# المدونات المستخدمة Corpus

تم استخدام المدونات الخاصة بالأمم المتحدة (Ziemski, Junczys-Dowmunt and Pouliquen 2016) The United Nations Parallel Corpus v1.0 في عملية تدريب و اختبار الشبكة العصبية و تقييمها

وهي عبارة عن السجلات الرسمية, و الوثائق و المناقشات لمؤسسات الأمم المتحدة و هذه السجلات مكتوبة باللغات الست الرسمية للأمم المتحدة، هذه الوثائق و السجلات قد تم جمعها و ترجمتها بشريا بين عامي 2014-1990 ، كما تم مقابلة هذه الجمل يدوياً أيضاً.

تتألف مدونات الأمم المتحدة من أجل ثنائيات الجمل العربية والإنكليزية من عدة وثائق بلغ عدد هذه الوثائق 111241 وثيقة , بلغ عدد ثنائيات الجمل المتناظرة 18,539,207، و كان عدد الكلمات العربية ضمن هذه الوثائق 456,552,223 كلمة، في حين عدد الكلمات الإنكليزية 512,087,009 كلمة، كما بلغ حجم هذه المدونة 6GigaByte.

السبب الرئيسي للاعتماد على هذه المدونة أن ترجمتها ذات نوعية جيدة كونها تصدر عن مؤسسات رسمية، إضافة إلى أن الأسلوب المتبع في الترجمة و اللغة الرسمية المستخدم يجعلها مصدراً جيداً للتدريب على اللغة العربية الفصحى الحديثة.

تم اختيار مليون زوج من الجمل المتناظرة من هذه المدونة لاستعمالها كبيانات تدريب (traing set), و تم اختيار 300,000 زوج كبيانات تحقق(development set) و اختيار 70000 زوج كبيانات اختيار(Test set) .

* تم إعادة ترتيب الأزواج عشوائياً للتخلص من أية علاقات بين الجملة و سابقاتها.
* تم اختيار الأزواج التي طول جملها من 5 إلى 30 كلمة فقط، و ذلك لكون زيادة طول الجملة سوف يؤدي إلى زيادة وقت التدريب.
* تم استبعاد الجمل التي تحتوي الأرقام.

أستخدم في بناء بعض النماذج و لحل مشكلة الكلمات النادرة خوارزمية Byte Pair encoding (Sennrich, Haddow and Birch 2016) وذلك لتدريب الشبكة العصبية على مستوى أجزاء الكلمة وذلك بالنسبة للغتين العربية و الإنكليزية.

أستعمل محلل (Abdelali, et al. 2016)FARASA محلل أجل معالجة كلمات اللغة العربية و استخلاص ملحقات الكلمة.

# بناء الشبكة العصبية والتدريب

تم بناء الشبكة العصبية باستخدام لغة البرمجة Python و بالاعتماد على مكتبة Tensorflow و هي عبارة عن مكتبة مفتوحة المصدر تم تطويرها من شركة Google لبناء أنظمة تعليم الآلة و الذكاء الصنعي.

أما بالنسبة لتدريب الشبكة العصبية و الاختبارات تم على حاسب ذي المواصفات التالية:

* المعالج :8th Generation Intel® Core™ i3 Processors
* الذاكرة 16 GB RAM DDR4
* بطاقة الاظهار(GPU):NIVIDIA GTX 1050ti

حيث أن بطاقة الاظهار(GPU) هي أهم عنصر لتدريب الشبكات العصبية لأنها تحتوي عدد أكبر من الأنوية مقارنة بالمعالج مما يمكن توزيع العمليات الحسابية على هذه الأنوية، وبالتالي ينتج لدينا تدريب سريع للشبكة العصبية مقارنة مع تدريبها على المعالج لوحده.

