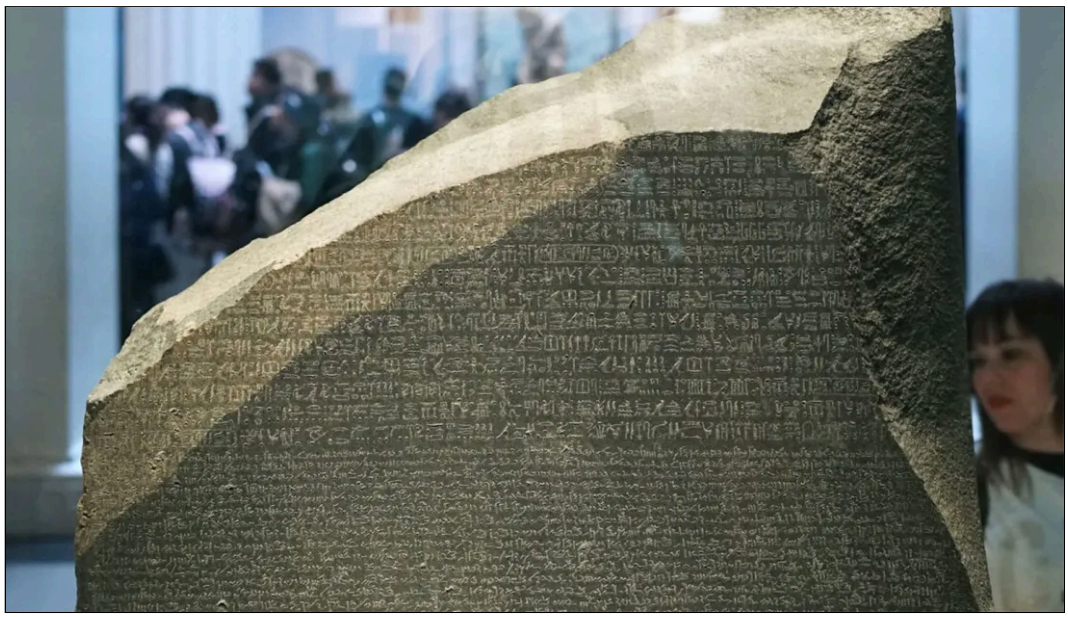
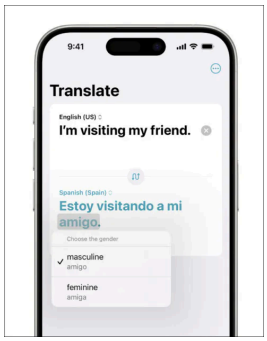
Apple’s Solution to Translating Gendered Languages

Originally published at https://www.unite.ai/apples-solution-to-translating-gendered-languages/

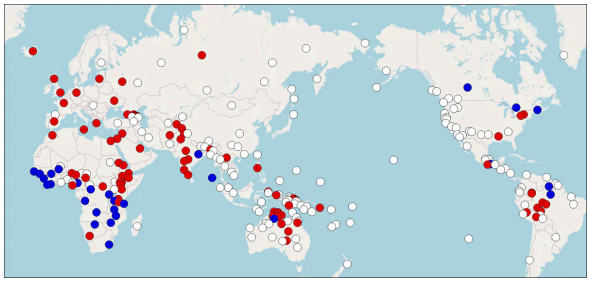


Apple AAPL 0.36% has just published a paper, in collaboration with USC, that explores the machine learning methods employed to give users of its iOS18 operating system more choice about gender when it comes to translation.



In iOS18, users can select alternative gender suggestions for a translated word in the native Translate app. Source: https://support.apple.com/guide/iphone/translate-text-voice-and-conversations-iphd74cb450f/ios

Though the issues tackled in the work (which Apple has announced here) engages, to a certain extent, in current topical debates around definitions of gender, it centers on a far older problem: the fact that 84 out of the 229 known languages in the world use a sex-based gender system.



The red dots indicate languages that use a sex-based gender system. Source: https://wals.info/feature/31A#map

Surprisingly, the English language falls into the sex-based category, because it assigns masculine or feminine singular pronouns.

By contrast, all Romance languages (including over half a billion Spanish speakers) – and multiple other popular languages, such as Russian – require gender agreement in ways that force translation systems to address sex-assignment in language.

The new paper illustrates this by observing all possible Spanish translations of the sentence *The secretary was angry with the boss*:

From the new paper, an example of the potential gender assignments in the sentence 'The secretary was angry with the boss', translating from English to Spanish. Source: https://arxiv.org/pdf/2407.20438

Naïve translation is far from sufficient for longer texts, which may establish gender at the start (*'He'*, *'She'*, etc.) and thereafter not refer to gender again. Nonetheless, the translation must remember the assigned gender of the participant *throughout the text*.

