



CASE STUDY: MT TRAINING DATA FOR PANGEANIC

Improving Adaptive MT Outputs by on Average 22% in BLEU Scores Across Five Languages

TAUS provided language data for Pangeanic, a leading European NLP and translation services company, to train their machine translation models for the COVID-19, pandemic and healthcare domain.



THE CHALLENGE

Finding high-quality data for MT training has always been a challenge on the path to generating high-performing MT output. The challenge increases when the language pairs are rare or when training data in a lesser-known domain is needed.

Due to the global pandemic caused by COVID-19, a domain that had not been so popular came into the spotlight. To enable faster, accurate and automated translations for the vital information on this topic, training datasets in the pandemic, COVID-19, viral illnesses and healthcare were required.

Machine translation systems do a great job at providing solutions for automated translation services when fed with the right training data. However, it was a challenge to find high-quality datasets necessary to build specialized automatic tools about this new topic in the healthcare domain.

SOLUTION

TAUS' expertise in domain-specific training data collection and creation was instrumental in Pangeanic's decision to partner with us.



TAUS stands out because of their capabilities in the language data space, but we were also impressed by their expertise in fine-tuning the datasets to match the exact domain requirements. Using the datasets provided by TAUS, we've run experiments in English to Spanish, German, Polish, Russian, and Chinese language pairs for the pandemic and healthcare domain



Mercedes García-Martínez, Chief Research Scientist at Pangeanic

DATA COLLECTION

TAUS provided Pangeanic a total of 1.8 million words of MT training data in English to Spanish, German, Polish, Russian, and Chinese language pairs.

DATA SELECTION

The translation units (TUs) in the pandemic and healthcare dataset provided by TAUS were sorted by their relevance to the domain. The most relevant TUs were placed on the top of the file from strictly coronavirus related TUs to more general TUs in the pandemic and healthcare domain. This method allowed the customer to filter the data based on how specific they wanted their MT engines to be in the given domain or how broad they wanted to go in it. Pangeanic also made use of this method and, after performing automatic cleaning, they carried out manual checks to choose the TUs most related to coronavirus for their training purposes.

RESULTS

22,6% average increase on BLEU scores

7% average increase on TER scores

8,6% average increase on ChrF scores

Using the data provided by TAUS, Pangeanic built COVID-19 domain-specific neural machine translation (NMT) models for the five language pairs on Pangeanic ECO user-friendly customer portal on which the user can adapt models using three levels of training settings.

The three levels of aggressivity is a proprietary Pangeanic technology that flexibly trains Deep Learning algorithms. Users can choose to simply add data to re-train the engine in the usual way as other ML companies (**conservative**), prioritize it (**normal**), or impact learning rates very deeply (**aggressive**). In Deep Adaptive Machine Translation's "aggressive mode", engines learn from the incoming material at much faster rates than by the traditional "addition" or "prioritization", which results in higher parity rates.

LANGUAGE-SPECIFIC RESULTS

The highest BLEU score improvement has been recorded in the English > Russian language pair with 50%, followed by English > Chinese with 26%, English > German with 20%, English > Spanish with 9%, and English > Polish with 8%.

EN>ES	MODEL	BLEU	TER	CHRF
	ENES base	44.02	42.58	68.51
	ENES conservative	46.52	40.77	69.97
	ENES normal	47.27	40.47	70.38
	ENES aggressive	47.80	39.88	70.70
	Improvement Percentage Accomplished	9%	7%	3%
EN>DE	MODEL	BLEU	TER	CHRF
	ENES base	30.72	59.63	61.42
	ENES conservative	34.91	55.79	64.14
	ENES normal	35.23	55.41	64
	ENES aggressive	36.86	53.76	65.15
	Improvement Percentage Accomplished	20%	1%	6%
EN>PL	MODEL	BLEU	TER	CHRF
	ENES base	35.47	53.31	61.73
	ENES conservative	37.56	51.79	63.35
	ENES normal	38.01	51.55	63.54
	ENES aggressive	38.17	51.54	63.62
	Improvement Percentage Accomplished	8%	3%	3%
EN>RU	MODEL	BLEU	TER	CHRF
	ENES base	19.37	70.73	49.27
	ENES conservative	27.14	64.65	56.22
	ENES normal	28.21	63.08	57.24
	ENES aggressive	29.02	62.16	57.81
	Improvement Percentage Accomplished	50%	14%	17%
EN>ZH	MODEL	BLEU	TER	CHRF
	ENES base	23.55	56.03	34.34
	ENES conservative	28.74	51.42	37.95
	ENES normal	29.17	51.71	38.51
	ENES aggressive	29.73	50.84	39.02
	Improvement Percentage Accomplished	26%	10%	14%

