining sequences?
   Proxies might be needed, but they bear their own issues. (Vafa et al. 2021, Schwarzenberg et al. 2021)
Open Questions in Seq2seq Explainability

❓ What is a good explanation for NMT?
   Faithfulness with model predictions vs. consistency with human intuition. (Ju et al. 2021, Madsen et al. 2021)

❓ What does “explaining a translation” means?
   XNLP approaches are classification-centric - how to extend them to autoregressive generation?

❓ How to faithfully attribute cross-attention? (Voita et al. 2021a, Voita et al. 2021b)

❓ How can we reduce the exponentially high cost of explaining sequences?
   Proxies might be needed, but they bear their own issues. (Vafa et al. 2021, Schwarzenberg et al. 2021)

❓ How can we maximize the practical effectiveness of explanations?
   New methods of visualizing, calibrating and interacting with systems might be needed. (Ye et al. 2021)
Interactivity for Human-in-the-loop XNMT

A dose decrease may be required in patients who stop smoking.

Bij patiënten

Een dosis verlaging kan noodzakelijk voor patiënten die stoppen met roken.

(Coppers et al. 2018)
Interactivity for Human-in-the-loop XNMT

Idea: exploiting the virtuous circle of NMT systems and post-editors at work.

How to exploit behavioral metrics to minimize the interaction overhead?

(Coppers et al. 2018)
Current Work

- Evaluating post-editing of machine translated content across typologically-diverse language pairs
- Collect data, edit analysis to highlight error types.
Current Work

- Evaluating post-editing of machine translated content across typologically-diverse language pairs
- Collect data, edit analysis to highlight error types.

- Library for interpretability on sequence generation models (Hugging Face, Fairseq).
- Feature and instance attribution methods for codebase robustness and reusability.

Open to collaboration! 😊
Thank you for your attention!

university of groningen / center for language and cognition / Dutch Research Council

@gsarti_
gsarti
g.sarti@rug.nl
gsarti.com
interpretingdl.github.io
References

References