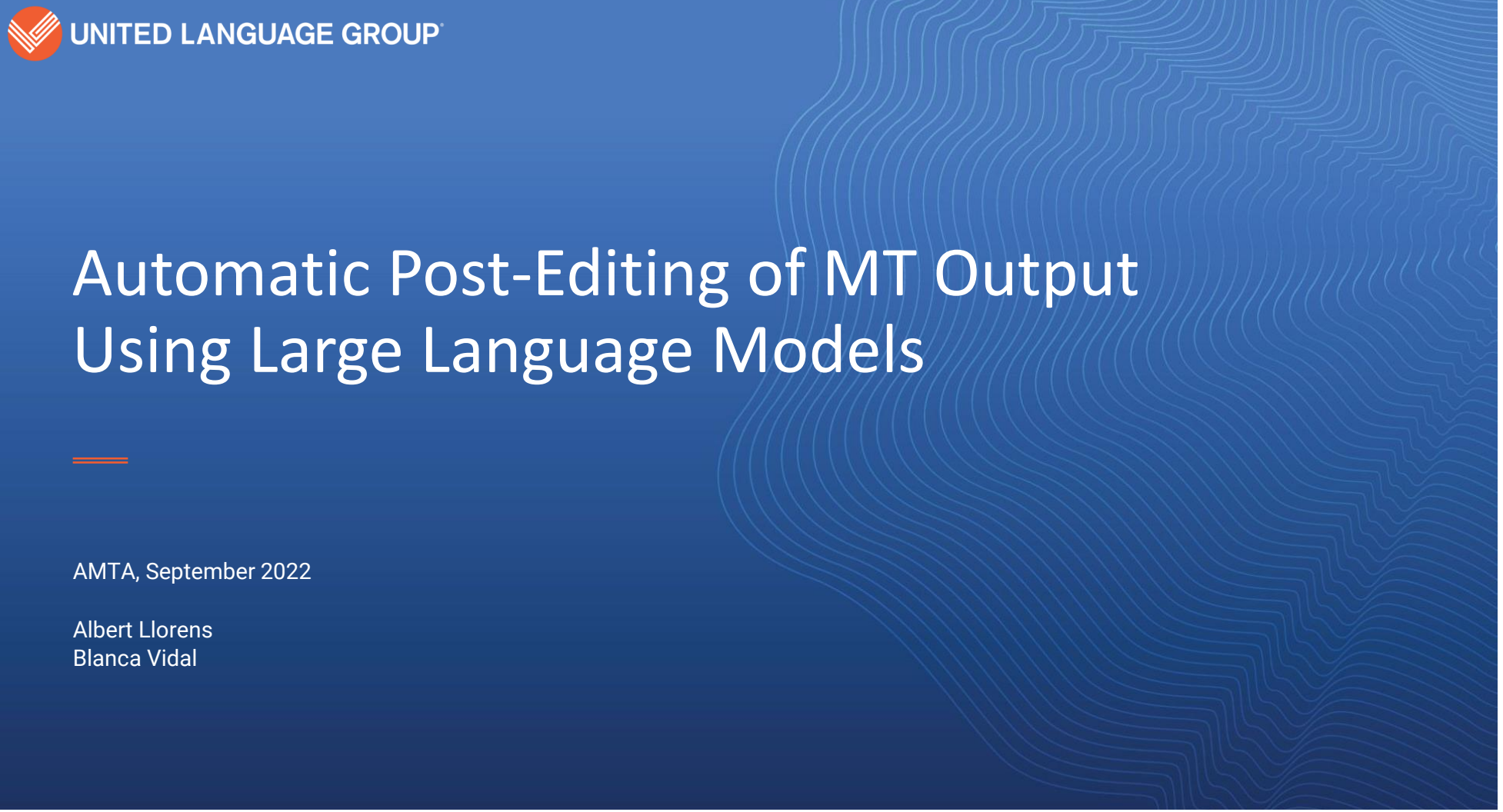

UNITED LANGUAGE GROUP
Automatic Post-Editing of MT Output
Using Large Language Models
AMTA, September 2022
Albert Llorens
Blanca Vidal

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)

UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

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

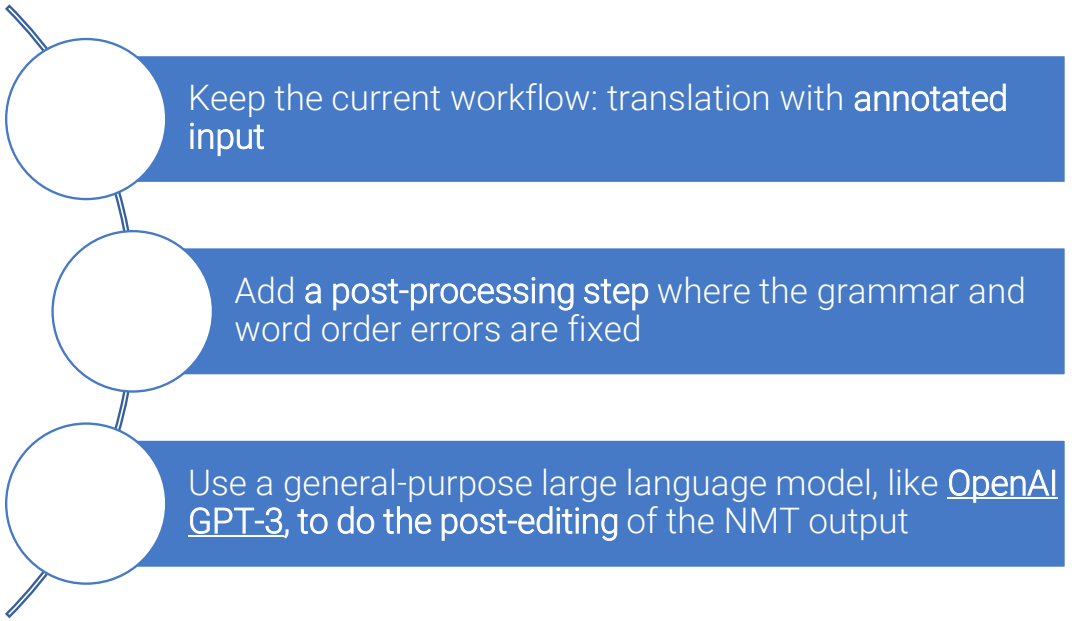
UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

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



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

UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Experiment Objectives

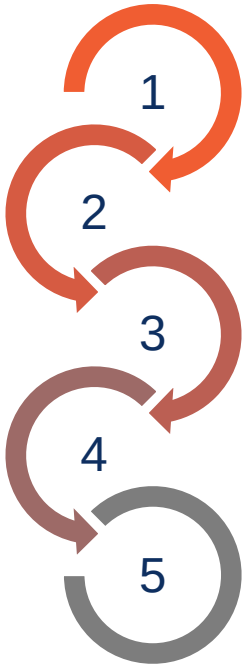
The experiment we implemented wanted to check the following points:

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



- Restricting the set to English-Spanish
- Filtering the data set by
  - Lemmatizing source and target segments
  - Removing all segments that don't match any pair in the glossary
- Data size after preparation: ~2,000 segments
- 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>
- Selecting a sample of 250 segments from the test data

Experiment Requests and Outputs

		Tasks, requests, prompts, outputs
1	ULG MT	Source file without annotation sent to ULG MT
2	ULG MT	Source file with Glossary annotation sent to ULG MT
3	GPT-3	Output of (2) sent to GPT-3 ‘edits’ endpoint with prompt "Corregir la gramática en español" ["temperature": 0 , engine="text-davinci-edit-001"]
4	GPT-3	Output of 3 sent to GPT-3 ‘edits’ endpoint with prompt "Corregir el orden de las palabras en español" ["temperature": 0, engine="text-davinci-edit-001"]
5	GPT-3	Source file without annotation sent to GPT-3 ‘completions’ endpoint with prompt "Traducir al español" ["temperature": 0, engine="text-davinci-002"]
6	GPT-3	Source file with Glossary annotation sent to GPT-3 ‘completions’ endopoint with prompt "Traducir al español con el glosario {source term}={target term}" ["temperature": 0, engine="text-davinci-002"]