|  |
| --- |
| gpu2 |
| **شكل(19)** مقارنة بين CPU و GPU من حيث عدد الأنوية **, يمتللك الـGPU عدد أكبر من الأنوية و بالتالي زيادة عدد العمليات الحسابية البسيطة التي يمكن إجرائها على التفرع.** |

|  |
| --- |
| https://image.slidesharecdn.com/w69ejoigsewjwetvbmwx-signature-600322f58dd0972f78d9395dd13259e83b846b0f09d6bd4f43956745ecd4e6f8-poli-150502101952-conversion-gate02/95/gpu-power-consumption-and-performance-trends-14-638.jpg?cb=1430578632 |
| **شكل(20)** مقارنة بين CPU و GPU من حيث البنية الفيزيائية (Pietro and Alessio n.d.) |

أصبح تدريب الشبكات العصبية باستخدام الـGPU من شركة Nvidia هو المعيار الرئيسي عند تدريب الشبكات العصبية الكبيرة DeepLearnign و ذلك باستخدام مكتبة CUDA التي تقدمها شركة Nvidia لتدريب الشبكات العصبية.

# النماذج التي تم بنائها

## ar\_en\_enc\_dec

تم بناء هذا النموذج للترجمة من العربية إلى الإنكليزية وفقاُ لتوصيف (Sutskever, Vinyals and Le 2014) و التي وضحناها بنيتها في الفقرة (رقم الفقرة), وذلك لبناء نموذج للترجمة الآلية بالاعتماد شبكات Encoder-Decooder للترجمة الآلية.

بالنسبة للبيانات التدريب فقد تم استخدام مدونات الأمم المتحدة كبيانات للتدريب حيث تم اختيار حوالي مليون جملة كأزواج لبيانات التدريب، ومن أجل مفردات اللغة تم أعتبارأول 50000 كلمة أكثر تكرارا بالنسبة للغة العربية و23000 كلمة بالنسبة للغة الإنكليزية أم الكلمات الأقل تكرار من ذلك فقد تم استبدالها بـ UNK، ولم تجرى أية معالجة خاصة على الكلمات سواءً العربية أو الإنكليزية.

أما بالنسبة للشبكة العصبية لها المكونات التالية:

بالنسبة للـencoder فيتألف من طبقة embbeding بـ 128 وحدة و أربع طبقات من خلايا LSTM عدد الخلايا في كل طبقة 128 خلية طبقتين لتمثيل الدخل بالاتجاه الصحيح و طبقتين لتمثيل الدخل بالاتجاه المعكوس

و بالنسبة للـdecoder فيتألف من أربعة طبقات في الخرج من خلايا LSTM.

تم تدريب الشبكة على دفعات batches كل batch يتألف من 128 زوج من الجمل و باستخدام طريقة Stochastic gradient descent وعدد خطوات التدريب مليون خطوة خطوة .

تم تهيئة أوزان الشبكة عشوائياً من المجال وباستخدام معدل تدريب 0.5 وانقاص معدل التدريب تدريجياً.

تم استخدام Dropout وبمعدل 0.2 و ذلك لتحسين التدريب ومنع الـoverfitting.

استغرق تدريب الشبكة العصبية 5 أيام تقريباً، وباستعمال 4Gbyte GPUعدد واحد فقط.

كانت نتائج تدريب النموذج وبحسب اختبارها على مجموعة الاختبار وفق معيار BLEU موزعة على أطوال الجمل كما يلي:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **طول الجملة (كلمة)** | **5-10** | **11-15** | **16-20** | **20-25** | **>25** | **الكلي** |
| **معيار BLEU** | 26.21 | 24.05 | 24.01 | 22.43 | 21.25 | 22.57 |



## en\_ar\_enc\_dec

تم بناء هذا النموذج للترجمة من الإنكليزية إلى العربية وفقاُ لتوصيف (Sutskever, Vinyals and Le 2014) و التي وضحناها بنيتها في الفقرة (رقم الفقرة), وذلك لبناء تموذج للترجمة الآلية بالاعتماد شبكات Encoder-Decooder للترجمة الآلية.