This can be challenging for token-based approaches that address translations in discrete chunks, and risk to lose the assigned gender-context throughout the duration of the content.

Worse, systems that provide alternative translations for biased gender assignments cannot do this indiscriminately, i.e., by merely substituting the gender noun, but must ensure that all other parts of language agree with the changed gender noun.

In this example from the Apple/USC paper, we see that though *Secretary* has been assigned a male gender, the singular past *was* has been left as feminine (*estaba*):

Brute-force gender substitutions can neglect necessary gender agreement. In this example, the word 'enojada' should be 'enojado', to agree with the masculine 'El secretario'.

A translation system must also cope with the eccentricities of particular languages in regard to gender. As the paper points out, the pronoun *I* is gendered in Hindi, which provides an uncommon clue to gender.

Gender Issues

In the new paper, titled *Generating Gender Alternatives in Machine Translation*, the Apple and USC researchers propose a semi-supervised method to convert gender-ambiguous entities into an array of entity-level alternatives.

The system, which was used to inform translation from the Apple Translate app in iOS18, constructs a language schema by both the use of large language models (LLMs), and by fine-tuning pre-trained open source machine translation models.

The results from translations from these systems were than trained into an architecture containing *gender structures* – groups of phrases that contain diverse forms of varying gendered nouns representing the same entity.

The paper states*:

> *'Gender biases present in train data are known to bleed into natural language processing (NLP) systems, resulting in dissemination and potential amplification of those biases. Such biases are often also the root cause of errors.*
>
> *'A machine translation (MT) system might, for example, translate doctor to the Spanish term médico (masculine) instead of médica (feminine), given the input "The doctor asked the nurse to help her in the procedure".*
>
> *'To avoid prescribing wrong gender assignment, MT systems need to disambiguate gender through context. When the correct gender cannot be determined through context, providing multiple translation alternatives that cover all valid gender choices is a reasonable approach.'*

The approach that the researchers arrive at effectively turns a translation from a single token to a user-controlled array.

(Though the paper does not mention it, this opens up the possibility, either in Apple Translate or in similar portals that offer translation services, for user choices to be fed back into later iterations of the model)

The model Apple and USC developed was evaluated on the GATE and MT-GenEval test sets. GATE contains source sentences with up to 3 gender-ambiguous entities, while MT-GenEval contains material where gender cannot be inferred, which, the authors state, aids in understanding when alternative gender options should not be offered to the user.

In both cases, the test sets had to be re-annotated, to align with the aims of the project.

To train the system, the researchers relied on a novel automatic data augmentation algorithm, in contrast to the aforementioned test sets, which were annotated by humans.

Contributing datasets for the Apple curation were Europarl; WikiTitles; and WikiMatrix. The corpora was divided into *G-Tag* (with 12,000 sentences), encompassing sentences with head words for all entities, together with a gender-ambiguous annotation; and *G-Trans* (with 50,000 sentences), containing gender-ambiguous entities and gender alignments.

The authors assert:

> *'To the best of our knowledge, this is the first large-scale corpus that contains gender ambiguities and how they effect gendered forms in the translation.'*

Datasets and diverse data for the project have been made available on GitHub. The data features five language pairs, pitting English against Russian, German, French, Portuguese and Spanish.

The authors leveraged a prior approach from 2019 to endow the model with the capability to output gender alignments, training with cross entropy loss and an additional alignment loss.

For the data augmentation routine, the authors eschewed traditional rule-based methods in favor of a data-centric approach, fine-tuning a BERT pre-trained language model on the G-Tag dataset.

Double-Take

For cases where ambiguous gender entities are detected, Apple and USC explored two methods – the fine-tuning of pre-trained language models, and the use of LLMs.

In regard to the first method, the paper states:

> *'We fine-tune a pre-trained MT model M on a bitext extracted from the G-Trans dataset. The source sentences of this bi-text contain ambiguous entities tagged as masculine or feminine using* <M>/<F> *tags, and the target translation has correct gender inflections given the gender tags.'*

