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
Breaking language barriers with Arabic language technology

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 Million In funding secured to date
+2 Billion Words processed
+10 Arabic dialects supported
98% Customer retention rate
85K Freelancers
5 On-ground offices in MENA
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
360° Linguistic Services, ONE Hub

Language Service platform

TMS&CAT tool with Arabic NMT

Talent Marketplace

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

tarjamé
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 “؟”).
## Available Tools for Arabic Sentence Segmentation

<table>
<thead>
<tr>
<th>Model</th>
<th>Approach</th>
<th>Support Arabic</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraNLP</td>
<td>ML</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>SAFAR</td>
<td>Rules-based + ML</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>pySBD</td>
<td>Rules-based</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>NLTK</td>
<td>unsupervised approach</td>
<td>✗</td>
<td>Modified to support the Arabic question mark.</td>
</tr>
</tbody>
</table>

**Table 1: Information on Available Arabic Sentence Segmentation Tools**
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!
<table>
<thead>
<tr>
<th>Multiple Spaces</th>
<th>01</th>
<th>Abbreviations</th>
<th>04</th>
</tr>
</thead>
<tbody>
<tr>
<td>المدة: 99 تبدأ من: 99/04/1349 هـ</td>
<td></td>
<td>ق.م. (قبل الميلاد)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multiple Full stops</th>
<th>02</th>
<th>Brackets</th>
<th>05</th>
</tr>
</thead>
<tbody>
<tr>
<td>نبذة عنا: 4</td>
<td></td>
<td>يتم دفع أي رسوم مقطوعة (على سبيل المثال: الرسوم السنوية).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Floating-Point Numbers</th>
<th>03</th>
<th>List Numbering</th>
<th>06</th>
</tr>
</thead>
<tbody>
<tr>
<td>يتوقع أن يصل الطلب على الأراضي الصناعية إلى حوالي 6.9 مليون متر مربع.</td>
<td></td>
<td>تعريف علامات الترقيم</td>
<td></td>
</tr>
</tbody>
</table>

Manual Unit Testing Use Cases (With Examples)
تسببت الحطام المحترق ومخلفات النفط بأضرار جسيمة في الساحل السريلانكي المجاور. إن حجم الضرر جعل من هذا الحادث من أسوأ الكوارث البيئية في سريلانكا. أغلقت البحرية السريلانكية 25 فرداً من أفراد طاقم سفينة الشحن بعد أن دمرت الانفجارات أجزاء منها. كما ساعدت البحرية الهندية في السيطرة على الحريق.

إن تعهدات أي من الطرفين بالتعويض مشروطة: (أ) بقيام الطرف الذي يمنح له التعويض بتزويد الطرف المانح للتعويض بإشعار خطي عاجل عن أي مطالبة (شرطية أن يعفي الإشعار بصورة عاجلة الطرف المانح للتعويض من تعهد فقط بالقدر التي يستطيع فيه أن يبين الضرر المادي من مثل ذلك الإخفاق)، (ب) بحثية الطرف الذي يمنح له التعويض للسيطرة والسليمة الحصرية فيما يتعلق بالدفاع والتسويء عن أي مطالبة من ذلك القبيل.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Multiple Spaces</th>
<th>Multiple Full stops</th>
<th>Floating-Point Numbers</th>
<th>Abbreviations</th>
<th>Brackets</th>
<th>paragraph with Full Stops</th>
<th>Paragraph without Full Stops</th>
<th>List Numbering</th>
<th>Multiple Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraNLP</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>36</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>PySBD</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>36</td>
<td>4</td>
<td>18</td>
<td>19</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>NLTK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

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 Arabic Sentence Segmentation Experiments

<table>
<thead>
<tr>
<th></th>
<th>Unsupervised Machine Learning (ML)</th>
<th>Deep Learning (DL)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Data Size</strong></td>
<td>1477813</td>
<td>863821</td>
</tr>
<tr>
<td><strong>Architecture</strong></td>
<td>Punkt (Kiss &amp; Strunk, 2006)</td>
<td>CNN bi-LSTM LSTM</td>
</tr>
</tbody>
</table>

Table 3: Information on Tarjama Arabic Sentence Segmentation Experiments
Tarjama DL Methodology for Arabic Sentence Segmentation

ترجمة منطقة حرة ذم م - أبو طبي

Segmentation ✗

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

Segmentation ✓
## Comparison: Available Tools Vs. Tarjama Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Multiple Spaces</th>
<th>Multiple Full stops</th>
<th>Floating-Point Numbers</th>
<th>Abbreviations</th>
<th>Brackets paragraph with Full Stops</th>
<th>Brackets paragraph without Full Stops</th>
<th>List Numbering</th>
<th>Multiple Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraNLP</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>36</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>PySBD</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>36</td>
<td>4</td>
<td>18</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>NLTK</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Unsupervised ML (Tarjama)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Deep Learning (Tarjama)</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

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

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.
### Alignment Scores

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of Aligned Segments</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASER</td>
<td>1643</td>
<td>94.21</td>
<td>88.25</td>
<td>91.13</td>
</tr>
<tr>
<td>BLEUALign</td>
<td>1649</td>
<td>94.60</td>
<td>88.93</td>
<td>91.68</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Original Data</th>
<th>Sentences</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>1502878</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>7793757</td>
</tr>
<tr>
<td></td>
<td>Arabic</td>
<td>7948912</td>
</tr>
</tbody>
</table>

Table 6: Coverage on English-Arabic Old Tarjama Data
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

https://translate.tarjama.com