Experiment Results

BLEU, TER and Terminology Consistency scores

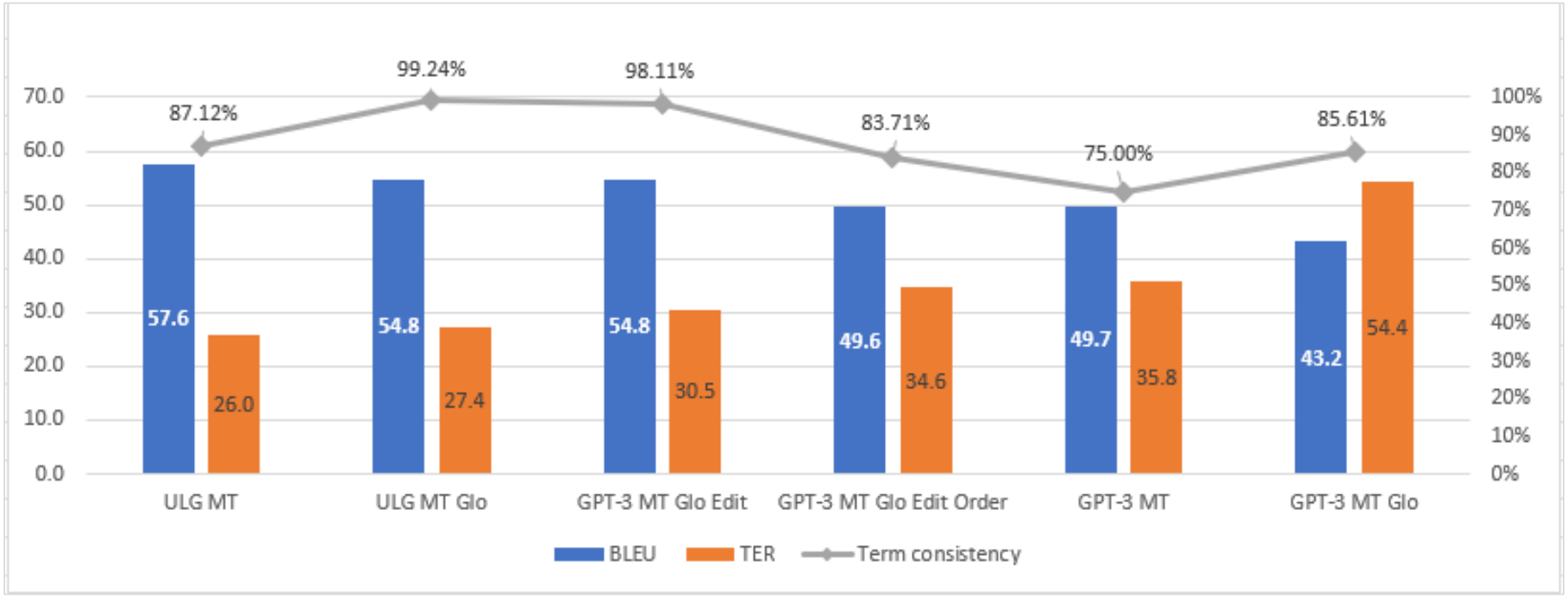
		BLEU	TER	TC
1	ULG MT without glossary	57.6	26.0	87.12%
2	ULG MT with glossary	54.8	27.4	99.24%
3	GPT-3 PE grammar	54.8	30.5	98.11%
4	GPT-3 PE grammar and order	49.6	34.6	83.71%
5	GPT-3 MT without glossary	49.7	35.8	75.00%
6	GPT-3 MT with glossary	43.2	54.4	85.61%

UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Results: Output Scores

BLEU, TER and Terminology consistency scores



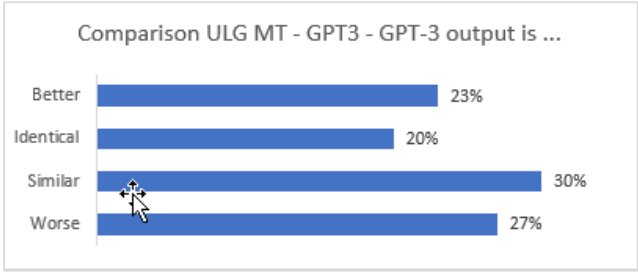
Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Results: Comparative Analysis