QUALITY ANALYSIS

Quality analysis was also done comparing translation examples from the base model and the aggressive model. COVID-19 specific words were spotted to check how the model has been adapted. Based on the analysis, it was discovered that in all cases the adaptive model provides more accurate translations and can deal with different linguistic challenges better after training with the datasets provided by TAUS. Here are some examples of the quality analysis on the translations:

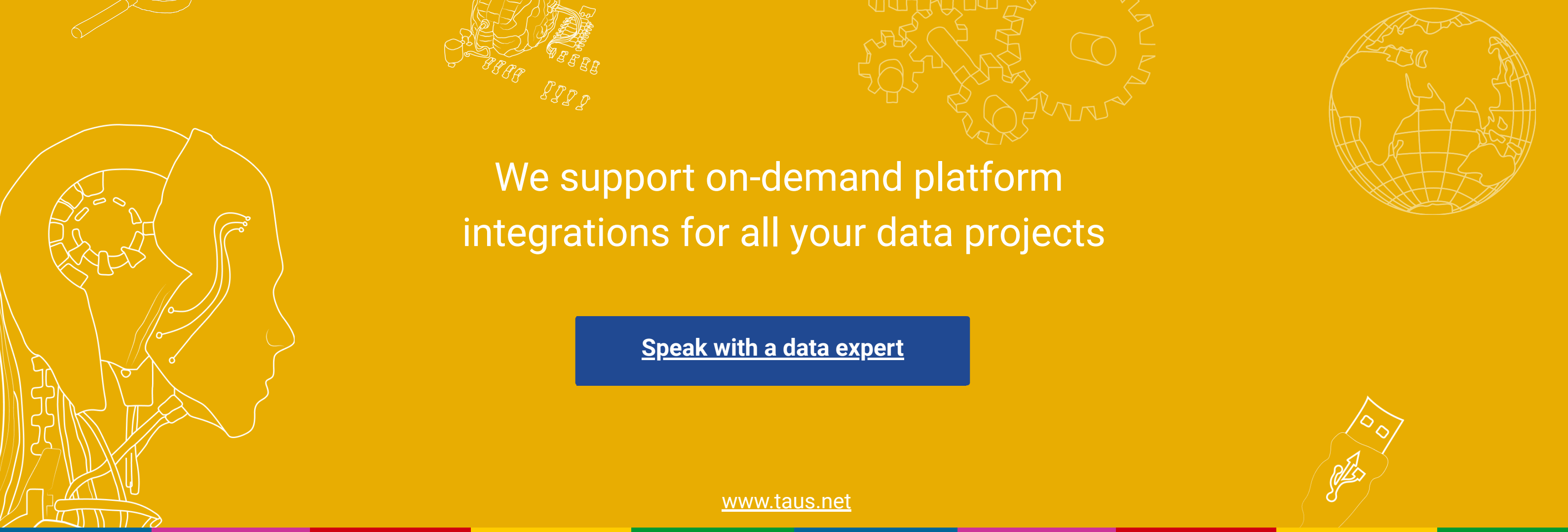
	SOURCE	REFERENCE	COMMENT
ENGLISH > SPANISH	These may signal serious side effects that may need urgent medical attention.	Estos pueden indicar efectos adversos graves que pueden requerir atención médica urgente .	In this example, it can be seen that the adaptive model is able to translate the word urgent, whereas the base model could not translate it.
Base model	Estos pueden indicar efectos secundarios graves que pueden necesitar atención médica.		
Back translation	These can indicate serious side effects that may need medical attention.		
Adaptive model	Estos pueden indicar efectos secundarios graves que pueden necesitar atención médica de urgencia .		
Back translation	These can indicate serious side effects that may need urgent medical attention.		

