tarjama

Deep Learning segmentation and alignment for Arabic

Unlocking the value of bilingual translated documents

Nour Al-Khdour – Applied ML Scientist
Dr. Rebecca Jonsson – Chief Product Officer

AMTA 2022 ORLANDO

Try our NMT at:
https://translate.tarjama.com

amt
arabic machine translation by tarjama

tarjama

Breaking language barriers with Arabic language technology

AI-Enabled LSP

Tarjama is the leading AI-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.

+600
Retained clients

+10 Arabic dialects supported

98% Customer retention rate

5 On-ground offices in MENA

+10 Million
In funding secured to date

85K
Freelancers

+2 Billion
Words processed

49%
Females

Arabic Language – Fun Facts

6th Most spoken language

1.1% Only of top 10M websites use Arabic

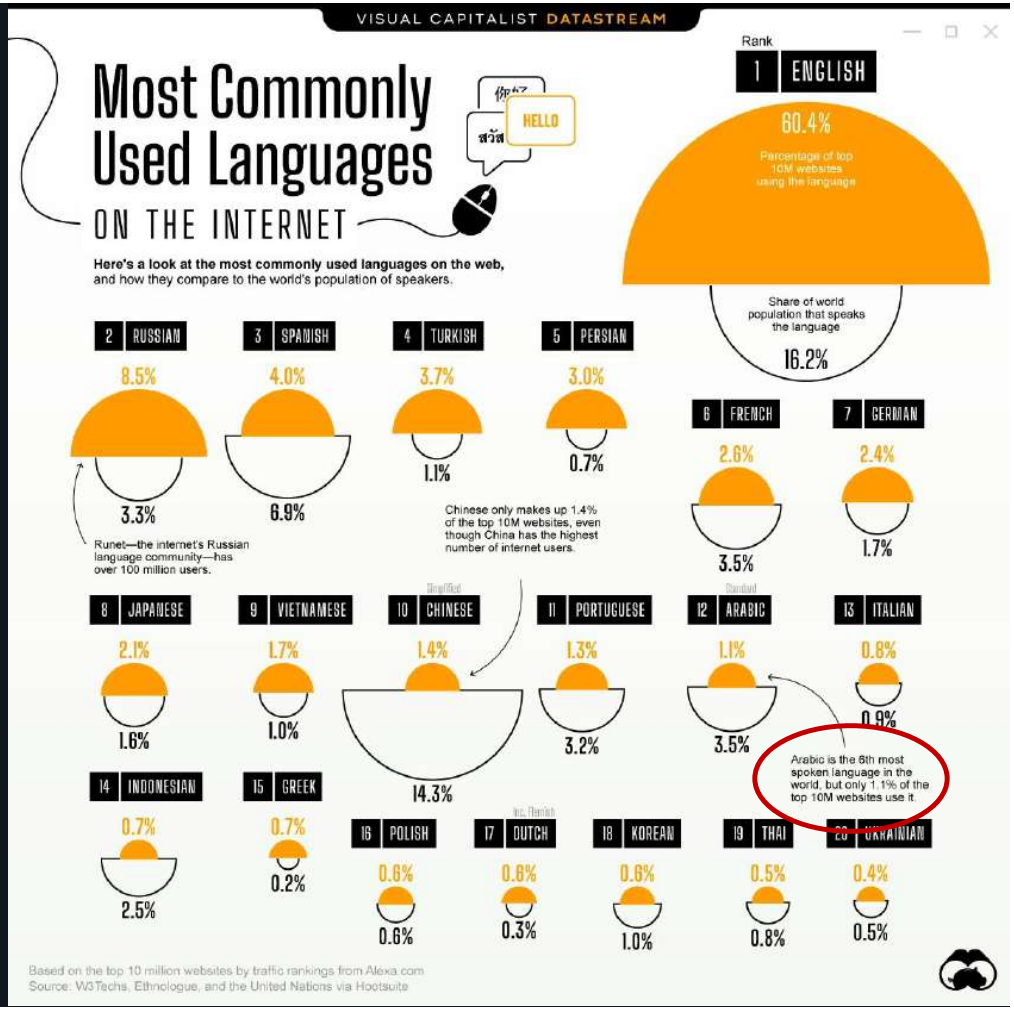
75% Internet penetration (but over 90% in UAE, Qatar). Big growth!

