EA-MT: Entity-Aware Machine Translation

Ina Ocelli, Lola Kovalski

1 Background:

A named entity is anything that can be referred to with a proper name, such as people, organizations, locations, creative works, dates and times, among other categories. Recognizing named entities is a crucial, albeit challenging, step in a variety of natural language processing tasks.

Current machine translation (MT) systems underperform when translating named entities, especially in cases requiring precise semantic and contextual understanding. By incorporating an entity-aware framework into MT, which leverages named entity recognition (NER) and alignment techniques, we aim to improve translation quality.

2 Hypotheses and Goals:

We expect that an entity aware MT system will improve overall translation accuracy. The primary goal of this project is to (1) implement an entity-aware translation model and (2) demonstrate its effectiveness by measuring improvements in both entity translation accuracy and overall BLEU/COMET scores compared to a baseline transformer base MT system.[1]

Goal 1: Develop a translation model that improves handling of rare, ambiguous, or culturally specific named entities.

Goal 2: Benchmark the performance of entity-aware machine translation models against standard translation systems

3 Sources of Data:

The training dataset for this project will be provided by SemEval-2025, for the Entity-Aware Machine Translation (EA-MT) task. This dataset will include diverse examples of sentences with named entities, designed to challenge machine translation systems.

[1]Note that we do not expect to see a significant improvement in BLEU/COMET scores as incorrect translations tend to hurt human readability rather than scores.

Unfortunately, we will not receive official validation and test sets in time for our experiments. To address this, we will either:

1. Split the Training Dataset: Create validation and test sets by dividing the training dataset into smaller subsets for evaluation.
2. Use external datasets: After receiving the training data, we can try to find compatible datasets for validation/test set, e.g. Mintaka, MultiNERD

4 Rough Plan of Analysis:

4.1 Identify Named Entities (NER)

Objective: Use a pretrained Named Entity Recognition (NER) model to identify entities in the source text.

- Use the **tomaarsen/span-marker-mbert-base-multinerd** NER model trained on the **MultiNERD dataset**.
- Label named entities with categories (e.g., PERSON, ORGANIZATION, LOCATION) for better context-aware translation.

4.2 Mask Identified Named Entities with Placeholders

Objective: Replace identified named entities with placeholders to isolate their impact during translation.

- Replace entities with unique tags such as `[ENTITY_PERSON_1]`, `[ENTITY_LOC_1]`.
- Ensure the mapping between original entities and placeholders is preserved for later translation and integration into original text.

4.3 Fine-Tune an Existing MT/NMT Model on the Masked Dataset

Objective: Train a neural machine translation (NMT) model to handle placeholder-based inputs and focus on sentence structure.

- Use a state-of-the-art MT model, such as **mBART**.

4.4 Reintegrate Original Entities into Translated Text

Objective: Map placeholders back to the translated text while deciding how to handle each entity.

- **Translate the Entity:** For culturally significant entities (e.g., "The Great Gatsby" to "Il grande Gatsby").
- **Keep as Is:** For universal names (e.g., "iPhone").
- How do we decide to translate a named entity? When do we keep the same? When do we change language? When do we use dictionary approach?

4.5 Evaluate the Impact of Named Entity Masking and Translation

Objective: Test the efficacy of the masking and reintegration approach.

- Run translation experiments with and without entity-aware handling.
- Use metrics such as:
- **Entity-Level Precision & Recall:** To measure the accuracy of named entity translations.
- Conduct an ablation study to understand the contribution masking+reintegration.

5 Anticipated sources of complication:

- **Ambiguity of Segmentation:** Difficulty in identifying precise boundaries of named entities, particularly in cases with overlapping or nested entities.
- **Type Ambiguity:** Challenges in determining the correct category of an entity (e.g., "Apple" as a company vs. a fruit) based on context, impacting recognition and translation accuracy.

6 Papers Sources:

- https://aclanthology.org/2024.findings-acl.691.pdf
- https://aclanthology.org/2022.findings-naacl.60.pdf
- https://arxiv.org/pdf/2305.07360
- https://aclanthology.org/W03-2201.pdf
- https://aclanthology.org/2022.eamt-1.17.pdf