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

RESULTS

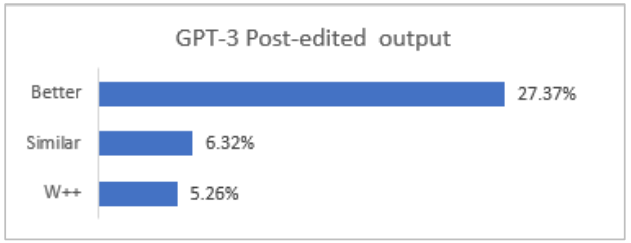
- Most outputs are either similar or identical to the ones of our current models.
- GPT-3 is less conservative in the preservation of the source.
- Makes changes that need to be contrasted with the input.
- Need to fix GPT-3's addition of extra dots (and blanks)

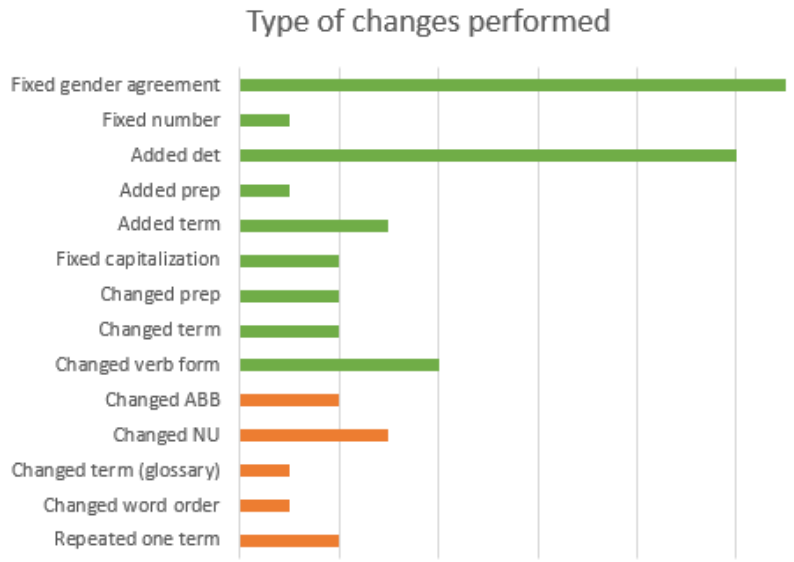


34	(for example, **argon**{***argón***})	34	(por ejemplo, argón)	<>	Por ejemplo, el argón.
35	**dry inert gas**{***gas inerte seco***}	35	gas inerte seco	<>	Gas inerte y seco.
36	Simultaneously available 4-20 mA **active output**{***salida activa***}, represents on-line density or temperature	36	Salida activa de 4-20 mA disponible simultáneamente, representa la densidad o la temperatura en línea	<>	Disponible simultáneamente 4-20 mA salida activa, representa en línea densidad o temperatura.

Results: Comparative Analysis

2 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

Results: Comparative Analysis

EXAMPLES OF IMPROVED OUTPUTS

★★★	78	Test points with **Product01**{*Product01*} tool	Puntos de prueba con Product01 herramienta	<>	Puntos de prueba con la herramienta Product01
★★★	19	For Model 1500, output voltage is +15 VDC ±3% with a 2,2 kohm internal **pull-up resistor**{*resistencia pull-up*}.	Para el modelo 1500, el voltaje de salida es de +15 VCC ±3% con un resistencia pull-up interno de 2,2 kohm.	<>	Para el modelo 1500, el voltaje de salida es de +15 VCC ±3% con una resistencia pull-up interna de 2,2 kohm.
★★☆	95	If you are installing a cable gland at the I.S. barrier, a standard cable gland with a sealing washer, **lock washer**{*arandela*}, and lock nut is required.	Si está instalando un prensaestopas en la barrera I.S., se requiere un prensaestopas estándar con una arandela de sellado, arandela y tuerca de bloqueo.	<>	Si está instalando un prensaestopas en la barrera I.S., se requiere un prensaestopas estándar con una arandela de sellado, una arandela y una tuerca de bloqueo.
★★☆	241	One exception is **nitric acid**{*ácido nítrico*}, for which 304 stainless steel has better corrosion resistance.	Una excepción es ácido nítrico, para el cual el acero inoxidable 304 tiene una mejor resistencia a la corrosión.	<>	Una excepción es el ácido nítrico, para el cual el acero inoxidable 304 tiene una mejor resistencia a la corrosión.
★★☆	104	**infrared port**{*puerto infrarrojo*} write-protection	puerto infrarrojo protección contra escritura	<>	puerto infrarrojo con protección contra escritura
★☆☆	66	To poll an **external temperature device**{*dispositivo de temperatura externo*}, configure polling for temperature as described in.	Para sondear un dispositivo de temperatura externo, configure el sondeo para la temperatura como se describe en.	<>	Para sondear un dispositivo de temperatura externo, configure el sondeo para la temperatura como se describe en la sección.
★☆☆	107	If you are using the **infrared port**{*puerto infrarrojo*}, position the infrared device for communication with the infrared port (see).	Si está utilizando el puerto infrarrojo, coloque el dispositivo infrarrojo para la comunicación con el puerto infrarrojo (consulte).	<>	Si está utilizando el puerto infrarrojo, coloque el dispositivo infrarrojo para la comunicación con el puerto infrarrojo (consulte la figura).

UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Results: Comparative Analysis

EXAMPLES OF WRONG CORRECTIONS

Changed glossary term

238	**config I/O custom**{*E/S configurables personalizadas* }	E/S configurables personalizadas	<>	E/S personalizadas configurables
239	**config I/O default**{*E/S configurables predeterminadas* }	E/S configurables predeterminadas	<>	E/S configurables por defecto

Acceptable worsening

133	Maximum distance between remote **booster amplifier**{*amplificador booster* } and sensor is 18m.	La distancia máxima entre el amplificador booster remoto y el sensor es de 18 m.	<>	La distancia máxima entre el amplificador booster remoto y el sensor es de 18 metros.
76	External means that an external **pull-up resistor**{*resistencia pull-up* } and source are required.	Externo significa que se requieren resistencia pull-up y una fuente externas.	<>	Externo significa que se requieren resistencia pull-up y una fuente externa.
18	**engineering unit of measurement**{*unidad de ingeniería de medición, unidad de ingeniería de medida* }	unidad de ingeniería de medición, unidad de ingeniería de medida	<>	unidad de ingeniería de medición, unidad de ingeniería de medida, unidad de ingeniería de mediciones

CORRECTION NOT SUPPORTED

63	For F-Series sensors with a junction box or an **extended core processor**{*procesador central extendido* }, the maximum ambient temperature is 60°C (140°F).	Para los sensores de la serie F con una caja de conexiones o un procesador central extendido, la temperatura ambiente máxima es de 60 ° C (140 ° F).	<>	Para los sensores de la serie F con una caja de conexiones o un procesador central extendido, la temperatura ambiente máxima es de 60 ° C (140 ° F).

Results: Comparative Analysis

3 Check if **GPT-3** can **improve its own Post-Editing** by requesting **word order correction**

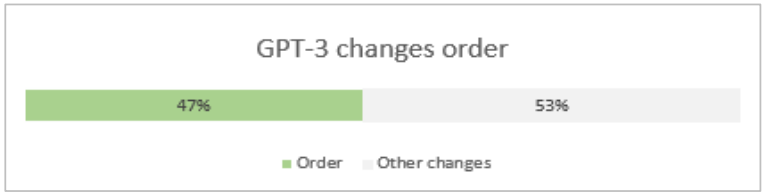
RESULTS

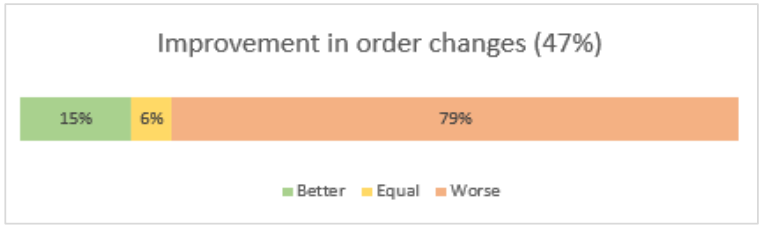
ORDER CHANGES

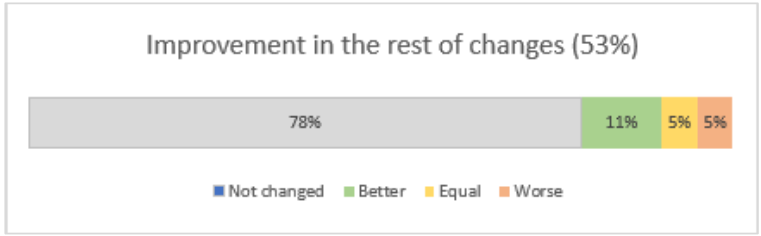
- Improve: apposition of proper names
- Similar: order of exchangeable noun modifiers
- Worsening: change order in glossary terms, change in term meaning, change of the translation of a glossary term