1 of 6 Official UN languages

Scarcity of parallel data & low quality

Scarcity of NLP tools & low quality

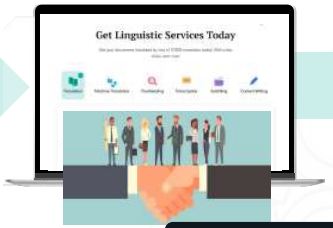
tarjama



tarjama Platform

360° Linguistic Services, **ONE** Hub

t-portal



t-staff

Language Service platform

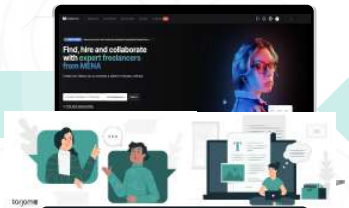
Language services

- Translation
- Transcreation
- Transcription
- Proofreading
- Content writing
- Copy editing
- Video editing
- Media editing
- DTP
- Stamping
- Diacritization
- Subtitling

CleverSo & amt

TMS&CAT tool with Arabic NMT

Ureed.com



Talent Marketplace

How to unlock the value of bilingual translated documents?

Potential docs to unlock

Old translated documents (before CAT tool usage)

Crawled corpora

Documents from clients

Old TMs

Potential Approaches

Manual alignment: time-consuming, tedious and expensive

Available sentence segmentation: for Arabic, performing so and so…

What to do?

tarjama

Arabic Sentence Segmentation

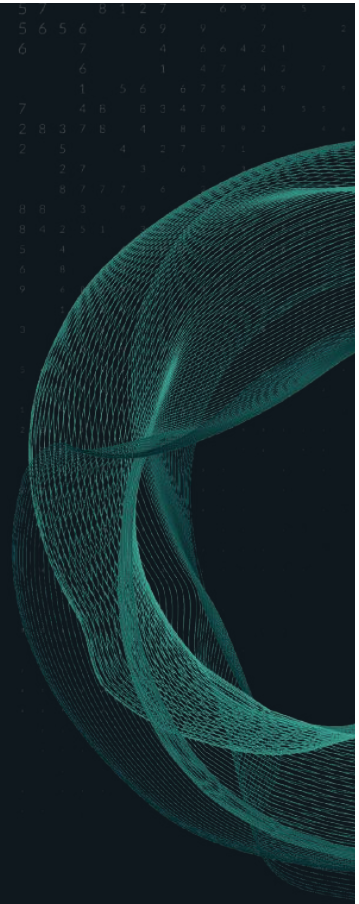
Challenges

Detect the sentence boundary based on the context, not rule-based.

Ambiguity of full stops.

Arabic has no capital letters.

Arabic has different punctuation marks, such as (comma “،”, and question mark “؟”).



tarjama

Available Tools for Arabic Sentence Segmentation

Model	Approach	Support Arabic	Notes
AraNLP	ML	✔	
SAFAR	Rules-based + ML	✔	
pySBD	Rules-based	✔	
NLTK	unsupervised approach	✖	Modified to support the Arabic question mark.

Table 1: Information on Available Arabic Sentence Segmentation Tools

tarjama

Evaluation

Automatic Unit Testing

Automatically synthesized testing set.

Comprises of ~4.5k examples.

Evaluation Metrics: Accuracy, Precision, Recall and F1 Score.

Use Cases: Exclamation and Question marks, Full Stop, Floating-Point Numbers, Abbreviations, List Numbering.

