Automatic Post-Editing of MT Output Using Large Language Models

AMTA, September 2022

Albert Llorens
Blanca Vidal
Why Glossaries Matter in the Translation Business

**Accuracy** of translation
- Not using the industry or company specific translation of a term may lead to inaccurate translations

Glossaries ensure the **consistency** of the translation of key terms, both within and across documents

**Client glossaries typically include**
- Product names
- Company names
- Ambiguous words
- Abbreviations
- Borrowed words
- Terminology (specialized industry/field terms)
Glossaries and Machine Translation

Pre-translation with NMT is widely used in the Translation business

NMT is a black box to users, developers, and researchers

NMT models can be trained, but not forced

Glossaries are more about “forcing” than “training”

It is not straightforward to “force” a NMT system to translate terms according to a glossary
ULG Use Case

- ULG main NMT provider handles glossaries by doing a brute force find-and-replace operation.
- This approach guarantees close to 100% consistency of machine translations with glossary translations.
- But it has negative side effects in the translation quality, mostly in:
  - Grammatical agreement (gender, number, case)
  - Word order
- ULG Glossaries are used in MT in two different ways:
  - As bilingual dictionaries that can be referenced at request level with a category id
  - At runtime, by annotating the terms that require a specific translation with xml tags in the input string of the request.
Proposed Solution

Keep the current workflow: translation with annotated input

Add a post-processing step where the grammar and word order errors are fixed

Use a general-purpose large language model, like OpenAI GPT-3, to do the post-editing of the NMT output
About OpenAI API and GPT-3 Models

- The OpenAI API can be applied to virtually any task that involves understanding or generating natural language.
- The API is powered by GPT-3, a set of models with different capabilities.
- The API requests are headed by a prompt that describes the task to be done by the model.

- The prompts used in the experiment are:
  - “Corregir la gramática en español”
  - “Corregir el orden de las palabras en español”
  - “Traducir al español con el glosario {}={}:\n\n.”

- The models used in the experiment are:
  - text-davinci-edit-001
  - text-davinci-002

- The endpoints used in the experiment are:
  - /completions: input text as a prompt, and get a text completion that matches the prompt instruction.
  - /edits: change existing text via a prompt, instead of completing it.
Experiment Objectives

The experiment we implemented wanted to check the following points:

1. Check if GPT-3 can be used as an MT engine
2. Check if GPT-3 can be used as an Automated Post-Editor
3. Check if GPT-3 can improve its own Post-Editing by requesting word order correction
4. Check if GPT-3 can be used as an MT engine using Glossary annotations
Test Data Selection

- Translation Memory and glossary of a ULG client
- Both TM and glossary must be big enough, and TM must be highly consistent with the glossary
- Choice of languages: English to German and to Spanish
- Data size: ~500k TM segments and ~600 glossary terms
Test Data Preparation

1. Restricting the set to **English-Spanish**

2. **Filtering** the data set by
   - Lemmatizing source and target segments
   - Removing all segments that don’t match any pair in the glossary

3. Data size after preparation: ~2,000 segments

4. **Annotating** source segments with glossary translations. Examples:
   - Side view `<term trans='disco de ruptura'>rupture disk</term>`
   - Sensor `<term trans='procesador central extendido'>extended core processor</term>`

5. Selecting a **sample of 250 segments** from the test data
### Experiment Requests and Outputs

**Tasks, requests, prompts, outputs**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ULG MT</strong></td>
<td>Source file without annotation sent to ULG MT</td>
<td>Source file with Glossary annotation sent to ULG MT</td>
<td>Output of (2) sent to GPT-3 ‘edits’ endpoint with prompt &quot;Corregir la gramática en español&quot; [&quot;temperature&quot;: 0, engine=&quot;text-davinci-edit-001&quot;]</td>
<td>Output of 3 sent to GPT-3 ‘edits’ endpoint with prompt &quot;Corregir el orden de las palabras en español&quot; [&quot;temperature&quot;: 0, engine=&quot;text-davinci-edit-001&quot;]</td>
<td>Source file without annotation sent to GPT-3 ‘completions’ endpoint with prompt &quot;Traducir al español&quot; [&quot;temperature&quot;: 0, engine=&quot;text-davinci-002&quot;]</td>
<td>Source file with Glossary annotation sent to GPT-3 ‘completions’ endpoint with prompt &quot;Traducir al español con el glosario {source term}={target term}&quot; [&quot;temperature&quot;: 0, engine=&quot;text-davinci-002&quot;]</td>
</tr>
</tbody>
</table>
## Experiment Results