بالنسبة للبيانات التدريب فقد تم استخدام مدونات الأمم المتحدة كبيانات للتدريب حيث تم اختيار حوالي مليون جملة كأزواج لبيانات التدريب، ومن أجل مفردات اللغة تم أعتبارأول 50000 كلمة أكثر تكرارا بالنسبة للغة العربية و23000 كلمة بالنسبة للغة الإنكليزية أم الكلمات الأقل تكرار من ذلك فقد تم استبدالها بـ UNK، ولم تجرى أية معالجة خاصة على الكلمات سواءً العربية أو الإنكليزية.

أما بالنسبة للشبكة العصبية لها المكونات التالية:

بالنسبة للـencoder فيتألف من طبقة embbeding بـ 128 وحدة و أربع طبقات من خلايا LSTM عدد الخلايا في كل طبقة 128 خلية طبقتين لتمثيل الدخل بالاتجاه الصحيح و طبقتين لتمثيل الدخل بالاتجاه المعكوس

و بالنسبة للـdecoder فيتألف من أربعة طبقات في الخرج من خلايا LSTM.

تم تدريب الشبكة على دفعات batches كل batch يتألف من 128 زوج من الجمل و باستخدام طريقة Stochastic gradient descent وعدد خطوات التدريب مليون خطوة خطوة .

تم تهيئة أوزان الشبكة عشوائياً من المجال وباستخدام معدل تدريب 0.5 وانقاص معدل التدريب تدريجياً.

تم استخدام Dropout وبمعدل 0.2 و ذلك لتحسين التدريب ومنع الـoverfitting.

استغرق تدريب الشبكة العصبية 5 أيام تقريباً، وباستعمال 4Gbyte GPUعدد واحد فقط.

كانت نتائج تدريب النموذج وبحسب اختبارها على مجموعة الاختبار وفق معيار BLEU اعتبارا لطول الجملة هو

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **طول الجملة** | **5-10** | **11-15** | **16-20** | **20-25** | **>25** | **الكلي** |
| **معيار BLEU** | 22.13 | 21.18 | 20.17 | 18.14 | 17.52 | 20.19 |

## ar\_en\_attention

تم بناء هذا النموذج للترجمة من العربية إلى الإنكليزية وفقاُ لتوصيف (Bahdanau, Cho and Bengio 2015) و التي وضحناها بنيتها في الفقرة (رقم الفقرة), وذلك لبناء تموذج للترجمة الآلية بالاعتماد شبكات , Encoder-Decooder المعتمدة على آلية الانتباه Attension Mechanism للترجمة الآلية.

بالنسبة للبيانات التدريب فقد تم استخدام مدونات الأمم المتحدة كبيانات للتدريب حيث تم اختيار حوالي مليون جملة كأزواج لبيانات التدريب، ومن أجل مفردات اللغة تم أعتبارأول 50000 كلمة أكثر تكرارا بالنسبة للغة العربية و23000 كلمة بالنسبة للغة الإنكليزية أم الكلمات الأقل تكرار من ذلك فقد تم استبدالها بـ UNK، ولم تجرى أية معالجة خاصة على الكلمات سواءً العربية أو الإنكليزية.

أما بالنسبة للشبكة العصبية لها المكونات التالية:

بالنسبة للـencoder فيتألف من طبقة embbeding بـ 128 وحدة و أربع طبقات من خلايا LSTM عدد الخلايا في كل طبقة 128 خلية طبقتين لتمثيل الدخل بالاتجاه الصحيح و طبقتين لتمثيل الدخل بالاتجاه المعكوس

و بالنسبة للـdecoder فيتألف من أربعة طبقات في الخرج من خلايا LSTM.

و بالنسبة لآلية الانتباهattension mechanism فقد تم الاعتماد على global attention الموضخ بالفقرة (رقم الفقرة)

تم تدريب الشبكة على دفعات batches كل batch يتألف من 128 زوج من الجمل و باستخدام طريقة Stochastic gradient descent وعدد خطوات التدريب مليون خطوة خطوة .

تم تهيئة أوزان الشبكة عشوائياً من المجال وباستخدام معدل تدريب 0.5 وانقاص معدل التدريب تدريجياً.