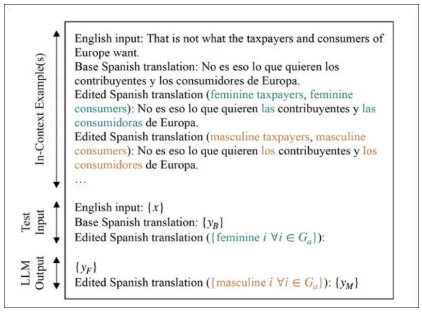
	Source	Target
G-Trans dataset	The doctor was angry with the patient doctor → Gender-Ambiguous patient → Gender-Ambiguous	(El doctor / La doctora) estaba (enojado / enojada) con (el / la) paciente
Fine-tuning bi-text	The doctor<M> was angry with the patient<M> The doctor<F> was angry with the patient<F> The doctor<M> was angry with the patient<F> The doctor<F> was angry with the patient<M>	*El doctor* estaba *enojado* con *el* paciente *La doctora* estaba *enojada* con *la* paciente *El doctor* estaba *enojado* con *la* paciente *La doctora* estaba *enojada* con *el* paciente

An illustration of the schema for extracting bi-text from the G-Trans dataset.

In the image above, we see the fine-tuned text in the lower middle column, and the desired output in the right column, with the underlying rationale illustrated above.

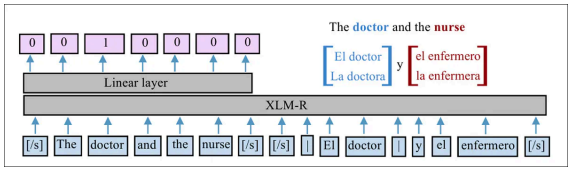
For this approach, the authors made use of a lattice rescoring method from an earlier 2020 work. To ensure that only the target domain (gender) was addressed, a constrained beam search was used as a filter.

For the LLM approach, the authors devised a strategy that uses an LLM as an editor, by re-writing the supplied translations to provide gender assignments.



The LLM is prompted using an in-context example in order to assign gender.

With results from both approaches concatenated, the model was subsequently fine-tuned to classify source tokens as *aligned* (indicated by '1' in the schema below) or *non-aligned* (indicated by '2' below).



A schema for the concatenation of results from both approaches.

Data and Tests

The *ambiguous entity* detector used for the project was developed by fine-tuning Facebook AI's xlm-roberta-large model, using transformers. For this, the combined G-Tag was used across all five language pairs.

In the first of the aforementioned two approaches, the M2M 1.2B model was trained on Fairseq, jointly with bi-text data from the G-Trans dataset, with gender inflections provided by Wiktionary.

For the LLM method, the authors used GPT-3.5-turbo. For the alignment of gender structures, xlm-roberta-large was again used, this time with gender alignments extracted from G-Trans.

Metrics for the evaluation of alternatives, structure (with *precision* and *recall*), and *alignment accuracy*.

Though the first two of these are self-explanatory, alignment accuracy measures the percentage of output gender structures that conform to the known correct source identity, and uses the δ-BLEU method, in accordance with the methodology for MT-GenEval.

Below are the results for the data augmentation pipeline:

Language Pair	Model	Alternatives Metrics ↑ Precision%	Alternatives Metrics ↑ Recall%	δ-BLEU ↓	Structure Metrics ↑ Precision%	Structure Metrics ↑ Recall%	Alignment ↑ Accuracy%
En–De	Fine-tuned M2M	**94**	89.7	4.7	87.8	91	93.7
	GPT	91.1	**92.7**	**2.8**	**89.8**	**94**	
En–Es	Fine-tuned M2M	**95.7**	91.6	3.3	**88.1**	**93**	91.5
	GPT	91.5	**93.7**	**2.7**	84.7	92.7	
En–Fr	Fine-tuned M2M	**93.8**	**92.5**	3.6	**88.1**	92.9	92.9
	GPT	89.4	91	**2.8**	85.8	**94.8**	
En–Pt	Fine-tuned M2M	**94.8**	**94.3**	**3.5**	88.3	92.4	93.6
	GPT	93.8	83.5	5.5	**89.6**	**95.2**	
En–Ru	Fine-tuned M2M	**89.4**	**89.3**	**5.7**	**87**	**87.7**	93.2
	GPT	83.5	58.2	10.6	83.1	85	
En–It	Fine-tuned M2M	95.4	87.9	8.2	79.4	75.3	94.1

Results from the data augmentation tests. Upward arrows indicates 'higher-the-better', downward 'lower-the-better'.

Here the authors comment*:

> *'Both M2M and GPT perform mostly on par with the exception of English-Russian, where GPT achieves much lower alternatives recall (58.7 compared to 89.3). The quality of generated gender structures is better for GPT on English-German and English-Portuguese and better for M2M on English-Spanish and English-Russian, as can be seen from the structure metrics.*
>
> *'Note that we don't have any G-Trans data for English-Italian, so the results of the M2M model and the alignment accuracy on English-Italian are purely due to zero-shot generalization of M2M and XLM models.'*

The researchers also compared the data augmentation system's performance, via M2M, against GATE's sentence-level gender re-writer, on GATE's own stated terms.

