tarjamo

The impact of Quality, Quantity, or the Right Type of Nutrients Feeding NMT a Healthy Diet

Abdallah Nasir – ML Tech Lead Sara Alisis – Al Linguistic QA Lead





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



Breaking language barriers with Arabic language technology

Tarjama AI-enabled LSP

Tarjama is the leading tech-enabled LSP in the MENA region, offering a variety of language services such as translation, localization, subtitling, transcription, interpretation and content creation. A female-led business founded in 2008.

On a mission to break language barriers in the MENA market with Arabic Language Technology and a proprietary AI-powered language service platform.

Tarjama is featured in Gartner's Market

Guide for AI-Enabled Translation Services

a

ERNST & YOUNG ENTREPRENEUR OF THE YEAR 2018



TOP 100

TOP 30

The Impact of Healthy Data

Our NMT models are manually evaluated by our Linguistic QA team with Adapted MQM approach.

Manual Evaluation of 255 segments (5970 words).

84.9%→87.7%

Of translations considered OK, Good and Perfect

NMT Results per Quality levels

	MT Quality	Model 1 Good Data	Model 2 Good+Healthy Data
	Perfect MT translation	46.2%	46.2%
	Good MT translation (minor errors)	0.7%	1.9%
	OK translation (few errors)	38%	39.6%
	Bad translation	8.2%	6.2%
	Nonsense translation	6.2%	5.8%

The Impact of Healthy Data

Our Healthy Data added to Model 2 was 18K parallel sentences only!

tarjamo

What defines Healthy Data?

Is it what makes clients happy? Is it what comes from premium data sources? Is it what was created by professional linguists? Is it what you get from a translation management system (TMS)?

tarjama



What is Domain-Specific Data? (Examples)

"In the researchers' new approach, some of the browser's own internal components – those responsible for the decoding of media files – would be shifted into WebAssembly sandboxes." "<mark>Anti-inflammatory</mark>: Medicine that reduces inflammation (swelling in the airway and <mark>mucus</mark> production)."

"By using a form of machine learning known as Convolutional Neural Networks (CNNs), the archaeologists created a computerized method that roughly emulates the thought processes of the human mind in analyzing visual information." "Percutaneous nephrolithotomy: When kidney stones can't be treated by the other procedures – either because there are too many stones, the stones are too large or heavy, or because of their location – percutaneous nephrolithotomy is considered."

Run-on Sentences (Example)

------مكتب (2)للمقاولات -----شركة (1)من :الطاعن أقام على كل / " ------ "تتحصل في أن على ما يبين من الحكم المطعون فيه وسائر الأوراق -وحيث إن الواقعات مدنى كلى أمام المحكمة الإبتدائية بطلب الحكم بفسخ عقد المقاولة المبرم بينه والمدعى عليهما والزامهما بأن 2010 / 1060 المطعون ضدهما ، الدعوي رقم /للإستشار ات الهندسية در هم مع غرامة التأخير اعتبارا من وحتى تاريخ الحكم ، وذلك تأسيسا على أنه اتفق مع المكتب المدعى عليه الثاني على (2.600000) بردا له بالتضامن فيما بينهما مبلغ مقدار ه أبرم عقد مقاوله مع الشركة 🛛 بمنطقه العوير الثانية، فقام الأخير باختيار المدعى عليها الأولى للقيام بتنفيذ أعمال المشروع ، 782 - 721 إنشاء فيلا سكنيه على قطعة الأرض رقم در هم، و على ان تكون هناك غرامه (4.200000)شهر ا مقابل مبلغ مالي مقطوع مقداره (12)المدعى عليها الأولى ثم بموجبه تم الإتفاق على أن تقوم بانجاز المشروع خلال در هم، بيد أن المدعى عليهما) 2.600000) در هم عن كل يوم تأخير ، وقد أوفي بكافه التزاماته بسداد الدفعات المتفق عليها والتي بلغت جملتها مبلغ (1800)تأخير بواقع مبلغ فشلا في تنفيذ المشروع في الميعاد المتفق عليه، وأن ما تم إنجازه من أعمال يعد قليلا بالنظر لما اتفق عليه بموجب العقد، ومن ثم فقد أقام الدعوي ، دفعت المدعى عليها الثانية بعدم قبول الدعوى لوجود شرط التحكيم، وحكمت المحكمة برفض هذا الدفع وباختصاصها بنظر الدعوى وندبت خبيرة، وبعد أن أودع تقريره، وجهت المدعى عليها الأولى إلى المدعي القضائية وحتى السداد التام، سنويا إعتبارا من تاريخ المطالبة % 12در هما وفوائده القانونية بواقع (2.132.236)طلبة عارضة للحكم بالزامه بأن يؤدي لها مبلغا مقداره من أعمال المقاولة الأصلية بالإضافة إلى أعمال إضافية قامت بتنفيذها بموجب ملحق العقد المبرم بينهما وقيمتها المبلغ المطالب به ، (% 70) على أنها قامت بإنجاز نحو تأسيسا بيد أن الخبير لم يحتسب نسبة الإنجاز الحقيقة واحتسب غرامات تأخيريه رغم أن سبب التأخير في تنفيذ الأعمال يرجع إلى المدعى مما تستحق معه قيمة هذه الأعمال ، وحكمت در هما والفائدة (319.854) الدعوى المتقابلة ، بالزام المدعى عليه بأن يؤدي للمدعية مبلغا مقداره /المحكمة أو لا بر فض الدعوى الأصلية ثانية في موضوع الطلب العارض المدعية تقابلا ، هذا الحكم في شقه المتعلق بالطلب /سنوية إعتبار إ من تاريخ المطالبة القضائية وحتى السداد التام، استأنفت المدعية في الطلب العارض 9% القانونية عنه بو اقع مدنى 2012 / 42مدنى، قضت المحكمة أو لا في موضوع الاستئناف رقم 2012 /78مدنى، كما استأنفه المدعى أصلياً بالاستئناف رقم 2012 / 42العارض بالاستئناف رقم في موضوع : المستأنف فيما عدا ذلك ، ثانيا 🛛 در هما وتأييد الحكم (454.254) للمقاولات ليصبح مبلغ بتعديل المبلغ المقضى به في الدعوى المتقابلة لصالح المستأنفة . مدنى بإلغاء الحكم المستأنف الصادر في الدعوى الأصلية والقضاء مجددا بفسخ عقد المقاولة المبرم بين المستأنف والمستأنف ضدها الأولى وملحقه ورفض ما عدا ذلك الاستئناف رقم من طلبات، طعن المدعى اصلية في هذا الحكم بالتمييز الماثل بموجب صحيفه أودعت قلم كتاب هذه المحكمة وطلب فيها نقضه، وأودع مستند لم يسبق طرحه أمام محكمة الموضوع ، وبعد أن غرض الطعن على المحكمة في غرفه 🛛 وذلك لما هو مقرر من أنه لا يقبل التحدي أمام محكمة التمييز بمستند لم يسبق عرضه على محكمة الموضوع 🕞 استبعدته المحكمة مشوره رأت انه جدیر بالنظر وحددت جلسة لنظره