تم استخدام Dropout وبمعدل 0.2 و ذلك لتحسين التدريب ومنع الـoverfitting.

استغرق تدريب الشبكة العصبية سبعة أيام تقريباً، وباستعمال 4Gbyte GPUعدد واحد فقط.

لحل مشكلة الرمز unkفي خرج الشبكة العصبية تم الاعتماد على آلية الانتباه attention mechanism و التي تعطي درجات المقابلة بين كلمات الجملة المصدر و جملة الخرج حيث تم الاعتماد على أعلى درجة مقابلة لتحديد الكلمة في الجملة المصدرية و المقابلة للرمز unk و بعدها تم ترجمة هذه الكلمة باستخدام قاموس ثنائي اللغة.

كانت نتائج تدريب النموذج وبحسب اختبارها على مجموعة الاختبار وفق معيار BLEU موزعة على أطوال الجمل كما يلي:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **طول الجملة (كلمة)** | **5-10** | **11-15** | **16-20** | **20-25** | **>25** | **الكلي** |
| **Unk بدون معالجة** | 37.21 | 35.05 | 34.98 | 35.15 | 35.07 | 35.4 |
| **No unk** | 33.21 | 37.05 | 33.23 | 36.12 | 35.47 | 35.69 |

## en\_ar\_attention

تم بناء هذا النموذج للترجمة من الإنكليزية إلى العربية وفقاُ لتوصيف (Bahdanau, Cho and Bengio 2015) و التي وضحناها بنيتها في الفقرة (رقم الفقرة), وذلك لبناء تموذج للترجمة الآلية بالاعتماد شبكات , Encoder-Decooder المعتمدة على آلية الانتباه Attension Mechanism للترجمة الآلية.

بالنسبة للبيانات التدريب فقد تم استخدام مدونات الأمم المتحدة كبيانات للتدريب حيث تم اختيار حوالي مليون جملة كأزواج لبيانات التدريب، ومن أجل مفردات اللغة تم أعتبارأول 50000 كلمة أكثر تكرارا بالنسبة للغة العربية و23000 كلمة بالنسبة للغة الإنكليزية أم الكلمات الأقل تكرار من ذلك فقد تم استبدالها بـ UNK، ولم تجرى أية معالجة خاصة على الكلمات سواءً العربية أو الإنكليزية.

أما بالنسبة للشبكة العصبية لها المكونات التالية:

بالنسبة للـencoder فيتألف من طبقة embbeding بـ 128 وحدة و أربع طبقات من خلايا LSTM عدد الخلايا في كل طبقة 128 خلية طبقتين لتمثيل الدخل بالاتجاه الصحيح و طبقتين لتمثيل الدخل بالاتجاه المعكوس

و بالنسبة للـdecoder فيتألف من أربعة طبقات في الخرج من خلايا LSTM.

و بالنسبة لآلية الانتباه فقد تم الاعتماد على global attention الموضخ بالفقرة (رقم الفقرة)

تم تدريب الشبكة على دفعات batches كل batch يتألف من 128 زوج من الجمل و باستخدام طريقة Stochastic gradient descent وعدد خطوات التدريب مليون وثلاثمئة ألف خطوة .

تم تهيئة أوزان الشبكة عشوائياً من المجال وباستخدام معدل تدريب 0.5 وانقاص معدل التدريب تدريجياً.

تم استخدام Dropout وبمعدل 0.2 و ذلك لتحسين التدريب ومنع الـoverfitting.

استغرق تدريب الشبكة العصبية ثمانية أيام تقريباً، وباستعمال 4Gbyte GPUعدد واحد فقط.

لحل مشكلة الرمز unkفي خرج الشبكة العصبية تم الاعتماد على آلية الانتباه attention mechanism و التي تعطي درجات المقابلة بين كلمات الجملة المصدر و جملة الخرج حيث تم الاعتماد على أعلى درجة مقابلة لتحديد الكلمة في الجملة المصدرية و المقابلة للرمز unk و بعدها تم ترجمة هذه الكلمة باستخدام قاموس ثنائي اللغة.