Manual Unit Testing

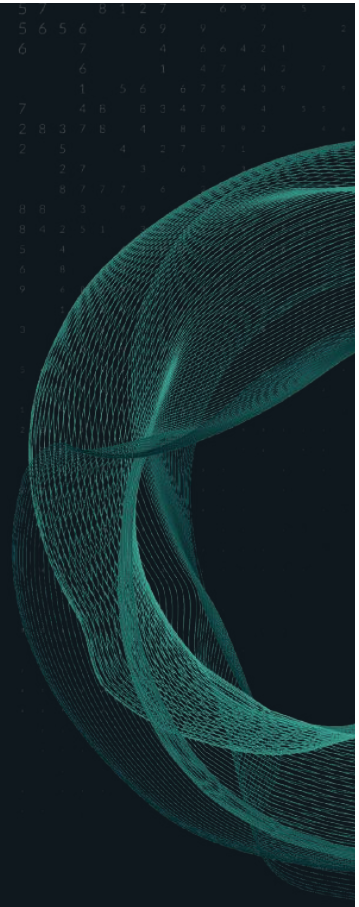
Manually prepared testing set.

Comprises of ~1.3k words.

Evaluation is done by Linguistic QA Experts

Use Cases: Next Slides!

tarjama



Manual Unit Testing Use Cases (With Examples)

Multiple Spaces

01

المدة: 99 تبدأ من: 17/04/1439هـ

Multiple Full stops

02

4 ..... نبذة عنا

Floating-Point Numbers

03

يُتوقع أن يصل الطلب على الأراضي الصناعية إلى حوالي 66.9 مليون متر مربع

Abbreviations

04

ق.م. (قبل الميلاد)

Brackets

05

يتم دفع أي رسوم مقطوعة(على سبيل المثال: الرسوم السنوية).

List Numbering

06

١. تعريف علامات الترقيم

tarjama

Manual Unit Testing Use Cases (With Examples)

07 Paragraphs with Full Stops

تسبب الحطام المحترق ومخلفات النفط بأضرار جسيمة في الساحل السريلانكي المجاور. إن حجم الضرر جعل من هذا الحادث من أسوأ الكوارث البيئية في سيرلانكا. أنقذت البحرية السريلانكية 25 فردًا من أفراد طاقم سفينة الشحن بعد أن دمرت الانفجارات أجزاءً منها. كما ساعدت البحرية الهندية في السيطرة على الحريق.

08 Paragraphs without Full Stops

إن تعهدات أي من الطرفين بالتعويض مشروطة : (أ) بقيام الطرف الذي يمنح له التعويض بتزويد الطرف المانح للتعويض بإشعار خطي عاجل عن أي مطالبة (شريطة أن يعفي الإخفاق في تقديم الإشعار بصورة عاجلة الطرف المانح للتعويض من تعهده فقط بالقدر التي يستطيع فيه أن يبين الضرر المادي من مثل ذلك الإخفاق)، (ب) بحيازة الطرف الذي يمنح له التعويض للسيطرة والسلطة الحصرية فيما يتعلق بالدفاع والتسوية عن أي مطالبة من ذلك القبيل

09 Multiple Cases

السماعات.. كيف تكون؟ على موقع يوتيوب، بثت قناة مهتمة بالشأن التقني تسجيلًا مصوراً يُظهر لأول مرة ما يعتقد أنها سماعات الأذن التي تمتاز بأنها تأتي مع وصلة Lightning بدلًا من موصل الصوت التقليدي 3.5 مم، وهي ما يشاع أنها ستأتي مع هاتف آبل المرتقب.

tarjama

Comparison: Available Arabic Sentence Segmentation Tools

Automated Unit Testing was conducted for the 4 tools.

SAFAR was the worst so it was excluded from the manual evaluation.