REST OF CHANGES

- Improve: added fluency (adding determiners)
- Worsening: adds a duplicated term







UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

Results: Comparative Analysis

EXAMPLES OF IMPROVED OUTPUTS

★★☆

249	**Product01|Product01| 5-pin Product02 connector in M20 housing**	Product01 conector Product02 de 5 pines en carcasa M20.	<>	Conector Product01 Product02 de 5 pines en carcasa M20.

EXAMPLES OF WRONG CORRECTIONS

!

236	(or pin 2-3 of Product02™ Product01|Product01|™ connector)	(o pin 2-3 del conector Product02™ Product01 ™).	<>	(o pin 2-3 del conector Product01 ™ Product02™).

!

2	Using a **DeviceNet**{***DeviceNet***} tool, three methods are available for totalizer and inventory control:	Utilizando una herramienta de DeviceNet, hay tres métodos disponibles para el totalizador y el control de inventario.	<>	Utilizando una herramienta de DeviceNet, hay tres métodos disponibles para el control de inventario y el totalizador.

!

13	**DIN rail enclosure**{***cubierta de carril DIN***}	cubierta de carril DIN.	<>	cubierta de DIN carril.
14	**core processor entity parameter**{***parámetro de entidad del procesador central***}	parámetro de la entidad del procesador central	<>	parámetro de la entidad central del procesador
15	**Construction Identification Code**{***código de identificación de construcción***}	código de identificación de construcción.	<>	código de construcción de identificación.
16	Side view with **rupture disk**{***disco de ruptura***}	Vista lateral con disco de ruptura.	<>	Vista lateral con ruptura de disco.
18	**engineering unit of measurement**{***unidad de ingeniería de medición, unidad de ingeniería de medida***}	unidad de ingeniería de medición, unidad de ingeniería de medida, unidad de ingeniería de mediciones	<>	unidad de medición de ingeniería, unidad de medida de ingeniería, unidad de mediciones de ingeniería

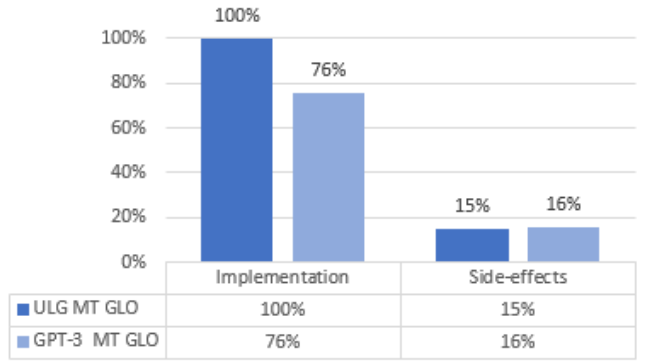
Results: Comparative Analysis

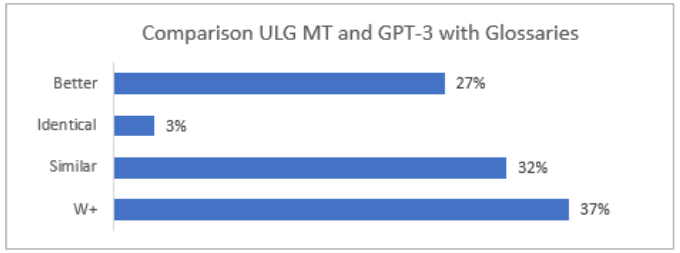
4

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





UNITED LANGUAGE GROUP

Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas, Orlando, USA, September 12-16, 2022. Volume 2: Users and Providers Track and Government Track