كانت نتائج تدريب النموذج وبحسب اختبارها على مجموعة الاختبار وفق معيار BLEU موزعة على أطوال الجمل كما يلي:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **طول الجملة (كلمة)** | **5-10** | **11-15** | **16-20** | **20-25** | **>25** | **الكلي** |
| **Unk بدون معالجة** | 28.88 | 28.47 | 27.49 | 28.45 | 27.66 | 27.45 |
| **No unk** | 28.45 | 29.55 | 28.97 | 29.87 | 29.45 | 28.41 |

## ar\_en\_bpe\_attention

تم بناء هذا النموذج للترجمة من العربية إلى الإنكليزية وفقاُ لتوصيف (Bahdanau, Cho and Bengio 2015) و التي وضحناها بنيتها في الفقرة (رقم الفقرة), وذلك لبناء تموذج للترجمة الآلية بالاعتماد شبكات , Encoder-Decooder المعتمدة على آلية الانتباه Attension Mechanism للترجمة الآلية، و الترجمة على مستوى أجزاء الكلمة (Sennrich, Haddow and Birch 2016) خيث أن هذه الطريقة لا تقوم بتوليد كلمات مجهولة.

حيث تقوم بتقسيم الكلمة إلى أجزاء أصغر باستخدام خوارزمية Byte Pair ENcoding Pbe , تم معالجة نصوص اللغة الإنكليزية باستخدام BPE , أم بالنسبة للغة العربية فقد تم استعمال محلل Farasa (Abdelali, et al. 2016) و من ثم تم استعمال BPE أيضاً على الخرج الناتج.

أما بالنسبة للشبكة العصبية لها المكونات التالية:

بالنسبة للـencoder فيتألف من طبقة embbeding بـ 128 وحدة و أربع طبقات من خلايا LSTM عدد الخلايا في كل طبقة 128 خلية طبقتين لتمثيل الدخل بالاتجاه الصحيح و طبقتين لتمثيل الدخل بالاتجاه المعكوس

و بالنسبة للـdecoder فيتألف من أربعة طبقات في الخرج من خلايا LSTM.

و بالنسبة لآلية الانتباه فقد تم الاعتماد على global attention الموضخ بالفقرة (رقم الفقرة)

تم تدريب الشبكة على دفعات batches كل batch يتألف من 128 زوج من الجمل و باستخدام طريقة Stochastic gradient descent وعدد خطوات التدريب مليون خطوة خطوة .

تم تهيئة أوزان الشبكة عشوائياً من المجال وباستخدام معدل تدريب 0.5 وانقاص معدل التدريب تدريجياً.

تم استخدام Dropout وبمعدل 0.2 و ذلك لتحسين التدريب ومنع الـoverfitting.

استغرق تدريب الشبكة العصبية 10 أيام تقريباً، وباستعمال 4Gbyte GPUعدد واحد فقط.

كانت نتائج تدريب النموذج وبحسب اختبارها على مجموعة الاختبار وفق معيار BLEU موزعة على أطوال الجمل كما يلي:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **طول الجملة (كلمة)** | **5-10** | **11-15** | **16-20** | **20-25** | **>25** | **الكلي** |
| **معيار BLEU** | 17.21 | 15.75 | 12.25 | 13.15 | 12.17 | 14.06 |

## en\_ar\_bpe\_attention

تم بناء هذا النموذج للترجمة من الانكليزية إلى العربية وفقاُ لتوصيف (Bahdanau, Cho and Bengio 2015) و التي وضحناها بنيتها في الفقرة (رقم الفقرة), وذلك لبناء تموذج للترجمة الآلية بالاعتماد شبكات , Encoder-Decooder المعتمدة على آلية الانتباه Attension Mechanism للترجمة الآلية، و الترجمة على مستوى أجزاء الكلمة (Sennrich, Haddow and Birch 2016) خيث أن هذه الطريقة لا تقوم بتوليد كلمات مجهولة.