	SOURCE	REFERENCE	COMMENT
ENGLISH > GERMAN	The truth is that the swine flu epidemic will be created BY THE VACCINE.	Die Wahrheit ist, dass die Schweinegrippeepidemie VON DER IMPFUNG erzeugt werden wird.	In the English-German translation example, the base model cannot correctly translate the article “the” because the next word is VACCINE which is not so frequent. However, the adaptive model can translate it correctly generating “der” for German translation.
Base model	Die Wahrheit ist, dass die Schweinegrippe-Epidemie von THE VACCINE verursacht wird		
Back translation	The truth is, THE VACCINE is causing the swine flu epidemic		
Adaptive model	Die Wahrheit ist, dass die Schweinegrippe-Epidemie von DER VACCINE erstellt werden		
Back translation	The truth is that the swine flu epidemic created by THE VACCINE		

	SOURCE	REFERENCE	COMMENT
ENGLISH > POLISH	It blocks a protein called CCR5, which is found on the surface of the cells in the body that HIV infects.	Blokuje on białko określane jako CCR5, które występuje na powierzchni komórek w organizmie zakażanych przez wirus HIV.	In this example, base model translation generated the word “ciele” from “body” instead of “organizmie”. However, this is solved using adaptive model which translates “body” correctly as “organizmie”.
Base model	Blokuje białko o nazwie CCR5, które znajduje się na powierzchni komórek w ciele, które zaraża HIV.		
Back translation	It blocks a protein called CCR5 that is found on the surface of cells in the body that infects HIV.		
Adaptive model	Blokuje białko o nazwie CCR5, które znajduje się na powierzchni komórek w organizmie, które zakaża HIV.		
Back translation	It blocks a protein called CCR5 that is on the surface of cells in the body that infects HIV.		

	SOURCE	REFERENCE	COMMENT
ENGLISH > RUSSIAN	We also hosted a conference on combating stigma and discrimination against HIV-positive people aimed at merit-based social inclusion and the promotion of social support for the victims of the virus.	Мы также организовали у себя конференцию по борьбе со стигматизацией и дискриминацией в отношении людей, инфицированных ВИЧ, направленную на включение их в социальную жизнь, где их роль определяется их заслугами, и на содействие социальной поддержке жертв этого вируса.	In this case, it can be seen that the translation with the adaptive model generates more correct words than the base model such as the correct acronym of HIV in Russian and “стигматизацией” and “этого” as the reference. The adaptive sentence translation is shorter than the reference but preserves the same meaning. However, the base model left the English acronym HIV in the Russian translation and the translation is not complete.
Base model	Мы также провели конференцию по борьбе со стигмой и дискриминацией в отношении HIV-положительных людей с целью охвата общественностью и содействия социальной поддержке жертв вируса.		
Back translation	We also hosted a conference on combating stigma and discrimination against people living with HIV to promote social inclusion and social support for victims of the virus.		
Adaptive model	Мы также провели конференцию по борьбе со стигматизацией и дискриминацией в отношении ВИЧ-инфицированных лиц, направленную на социальную интеграцию и содействие социальной поддержке жертв этого вируса		
Back translation	We also hosted a conference on combating stigma and discrimination against people living with HIV to promote social inclusion and social support for victims of the virus.		

	SOURCE	REFERENCE	COMMENT
ENGLISH > CHINESE	It blocks a protein called CCR5, which is found on the surface of the cells in the body that HIV infects.	告诉我病毒的事-你想知道什么？	In this example, it's observed that the base model cannot translate the English word "Tell" and generates two wrong dots in the middle of the sentence. However, a big improvement is seen in the translation with the adaptive model generating a very close translation compared to the reference.
Base model	Tell 我关于病毒。-好的。你想知道什么？		
Back translation	Tell I am about viruses. -Ok . What do you want to know ?		
Adaptive model	告诉我病毒-好的你想知道什么？		
Back translation	Tell me about the virus-ok what do you want to know?		



We support on-demand platform integrations for all your data projects

[Speak with a data expert](#)

www.taus.net

TAUS AT A GLANCE

Trusted Data Expert

Since 2008

- ✓ Helping businesses build high-quality AI and ML models
- ✓ NLP-powered scalable data solutions

Largest Language Data Repository

With 35B Words

- ✓ Data in 600+ language pairs
- ✓ Platform for language data creation and annotation

Worldwide Community of Data Contributors

3000+ Workers

- ✓ Native and qualified workforces formed on-demand for each project
- ✓ Based in 15+ countries, completing 5,000+ microtasks monthly

www.taus.net