Model	Multiple Spaces	Multiple Full stops	Floating-Point Numbers	Abbreviations	Brackets	paragraph **with** Full Stops	Paragraph **without** Full Stops	List Numbering	Multiple Cases
AraNLP	**3**	**6**	**3**	**36**	**3**	**3**	**3**	**11**	**3**
PySBD	**6**	**7**	**0**	**36**	**4**	**18**	**19**	**5**	**4**
NLTK	**0**	**0**	**0**	**21**	**0**	**0**	**0**	**3**	**2**

Table 2: Performance of the available Arabic Segmentation Tools on the Manual Unit Testing

Our Linguistic QA Experts report poor performance of available tools as shown in the table! Hence, build our own!

tarjama

Tarjama Arabic Sentence Segmentation Experiments

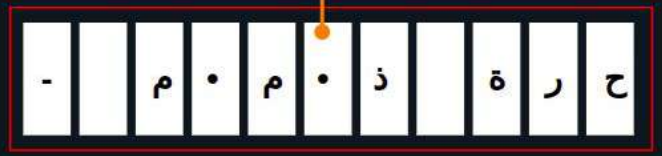
	Unsupervised Machine Learning (ML)	Deep Learning (DL)
Training Data Size	1477813	863821
Architecture	Punkt (Kiss & Strunk, 2006)	CNN bi-LSTM **LSTM**

Table 3: Information on Tarjama Arabic Sentence Segmentation Experiments

tarjama

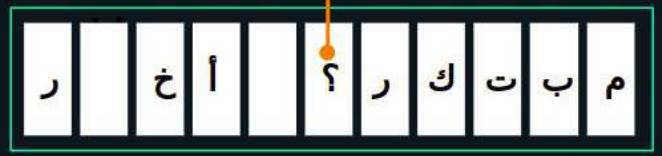
Tarjama DL Methodology for Arabic Sentence Segmentation

ترجمة منطقة حرة ذ.م.م - أبو ظبي



Segmentation ✖

هل أنت شخص مبتكر؟ أخبرني عن شيء مبتكر تفكر فيه الآن.



Segmentation ✓

tarjama

Comparison: Available Tools Vs. Tarjama Models

Model	Multiple Spaces	Multiple Full stops	Floating-Point Numbers	Abbreviations	Brackets	paragraph **with** Full Stops	Paragraph **without** Full Stops	List Numbering	Multiple Cases
AraNLP	**3**	**6**	**3**	**36**	**3**	**3**	**3**	**11**	**3**
PySBD	**6**	**7**	**0**	**36**	**4**	**18**	**19**	**5**	**4**
NLTK	**0**	**0**	**0**	**21**	**0**	**0**	**0**	**3**	**2**
Unsupervised ML (Tarjama)	**0**	**0**	**0**	**17**	**0**	**0**	**0**	**3**	**2**
Deep Learning (Tarjama)	**0**	**4**	**0**	**3**	**0**	**0**	**0**	**3**	**1**

Table 4: Appending Tarjama Arabic Segmentation Models Results on the Manual Unit Testing

tarjama

Tarjama Deep Learning Model highly Outperforms available Arabic Segmentation Tools!

Alignment Approaches

LASER (Language Agnostic Sentence Representations)

Extracted the embedding for both source and target files

Calculate the cosine similarity between the segment in the source file with five segments above and below the target segment.

Chose the aligned sentences based on the highest similarity score.

Different cosine similarity threshold experimented, the best threshold was 0.70.

BLEUAlign

Translate the source file into the target file language using MT.