حيث تقوم بتقسيم الكلمة إلى أجزاء أصغر باستخدام خوارزمية Byte Pair ENcoding Pbe , تم معالجة نصوص اللغة الإنكليزية باستخدام BPE , أم بالنسبة للغة العربية فقد تم استعمال محلل Farasa (Abdelali, et al. 2016) و من ثم تم استعمال BPE أيضاً على الخرج الناتج.

أما بالنسبة للشبكة العصبية لها المكونات التالية:

بالنسبة للـencoder فيتألف من طبقة embbeding بـ 128 وحدة و أربع طبقات من خلايا LSTM عدد الخلايا في كل طبقة 128 خلية طبقتين لتمثيل الدخل بالاتجاه الصحيح و طبقتين لتمثيل الدخل بالاتجاه المعكوس

و بالنسبة للـdecoder فيتألف من أربعة طبقات في الخرج من خلايا LSTM.

و بالنسبة لآلية الانتباه فقد تم الاعتماد على global attention الموضخ بالفقرة (رقم الفقرة)

تم تدريب الشبكة على دفعات batches كل batch يتألف من 128 زوج من الجمل و باستخدام طريقة Stochastic gradient descent وعدد خطوات التدريب مليون خطوة خطوة .

تم تهيئة أوزان الشبكة عشوائياً من المجال وباستخدام معدل تدريب 0.5 وانقاص معدل التدريب تدريجياً.

تم استخدام Dropout وبمعدل 0.2 و ذلك لتحسين التدريب ومنع الـoverfitting.

استغرق تدريب الشبكة العصبية 10 أيام تقريباً، وباستعمال 4Gbyte GPUعدد واحد فقط.

كانت نتائج تدريب النموذج وبحسب اختبارها على مجموعة الاختبار وفق معيار BLEU موزعة على أطوال الجمل كما يلي:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **طول الجملة (كلمة)** | **5-10** | **11-15** | **16-20** | **20-25** | **>25** | **الكلي** |
| **معيار BLEU** | 12.05 | 10.25 | 8.26 | 9.15 | 8.39 | 9.47 |

# مقارنة النتائج مع دراسات أخرى

## Encoder-decoder Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | النماذج التي قمنا ببنائها | | (Sutskever, Vinyals and Le 2014) | (Menacer, et al. 2017) |
| اللغة | EN->AR | Ar->EN | EN->FR | EN-AR |
| حجم بيانات التدريب | مليون | مليون | 12 مليون | 100K |
| حجم المفردات | 23k EN,50k AR | 23k EN,50k AR | 160k EN,80k FR | 20k EN,74k AR |
| Hardware | 1GPU | 1GPU | 8GPU | - |
| Training | 5 days | 5 days | 10 days | - |
| Neural Network | Encoder:2 layers bidirectioal lstm \*128 unit  Decoder:4 layers lstm \*128 unit | Encoder:2 layers bidirectioal lstm \*128 unit  Decoder:4 layers lstm \*128 unit | Encoder:2 layers reversed LSTM units \*1000  Decoder:2 layers STM units \*1000 | Encoder:1 layers lstm \*100 unit  Decoder:1 layer lstm \*100 unit |
| BLEU | 20.19 | 22.57 | 34.81 | 5.35 |

## Encoder-decoder with Attension Mechanism Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | النماذج التي قمنا ببنائها | | (Bahdanau, Cho and Bengio 2015) | (Menacer, et al. 2017) |
| اللغة | EN->AR | Ar->EN | EN->FR | EN-AR |
| حجم بيانات التدريب | مليون | مليون | 36 مليون | 100K |
| حجم المفردات | 23k EN,50k AR | 23k EN,50k AR | 30k EN,30k FR | 20k EN,74k AR |
| Hardware | 1 GPU GTX 1050-Ti | 1 GPU GTX 1050-Ti | 1 GPU Quadro K-6000 | - |
| Training | 7 days | 7 days | 10 days | - |
| Neural Network | Encoder:2 layers bidirectioal lstm \*128 unit  Decoder:4 layers lstm \*128 unit  With global attension | Encoder:2 layers bidirectioal lstm \*128 unit  Decoder:4 layers lstm \*128 unit  With global attension | Encoder:1 Layer bidirectioal GRU units \*1000  Decoder: 1 Layer GRU units \*1000 | Encoder:1 layers bidirectioal lstm \*100 unit  Decoder:1 layer lstm \*100 unit |
| BLEU | 27.45 | 35.4 | 28.45 | 20.63 |
| BLEU with no Unk | 28.41 | 35.69 | 36.15 | 21.03 |