### BLEU, TER and Terminology Consistency scores

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TER</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULG MT without glossary</td>
<td>57.6</td>
<td>26.0</td>
<td>87.12%</td>
</tr>
<tr>
<td>ULG MT with glossary</td>
<td>54.8</td>
<td>27.4</td>
<td>99.24%</td>
</tr>
<tr>
<td>GPT-3 PE grammar</td>
<td>54.8</td>
<td>30.5</td>
<td>98.11%</td>
</tr>
<tr>
<td>GPT-3 PE grammar and order</td>
<td>49.6</td>
<td>34.6</td>
<td>83.71%</td>
</tr>
<tr>
<td>GPT-3 MT without glossary</td>
<td>49.7</td>
<td>35.8</td>
<td>75.00%</td>
</tr>
<tr>
<td>GPT-3 MT with glossary</td>
<td>43.2</td>
<td>54.4</td>
<td>85.61%</td>
</tr>
</tbody>
</table>
Results: Output Scores

BLEU, TER and Terminology consistency scores

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>TER</th>
<th>Term Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>ULG MT</td>
<td>57.6</td>
<td>26.0</td>
<td></td>
</tr>
<tr>
<td>ULG MT Glo</td>
<td>54.8</td>
<td>27.4</td>
<td></td>
</tr>
<tr>
<td>GPT-3 MT Glo</td>
<td>54.8</td>
<td>30.5</td>
<td></td>
</tr>
<tr>
<td>GPT-3 MT Glo Edit</td>
<td>49.6</td>
<td>34.6</td>
<td></td>
</tr>
<tr>
<td>GPT-3 MT Glo Edit Order</td>
<td>49.7</td>
<td>35.8</td>
<td></td>
</tr>
<tr>
<td>GPT-3 MT</td>
<td>43.2</td>
<td>34.4</td>
<td></td>
</tr>
</tbody>
</table>

Check if GPT-3 can be used as an MT engine

RESULTS

• Most outputs are either similar or identical to the ones of our current models.
• GPT-3 is less conservative in the preservation of the source.
• Makes changes that need to be contrasted with the input.
• Need to fix GPT-3’s addition of extra dots (and blanks)
Check if GPT-3 can be used as an Automated Post-Editor

**RESULTS**
- GPT-3 managed to fix gender agreement problems
- It also fixed number agreement
- Addition of determiners and prepositions added fluency
- Addition of other terms also added fluency

![Graph showing type of changes performed](image)

- Fixed gender agreement
- Fixed number
- Added det
- Added prep
- Added term
- Fixed capitalization
- Changed prep
- Changed term
- Changed verb form
- Changed ABB
- Changed NLU
- Changed term (glossary)
- Changed word order
- Repeated one term
## Results: Comparative Analysis

**Examples of Improved Outputs**

<table>
<thead>
<tr>
<th>Example</th>
<th>Improved Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>78</td>
<td>Two points with Product A and Product B tool.</td>
</tr>
<tr>
<td>18</td>
<td>For Model 500, output voltage is 15 VDC ±1.5% with a 2.2 kΩ pull-up resistor (resistencia pull-up).</td>
</tr>
<tr>
<td>95</td>
<td>If you are installing a cable gland at the I/O barrier, a standard cable gland with a sealing washer, back-up washer, and lock nut is required.</td>
</tr>
<tr>
<td>241</td>
<td>One exception is nitric acid (ácido nítrico), for which 304 stainless steel has better corrosion resistance.</td>
</tr>
<tr>
<td>104</td>
<td>Inserted port (puerto insertado) wire protection.</td>
</tr>
<tr>
<td>65</td>
<td>Top an external temperature device (dispositivo de temperatura externo) with a high temperature signal.</td>
</tr>
<tr>
<td>101</td>
<td>If you are using an infrared port (puerto infrarrojo) for communication with the infrared port (ase).</td>
</tr>
</tbody>
</table>
## Results: Comparative Analysis

### Changed glossary term

<table>
<thead>
<tr>
<th>ID</th>
<th>Original Term</th>
<th>Corrected Term</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>208</td>
<td>config ID custom ESP-configureables personalizadas</td>
<td>ESP-configurable personalizadas</td>
<td>ESP personalizadas configureables</td>
</tr>
<tr>
<td>209</td>
<td>config ID default ESP-configurable personalizadas</td>
<td>ESP-configurable predefinidas</td>
<td>ESP-configurable por defecto</td>
</tr>
</tbody>
</table>