LP	Direction	Model	P%	R%	F0.5
En–Es	M→F	GATE	**95**	40	0.75
		Ours	89.6	**69.2**	**0.85**
	F→M	GATE	**97**	50	0.82
		Ours	94.5	**73.7**	**0.89**
En–Fr	M→F	GATE	**91**	27	0.62
		Ours	89.3	**72.5**	**0.85**
	F→M	GATE	**97**	28	0.65
		Ours	96.1	**79.3**	**0.92**
En–It	M→F	GATE	**91**	32	0.66
		Ours	78.7	**58.8**	**0.74**
	F→M	GATE	**96**	47	0.79
		Ours	92	**75.1**	**0.88**

The Apple/USC data augmentation pipeline pitted against the GATE sentence-level method.

Here the paper states:

> *'We see significant improvements in recall at the cost of relatively small degradation in precision (except English-Italian). Our system is able to outperform GATE on their proposed F.5 metric on all 3 language pairs.'*

Finally, the authors trained diverse 'vanilla' multilingual models into *vanilla bi-text*. The contributing datasets were WikiMatrix, WikiTitles, Multi-UN, NewsCommentary, and Tilde.

Two additional vanilla models were trained, one incorporating the G-Trans dataset with the prefixed tag *<gender>*, which was employed as the supervised baseline; and a third, incorporating gender structure and alignments (on the smaller local model, since using GPT's API-based services would have been very expensive for this purpose).

The models were tested against the 2022 FloRes dataset.

Language Pair	Model	Alternatives Metrics ↑ P%	Alternatives Metrics ↑ R%	δ-BLEU ↓	Structure Metrics↑ P%	Structure Metrics↑ R%	Alignment Accuracy ↑ %	FLoRes BLEU ↑
En–De	Vanilla	-	-	8.6	-	-	-	31.6
	Supervised	74.4	71.5	2.4	**55.2**	**57.5**	89.1	**31.9**
	w/ Augmented Data	**86.7**	**87.5**	**0.8**	48.2	55.6	**94.2**	31.6
En–Es	Vanilla	-	-	10.4	-	-	-	**26**
	Supervised	78.9	77.3	2.8	60.5	60.6	85.2	25.9
	w/ Augmented Data	**94.3**	**92**	**1**	**62.4**	**66.4**	**92.5**	**26**
En–Fr	Vanilla	-	-	8.1	-	-	-	**46.3**
	Supervised	74.5	67.8	3.1	**60.7**	61.7	82.1	44.9
	w/ Augmented Data	**87.3**	**86.7**	**0.8**	59	**67.3**	**92.5**	45.8
En–Pt	Vanilla	-	-	12.5	-	-	-	**44.6**
	Supervised	83.4	82.6	3.1	**60**	60.9	86.9	43.7
	w/ Augmented Data	**92.2**	**94.4**	**1.1**	59.5	**63.5**	**94.2**	44.1
En–Ru	Vanilla	-	-	5.3	-	-	-	25.6
	Supervised	70.6	54.5	2.4	**42**	39.5	83.7	**26.4**
	w/ Augmented Data	**80.7**	**77.2**	**1.5**	37.6	**39.8**	**91**	24.9
En–It	Vanilla	-	-	11.6	-	-	-	**27.9**
	w/ Augmented Data	93.7	89.4	**3.2**	53	50.9	94.6	27.6

End-to-end vanilla machine translation models tested (P = precision, R = recall).

The paper summarizes these results:

> *'The vanilla model cannot generate alternatives and shows a huge bias towards generating masculine forms (δ-BLEU ranging from 5.3 to 12.5 points).*
>
> *'This bias is greatly reduced by the supervised baseline. The model trained on augmented data further reduces the bias and obtains the best performance in terms of alternative metrics, alignment accuracy, and δ-BLEU.*
>
> *'This shows the effectiveness of the data augmentation pipeline. Augmented data also allows us to train a competitive system for English-Italian which lacks supervised data.'*

The authors conclude by noting that the success of the model has to be considered in the broader context of NLP's struggle to rationalize gender assignment in a translation method; and they note that this remains an open problem.

Though the researchers consider that the results obtained do not fully achieve the aim of the generation of entity-level gender-neutral translations and/or disambiguations regarding gender, they believe the work to be a ‘powerful instrument’ for future explorations into one of the most challenging areas of machine translation.

* *My conversion of the authors’ inline citations to hyperlinks*

First published Tuesday, October 8, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More