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

UNITED LANGUAGE GROUP

Credits

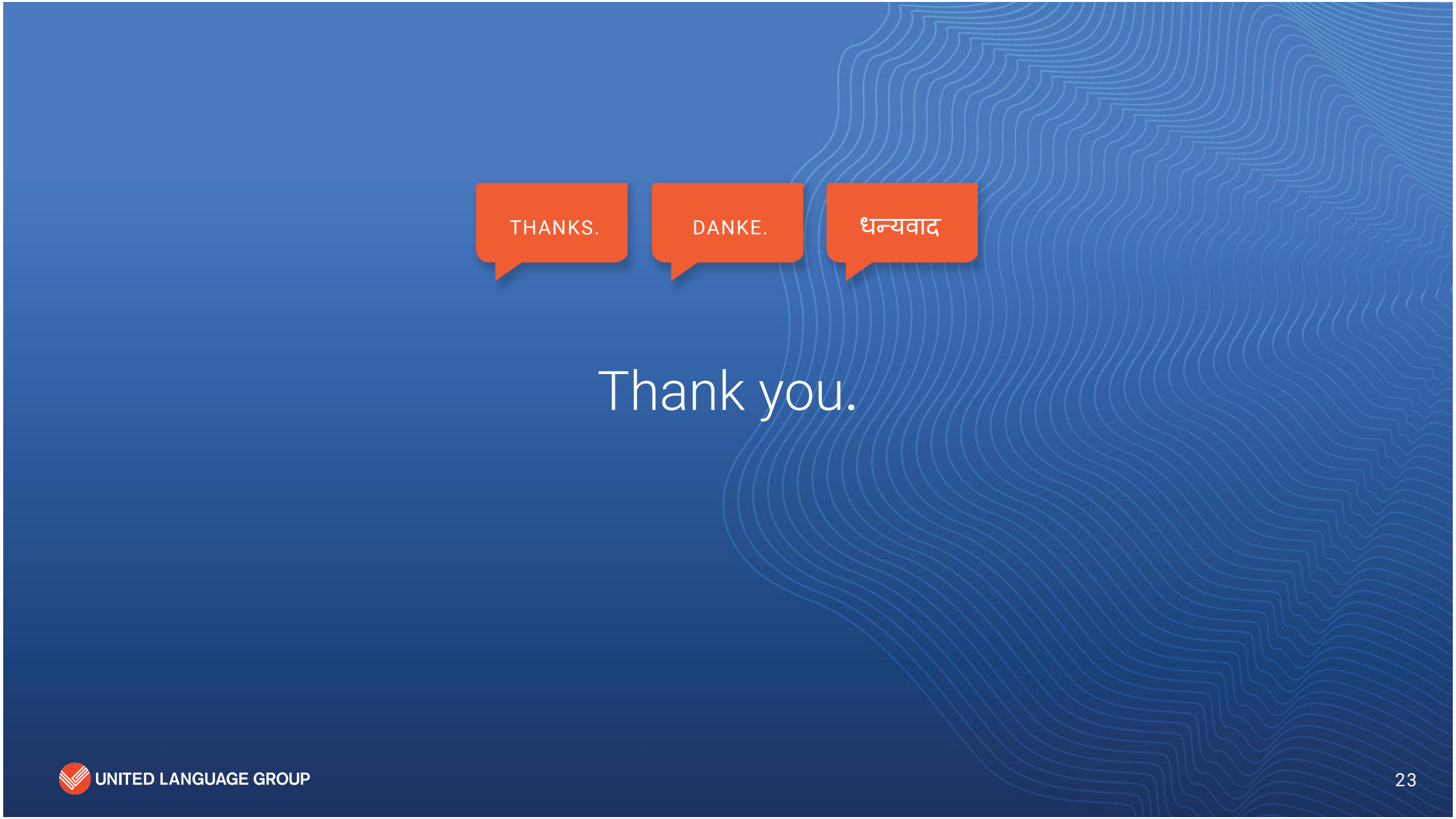
Alonso, Juan Alberto
juan.alonso@ulgroup.com

Llorens, Albert
albert.llorens@ulgroup.com

Madan, Mehul
mehul.madan@ulgroup.com

Vidal, Blanca
blanca.vidal@ulgroup.com

UNITED LANGUAGE GROUP


THANKS.
DANKE.
धन्यवाद
Thank you.
UNITED LANGUAGE GROUP
23