Adaptive CAT-embedded MT for low-memory, low-compute end-user devices

Marek Sabo
STAR AG
Wiesholz 35
CH-8262 Ramsen
marek.sabo@star-group.net

Abstract

We present ACATMT, a compact bilingual encoder-decoder NMT system for English and Swedish, designed for professional computer-assisted translation (CAT) tools. It runs on-device in ONNX format, under 1 GB of RAM with no GPU needed, and features real-time post-edit based terminology adaptation. It also supports translation memory conditioning via decoder prefilling. Evaluation on 5,021 technical segments unseen during training shows significant improvements in COMET and BLEU when using glossaries.

1 Introduction

Adapting MT in realtime by updating model weights after each segment is impractical on average translator's hardware. Constrained decoding and source augmentation still require explicit glossary management from the translator. ACATMT instead detects terminology changes automatically from post-edits, storing updated pairs in a glossary that is automatically applied to subsequent translations.

2 System Description

2.1 Architecture

ACATMT is an encoder-decoder model initialized from mT5-base (Xue et al., 2020) with tied embeddings. A unigram tokenizer trained on English and Swedish was used to reduce the vocabulary to 50,000 tokens and the original embeddings were aligned to the new tokenizer where possible. The model was trained on bilingual masked span prediction, then fine-tuned on translations with and without bilingual term entries. If glossaries are provided, special tokens signal the start of a term pair and the start of the target term. In training, the glossaries were generated from parallel corpora using the Stanza NLP pipeline's syntactic parsing and our own rules to extract noun phrase candidates. Target term candidates were aligned using LaBSE. To better handle incomplete input from human's typing during post-editing, a later training stage introduced more varied unigram sampling. Glossary extraction uses another special token. For this task we used a few thousand bilingual texts from which DeepSeek API extracted term pairs. Translation memories are supported via decoder prefilling with TM translations provided as decoder context before generation and stripped from the displayed output. Exporting and quantizing in ONNX enabled a faster generation (tokens/second): 16 with max battery saving, 21 with a balanced setting, 33 with max performance on battery, and 46 when plugged in. Tested on a laptop with Intel Core i5-1135G7 (no GPU) and 16 GB RAM.

2.2 Interactive Post-Editing

Our localhost browser demo interface streams output token by token from the on-device Python server. The translator can click any word mid-generation to interrupt and request alternatives from that position, optionally typing the first characters of a desired word to filter suggestions. Pressing *Enter* accepts the typed input and the model regenerates the translation from that point; while *Escape* accepts the edit locally without regeneration, suitable for minor corrections.

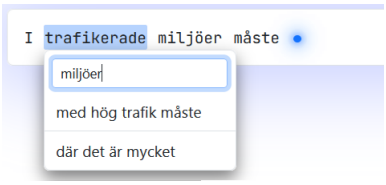

Figure 1: ACATMT interrupted when a user clicks on a word to rephrase it, types the beginning of an alternative, and ACATMT suggests completions in a dropdown.

2.3 Implicit Terminology Adaptation

The system's most distinctive feature operates without any deliberate user action. After translating a segment, the model extracts a bilingual glossary from its own output. If the translator post-edits the translation, the extraction is re-run. By comparing the two extraction results, the system detects changes in terminology and saves updated term pairs into a post-edit-based glossary, which is separate from any human-curated glossaries. If a new source text contains any captured term pairs, the system supplies them to the model. Term extraction is a trained behaviour, which works by providing both the source and the target to the encoder with a dedicated token for glossary generation.

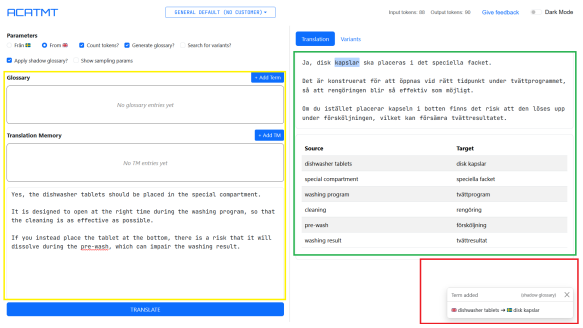

Figure 2: Source text (yellow) with optional user provided terms or translation memory, ACATMT output post-edited by the user (green) and the automatically extracted glossary. Terms automatically captured into the glossary appear in a toast notification (red)

3 Evaluation

Glossary usage is automatically evaluated on 5,021 segments from a customer's reference translation in medical technology. No human post-editors were involved. Segments where the source consisted solely of glossary terms were excluded from the evaluation. The customer data was not seen during training. We compare four ACATMT conditions against Helsinki-NLP's opus-mt-en-sv as a strong bilingual baseline. Terms used in the evaluation were derived from the same reference material using ACATMT's glossary extraction mode. We compare MT without glossary (vanilla), with the glossary, and with decoder prefilling using terms, and report the respective COMET (wmt22-comet-da) and sacreBLEU scores. When terminology is provided, ACATMT gains significant improvements in BLEU and COMET scores and outperforms Helsinki using decoder prefilling with terms.

Condition	COMET	BLEU
Helsinki opus-mt-en-sv		
Vanilla	0.8368	35.68
+ Decoder pref. with terms	0.8436	37.29
ACATMT		
Vanilla	0.8195	35.23
+ Glossary mode	**0.8719**	**45.19**
+ Glossary & pref.	0.8675	45.96
+ Decoder pref. with terms	**0.8733**	**47.57**

Table 1: Mean COMET and BLEU on 5,021 customer segments.

A formal evaluation of the implicit adaptation loop, measuring term capture rate and apply rate across post-edits in real human-facing interactions, remains as future work.

4 Pricing, licensing, and availability

ACATMT is intended for integration into STAR Transit, a commercially available CAT tool by STAR AG. Licensing and pricing are pending; the codebase is proprietary and cannot be released.

5 Discussion and Future Work

When the user clicks on a word to display alternatives, the system rarely surfaces source-language suggestions. If the model doesn't apply the supplied glossary term, the term often appears among the suggested alternatives.

Some terms are used very infrequently in customer texts. Even in the model variants that have been fine-tuned on specific customer's data, glossary input can occasionally further help the model.

We think that small, adaptive, privacy-preserving models represent a compelling direction for professional post-editing workflows, and hope ACATMT serves as a useful reference for the community.

References

Xue, Linting and others. 2020. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *arXiv preprint arXiv:2010.11934*.