## Encoder-decoder with Attension Mechanism Models On subword Level

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | النماذج التي قمنا ببنائها | | (Sennrich, Haddow and Birch 2016) | (Almahairi, et al. 2016) | |
| اللغة | EN->AR | Ar->EN | EN->DE | EN->AR | Ar->EN |
| حجم بيانات التدريب | مليون | مليون | - | 1.2M | 1.2M |
| حجم المفردات | 23k EN,50k AR | 23k EN,50k AR | - | - | - |
| معالجة البيانات | EN:Bpe  Ar:Frasra+bpe | EN:Bpe  Ar:Frasra+bpe | - | EN:Bpe  Ar: MADAMIRA +bpe | EN:Bpe  Ar: MADAMIRA +bpe |
| Hardware | 1 GPU GTX 1050-Ti | 1 GPU GTX 1050-Ti | - |  |  |
| Training | 10 days | 10 days | 7 days | 7 days | 7 days |
| Neural Network | Encoder:2 layers bidirectioal lstm \*128 unit  Decoder:4 layers lstm \*128 unit  With global attension | Encoder:2 layers bidirectioal lstm \*128 unit  Decoder:4 layers lstm \*128 unit  With global attension | Encoder:1000 unit  Decoder:1000 unit | Encoder:1 layer bidirectioal GRU \*512 unit  Decoder:1 layers lstm \*512 unit | Encoder:1 layer bidirectioal GRU \*512 unit  Decoder:1 layers lstm \*512 unit |
| BLEU | 9.47 | 16.06 | 25.3 | 33.62 | 49.7 |

# ملخصات الأبحاث

## أبحاث الترجمة الآلية العصبية لغير اللغة العربية

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **اللغة** | **data** | | | **Hardware** | **Training** | **Neural Network** | **BLEU** |
| **بيانات التدريب** | **المفردات** | **المعالجة** |
| (Sutskever, Vinyals and Le 2014) | EN->FR | 12M | 160k EN,80k FR |  | 8GPU | 10 d | Encoder:2 layers reversed LSTM units \*1000  Decoder:2 layers STM units \*1000 | 34.81 |
| (Bahdanau, Cho and Bengio 2015) | EN->FR | 36M | 30k EN,30k FR |  | 1 GPU Quadro K-6000 | 10 d | Encoder:1 Layer bidirectioal GRU units \*1000  Decoder: 1 Layer GRU units \*1000  With attension | 28.45 |
| No unk  36.15 |
| (Sennrich, Haddow and Birch 2016) | EN->DE | - | - | BPE | - | 7 days | Encoder:1000 unit  Decoder:1000 unit  With attension | 25.3 |
| (Cho, et al. 2014) | EN->FR | 348M WORD | 30K |  |  | 5 d | Encoder: 2000 unit GRU  Decoder: 2000 unit  GRU attension | 13.92 |
| No unk  23.45 |
| (Jason, Cho and Hofmann 2017) | DE->EN | 4.5M | - | No processing NN works on char level | 1GPU | - | 8 convolution layers  Encoder:1 layer 512 gru  Decoser:2 layer 1024 gru | 25.01 |
| CS->EN | 12.1M | - | - | 15.96 |
| FI->EN | 1.9M | - | - | 15.74 |
| RU->EN | 2.3M | - | - | 26.33 |
| (Zheng, et al. 2018)، | DE->EN |  |  |  |  |  | Encoder-decoder with attension  Decoder has 2 extra layers to  represent sFutur and pastr | 29.5 |
| EN->DE |  |  |  |  |  | 24.3 |
| CH->EN |  |  |  |  |  | 37.84 |