### Acceptable worsening

<table>
<thead>
<tr>
<th>ID</th>
<th>Original Term</th>
<th>Corrected Term</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Minimum distance between booster amplifier (amplificador booster) and sensors (sensores)</td>
<td>La distancia mínima entre el amplificador booster remoto y el sensor es de 30 m</td>
<td>La distancia mínima entre el amplificador booster remoto y el sensor es de 10 metros</td>
</tr>
<tr>
<td>02</td>
<td>External use of an external pull-up resistor (resistencia pull-up) and source are required</td>
<td>Existe un signo que se requieren resistencia pull-up y una fuente externa</td>
<td>Existe un signo que se requieren resistencia pull-up y una fuente externa</td>
</tr>
<tr>
<td>03</td>
<td>Engineering unit of measurement/units of measurement: unit of measure (unidad de medida)</td>
<td>unidad de ingeniería de medida, unidad de ingeniería de medida</td>
<td>unidad de ingeniería de medida, unidad de ingeniería de medida</td>
</tr>
<tr>
<td>04</td>
<td>For F-Serie sensors with a pressure booster in extended case processing/pressurized central extended case (extender caso)</td>
<td>Para los sensores de la serie F con una caja de conexiones con un procesador central en caso extendido, la temperatura ambiente máxima es de 60°C (94°F)</td>
<td>Para los sensores de la serie F con una caja de conexiones con un procesador central extendido, la temperatura ambiente máxima es de 60°C (94°F)</td>
</tr>
</tbody>
</table>
Check if GPT-3 can improve its own Post-Editing by requesting word order correction

RESULTS

ORDER CHANGES
- Improve: apposition of proper names
- Similar: order of exchangeable noun modifiers
- Worsening: change order in glossary terms, change in term meaning, change of the translation of a glossary term

REST OF CHANGES
- Improve: added fluency (adding determiners)
- Worsening: adds a duplicated term
Results: Comparative Analysis

### Examples of Improved Outputs

<table>
<thead>
<tr>
<th>Example</th>
<th>Improved Output</th>
<th>Example</th>
<th>Improved Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 01</td>
<td>Product 02</td>
<td>Product 01 connector in M20 housing</td>
<td>Product 01 connector</td>
</tr>
<tr>
<td>Product 02</td>
<td>Product 03</td>
<td>Connector Product 02 of S pines en carcasa M20</td>
<td>Connector Product 02 of S pines en carcasa M20</td>
</tr>
</tbody>
</table>

### Examples of Wrong Corrections

1. **Example:** Using a FreshnetNet (FreskenNet) test, there are 794 methods available for localization and inventory control.

   **Corrección:** Utilizando una herramienta de Localización, hay tres métodos disponibles para el localizador y el control de inventario.

2. **Example:**

<table>
<thead>
<tr>
<th>Example</th>
<th>Improved Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>DIN rail enclosure (cubierta de carril DIN)</td>
</tr>
<tr>
<td>14</td>
<td>core processor entity parameter (parámetro de entidad del procesador central)</td>
</tr>
<tr>
<td>15</td>
<td>Construction Identification Code (código de identificación de construcción)</td>
</tr>
<tr>
<td>16</td>
<td>Side view with rupture disk (Vista lateral con disco de ruptura)</td>
</tr>
<tr>
<td>17</td>
<td>engineering unit of measurement (unidad de ingeniería de medida)</td>
</tr>
</tbody>
</table>

### United Language Group

Check if GPT-3 can be used as an MT engine using Glossary annotations

RESULTS

- ULG MT Glossary gets applied in all segments (100%)
- GPT-3 is only applied in 76% of the segments due to different reasons
- In both cases there are side effects already found in previous tests
Conclusions

1. Using GPT-3 as an MT Engine shows interesting improvements in style and readability, but important "creativity" problems.

2. Using GPT-3 for Post-Editing shows very promising results, with a clear improvement in the outputs.

3. Using GPT-3 to fix Word Order problems results in many unnecessary and sometimes incorrect changes.

4. Using GPT-3 as an MT Engine with Glossary annotations results in many "creativity" problems and consistency errors.
Future Work

- Prompt engineering
  - Prompt language makes a difference
  - Adding examples, find the most appropriate wording of the instructions

- Adjusting request parameters
  - temperature: lower temperature, less risks
  - top_p: select tokens with top probability mass

- Fine-tuning the GPT-3 models with ULG data

- Using logprobs and beam search with higher temperature to filter undesired responses
Credits

Alonso, Juan Alberto
juan.alonso@ulgroup.com

Llorens, Albert
albert.llorens@ulgroup.com

Madan, Mehul
mehul.madan@ulgroup.com

Vidal, Blanca
blanca.vidal@ulgroup.com
Thank you.