Short Sentences (Examples)

Hello

Good company.

Yes, I can

This is not helpful.

torjama

Common Issues

Following are few of the common issues we found while acquiring Good Data

tarjamø

High-Quality for Clients but Low-Quality for NMT

اسم الطالبة هو سالي .سالي بنت ذكية تدرس كثيرا .إنها تذهب إلى المكتبة في عطلة نهاية الأسبوع .

What makes clients Happy	What makes NMT happy
The student's name is Sally	The student's name is Sally
	Sally is a smart kid who studies a lot.
She studies a lot and goes to the library on weekends.	She goes to the library on weekends.

Transcreation (Examples)

 7
 8
 1
 2
 7
 6
 9
 7
 2

 6
 5
 6
 6
 9
 9
 7
 2

 7
 4
 6
 6
 4
 2
 1

 6
 1
 4
 7
 4
 2
 7

 1
 5
 6
 6
 7
 5
 4
 3
 9

 4
 8
 8
 3
 4
 7
 4
 5
 3

 8
 3
 7
 8
 4
 1
 1
 6

 5
 4
 2
 7
 7
 1

 2
 7
 3
 6
 3

 8
 7
 7
 6

Source	Human Transcreation	Literal Translation
أحد الأسئلة العميقة التي باتت ترتفع في سماوات العلاقات العربية - الأميركية عامة، والسعودية الأميركية خاصة	One of the biggest questions facing Arab- US relations in general, and Saudi-US ties in particular.	One of the deep questions that is rising in the heavens of Arab-American relations in general, and Saudi-US relations in particular.
يمكن القطع بأن المخاوف مشروعة ولا شك، ولا توجد عملية سي اسية تجري في فردوس للأطهار ، بل على أرض الأشواك، حيث ال خير والشر يتلازمان منذ بداية الأيام إلى أن يرث الله الأرض ومن عليها	These fears are assuredly legitimate, as no political process is ideal and error-free in this world where good and evil have existed since the dawn of time.	There is no political process taking place in Ferdous, but on the land of the thorns, where good and evil have been in flux from the beginning of the days until God inherits the land and those on it.

torjama

Client-Specific Requests?

- Light Post Editing
- Special Terminology
- Extreme localization
- Specific dialects

tarj<u>am</u>e

Huge Single Data Sources - Make sure your diet is varied!

Examples:

- UN data
- Huge projects

It will result in:

- Repeated mistakes
- Repeated topics/information
- If the entire data is used, the result will be client specific instead of generic. overfitting
- Will affect the Terminology usage
- * Do not eat a lot of the same thing, even if it is healthy. That is not a healthy diet!

Do you need BIG Data?

- We throw data more than we keep.
- Small but healthy.

tarjama

Data Creation

* Happy customer ≠ Happy NMT <u>WHY not to consider NMT as our customer!</u>

tarjama

Creation Vs. Acquisition

Data Acquisition		Data Creation
Cheap - Cost is paid by Clients		Expensive - Cost is on our Budget
Linguists are trained to create this type of data		Special guidelines that can easily be missed
Alignment issues		No alignment issues
Transcreation)) –	No transcreation
Domain depends on clients' projects	Ø	Carefully selected data and domains

torjama

Guidelines for NMT Data Creation

- Select domains that you lack.
- Avoid generic articles. Aim for specialized ones for richer terminology.
- Educate linguists and PMs on how NMT learns.
- Explain common data issues: Like transcreation, alignment ...
- Ensure that the source is high quality. Proofread the source if needed.
- Do not use MT
- Iterate: Do not operate a big project with a huge budget for a specific domain. Start small.

tarjam<u>e</u>

A Cheaper Solution

Filter existing good data using Data Creation guidelines!







https://translate.tarjama.com

omt

arabic machine translation by tarjama