## أبحاث الترجمة الآلية العصبية بالنسبة للغة العربية

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **اللغة** | **data** | | | **Hardware** | **Training** | **Neural Network** | **BLEU** |
| **بيانات التدريب** | **المفردات** | **المعالجة** |
| (Almahairi, et al. 2016) | AR->EN | 1.2M | - | EN:Bpe  Ar: MADAMIRA +bpe | - | 7d | Encoder:1 layer bidirectioal GRU \*512 unit  Decoder:1 layers lstm \*512 unit  With attension | 49.7 |
| EN->AR | 7d | 33.62 |
| (Ziemski, Junczys-Dowmunt and Pouliquen 2016) | EN-AR | 100K | 20k EN,74k AR | - | - | - | Encoder:1 layers lstm \*100 unit  Decoder:1 layer lstm \*100 unit | 5.35 |
| EN-AR | 100K | 20k EN,74k AR | - | - | - | Encoder:1 layers bidirectioal lstm \*100 unit  Decoder:1 layer lstm \*100 unit | 20.63 |
| No unk  21.03 |

## النماذج التي قمنا ببنائها

بالنسبة للبيانات التدريب فقد تم التدريب على مليون زوج من الجمل، ومن أجل مفردات اللغة تم أعتبارأول 50000 كلمة أكثر تكرارا بالنسبة للغة العربية و23000 كلمة بالنسبة للغة الإنكليزية،وكانت نتائج النماذج التي قمنا بتدريبها موضحة بالجدول التالي:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| اللغة | النموذج | البنية | التدريب | معالجة بيانات التدريب | BLUE |
| AR->EN | ar\_en\_enc\_dec | Encoder-Decoder | 5d | - | 22.57 |
| **ar\_en\_attention** | Encoder\_Decoder with Attension | 7d | - | 35.4 |
| **35.69\*** |
| n ar\_en\_bpe\_attentio | Encoder\_Decoder with Attension | 10d | EN:bpe  Ar:Frasra+bpe | 14.06 |
| EN->AR | en\_ar\_enc\_dec | Encoder-Decoder | 5d | - | 20.19 |
| **en\_ar\_attention** | Encoder\_Decoder with Attension | 7d | - | 27.45 |
| **28.41\*** |
| n en\_ar\_bpe\_attentio | Encoder\_Decoder with Attension | 10d | EN:bpe  Ar:Frasra+bpe | 9.47 |

\* تم معالحة الرمز Unk في الخرج

# المراجع

Abdelali, Ahmed, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. "Farasa: A fast and furious segmenter for arabic." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.* 2016. 11-16.

Amjad Almahairi، Kyunghyun Cho، Nizar Habash، و Aaron Courville. "First result on Arabic neural machine translation." *arXiv* ، 2016.

Dzmitry Bahdanau، Kyunghyun Cho، و Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *ICLR*، 2015.

Kyunghyun Cho، Bart Van Merriënboer، Dzmitry Bahdanau، و Yoshua Bengio. "On the properties of neural machine translation: Encoder-decoder approaches." *arXiv preprint arXiv:1409.1259*، 2014.

Mohamed Amine Menacer، David Langlois، Odile Mella، Dominique Fohr، و Denis Jouvet. "Is statistical machine translation approach dead." *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing.* 2017.

Pietro, Piscione, and Villardita Alessio. "GPU power consumption and performance trends." n.d. https://www.slideshare.net/AlessioVillardita/ca-1st-presentation-final-published.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." *arXiv preprint arXiv:1508.07909*, 2016.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems.* 2014. 3104-3112.

Ziemski, Michael, Marcin Junczys-Dowmunt, and Bruno Pouliquen. "The United Nations Parallel Corpus." *Language Resources and Evaluation. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia.* 2016.