

A RAG-Based Context-Aware Term Translation & Gloss Generation Engine for CAT Tools

Overview & Motivation

Computer-assisted translation (CAT) tools increasingly rely on AI to support terminology management and translation decisions. However, existing term lookup systems are largely based on static dictionaries and translation memories that fail to capture the contextual meaning of ambiguous or polysemous terms—words or phrases whose meaning varies across domains and discourse contexts. This limitation increases cognitive load, disrupts translator focus, and contributes to inconsistent or inaccurate translations.

We propose a **Context-Aware Term Translation & Gloss Generation Engine** built on a **Retrieval-Augmented Generation (RAG)** architecture. The system dynamically retrieves evidence from multiple linguistic sources using a **hybrid retrieval strategy** that blends traditional keyword-based search with semantic (embedding-based) search, and synthesizes **contextualized glosses with citations** tailored to the specific usage of a term within a translation segment. To support translator workflow, the system employs a **progressive disclosure user interface** that minimizes distraction while enabling deeper semantic inspection when needed.

Research Foundation

Contextual Gloss Generation

Prior work demonstrates that word meaning can be modeled through gloss generation conditioned on local and global context rather than relying on fixed sense inventories. Ishiwatari et al. show that contextual cues enable generation of descriptions for unknown or ambiguous phrases directly from usage context, providing a robust mechanism for resolving polysemy.

Query + Context Conditioning

LitMind Dictionary introduces a framework in which both a target term and its surrounding context jointly condition the gloss generator. This aligns with CAT workflows in which translators encounter terms embedded in discourse and require meaning tailored to the immediate translation segment.

Glosses as Interpretable Semantic Representations

Recent work on interpretable word sense representations demonstrates that generated glosses act as human-readable semantic explanations, bridging neural representations and translator decision-making.

Beyond Static Sense Inventories

The Generatory project shows that neural gloss generation can replace rigid sense inventories by producing glosses for multi-word expressions and rare or evolving terminology, supporting dynamic language use in technical and press domains.

System Architecture: Retrieval-Augmented Generation (RAG)

The engine combines federated retrieval with context-conditioned generation:

1. **Input Capture**

The system receives a selected term or phrase and its surrounding sentence or paragraph context from the CAT tool.

2. **Federated Retrieval Layer (R in RAG)**

Parallel queries are issued to:

- Monolingual and bilingual dictionaries
- Translation memories and aligned corpora
- Press articles and domain-specific sources
- Web-scale linguistic resources

3. Retrieval is performed using a **hybrid strategy** that blends:

- **Keyword-based search** (exact matches, morphological variants)
- **Semantic search** (embedding-based similarity over term usage and context)

4. This hybrid approach improves recall and precision for polysemous and domain-specific terminology.

5. **Context Encoder & Gloss Generator (G in RAG)**

A transformer-based model jointly encodes:

- The term
- Local context
- Retrieved evidence

6. It generates multiple candidate **contextual glosses** optimized to disambiguate the polysemous meaning in the given usage context.

7. **Gloss Ranking & Synthesis**

Candidate glosses are ranked by contextual relevance and synthesized with aligned translation candidates and supporting citations.

8. **Citation & Transparency Layer**

Each generated gloss and translation is linked to dictionary entries, corpus examples, and press excerpts to ensure traceability and trust.

Progressive Disclosure User Interface

To reduce cognitive load and preserve translator focus, the UI employs a **three-layer progressive disclosure model**:

Layer 1 – Ranked AI Gloss & Translation Shortlist (Default View)

The primary interface displays a short ranked list (typically **three**) of the most likely:

- Contextualized glosses
- Corresponding translation candidates

This lightweight shortlist enables translators to rapidly resolve confusing terms while staying focused on the translation task, without navigating long dictionary entries or external resources.

Layer 2 – Cited Evidence View (On Demand)

If the translator seeks greater confidence, they can expand any candidate to view:

- Extracted dictionary glosses
- Corpus and press examples
- Translation memory matches

Each item is presented with inline citations showing the source and usage context.

Layer 3 – Full Federated Results (Deep Inspection)

For complex or terminologically sensitive cases, the translator can access the complete federated retrieval output, including:

- All retrieved source documents
- Ranked gloss candidates
- Usage distributions across domains
- Alternative semantic interpretations

This level supports expert validation, terminology management, and QA workflows.

This progressive disclosure strategy aligns with cognitive load theory by presenting essential semantic options first and revealing complexity only when explicitly requested.

Key Capabilities

- **Polysemy Resolution:** Disambiguates context-dependent term meanings through gloss generation.
 - **Hybrid Retrieval:** Combines keyword and semantic search within RAG.
 - **Explainable AI:** Produces ranked, human-readable glosses with citations.
 - **Workflow-Aware UX:** Minimizes interruption through progressive disclosure.
 - **Domain Adaptability:** Supports specialized translation domains.
 - **Continuous Learning:** Integrates user feedback to refine retrieval and ranking.
 - **Seamless Glossary Integration** The system is designed to make glossary creation and maintenance frictionless for translators. At every stage of interaction, users can promote a selected gloss–translation pair directly into their project glossary with a single action. From the ranked shortlist view, translators can pin the preferred contextual gloss and corresponding translation with one click, automatically creating a validated glossary entry linked to its source context and supporting evidence
-

Conclusion

By integrating **Retrieval-Augmented Generation**, **hybrid keyword and semantic retrieval**, and a **progressive disclosure interface with ranked gloss–translation pairs**, this system transforms terminology lookup into an intelligent semantic assistant. It addresses the challenge of **polysemous term interpretation** while reducing cognitive load and maintaining translator flow. The result is a CAT tool component that combines AI interpretability, linguistic rigor, and user-centered design to improve translation accuracy, transparency, and productivity.