Chose the aligned sentences based on the modified BLEU score

Both direction are experimented (English-Arabic, and Arabic-English), the best was using Arabic-English.

tarjama

Alignment Scores

Model	No. of Aligned Segments	Precision	Recall	F1-Score
LASER	1643	94.21	88.25	91.13
BLEUALign	1649	94.60	88.93	91.68

Table 6: Results of Alignment Approaches on Automatic Evaluation Test Set

tarjama

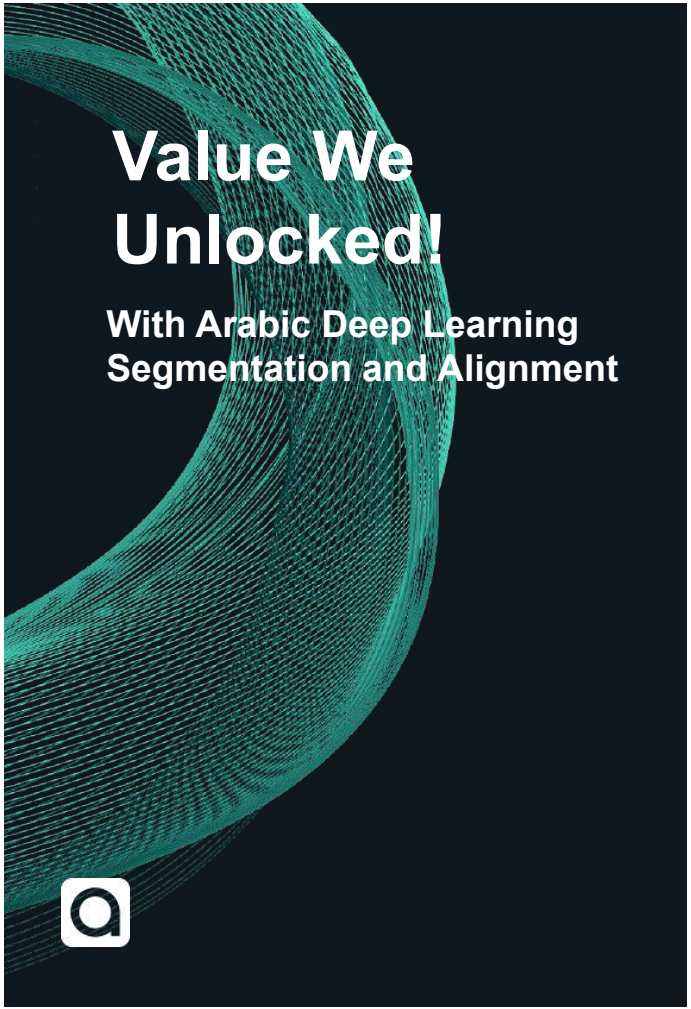
Unlocking OLD Tarjama Data

Unlocking the value of ~60 GB of archived Bilingual documents translated by Tarjama before usage of CAT tools (2008-2016). Data was extracted, segmented and aligned by our Deep Learning model to produce TMs and Parallel Data.

	Sentences		Tokens			
	English	Arabic	English	%	Arabic	%
Original Data	1502878	1482443	7793757	100	7948912	100
BLEUAlign	1164634		6362165	81.6	6415418	80.7

Table 6: Coverage on English-Arabic Old Tarjama Data

tarjama



Old Tarjama data

Unlocking the value of ~60 GB of archived Bilingual documents translated by Tarjama before usage of CAT tools (2008-2016). Data was extracted, segmented and aligned by our Deep Learning model to produce TMs and Parallel Data.

Crawled Comparable corpora

Unlocked GlobalVoices and WorldBank crawled comparable corpora. Allowed us to feed our Generic NMT with this data for EN->AR

Creating TMs from Bilingual Docs

Allows us to create TMs from previous data that a client has translated outside CAT tools. Something quite common in the MENA region.

First automated TM for an e-commerce client

E-commerce client shared translated product descriptions which they wanted to be imported into our CAT tool as a TM. The problem: each entry was a large bulk of non-segmented text. In order to make use of this as a TM in our CAT tool, each entry had to be segmented and aligned into a new TM automatically.

With our new Arabic Deep Learning Segmentation and Alignment approach, we aligned these documents of over 300K words in one day. Would have taken weeks or even months to do manually! Happy client!

THANK YOU

شكرًا

https://translate.tarjama.com