مذكرة حول اللغة العربية والترجمة الآلية

A Note on Arabic and Machine Translation

ناقشنا في الفصول السابقة من هذا الكتاب اللغة العربية من وجهة نظر أحادية (monolingual). على النقيض من ذلك، يتناول هذا الفصل قضايا تعددية اللغات عند العمل مع اللغة العربية، وعلى وجه التحديد، تطبيق الترجمة الآلية (Machine Translation – MT). وبما أن التطبيق قد لا يكون مألوفاً لبعض القراء، سنعطي مقدمة قصيرة في القسم التالي لتعريف المصطلحات والمفاهيم الأساسية. كما ننصح القراء بالاطلاع على بعض المقالات العديدة، والكتب والمواقع التي توفر مقدمة أكثر شمولية في مجال الترجمة الآلية. وسيوفر ما تبقى من هذا الفصل نقاشاً حول السمات اللغوية العربية، من وجهة نظر مقارنة، مع وضع الترجمة الآلية في الاعتبار. يعقب ذلك مسح للموارد المتاحة وعرض لحالة مجال الترجمة الآلية العربية (من اللغة العربية وإلى اللغة العربية).

٨,١ مفاهيم أساسية في الترجمة الآلية

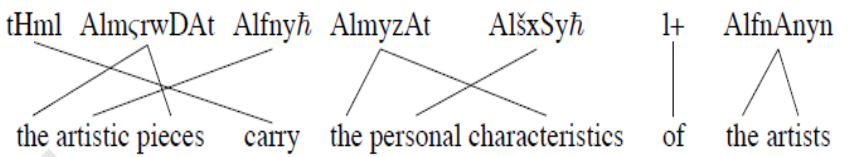
الترجمة الآلية هي تطبيق يعمل على التحويل من لغة (اللغة المصدر) لأخرى (اللغة الهدف). ويمكن تصنيف الطرق المختلفة للترجمة الآلية إلى منهجين: الترجمة

الرمزية/المستندة على القواعد (Rule-Based Machine Translation - RBMT) والترجمة الإحصائية / المعتمدة على المدونة (Statistical Machine Translation- SMT).

تتميز الترجمة بطريقة (RBMT) باستخدام ظاهر لقواعد وتمثيلات لغوية. وفي شكلها الخالص، تتضمن (RBMT) تقنيات مثل الترجمة الآلية المنقولة التي تربط بين اللغات في بعض المستويات النحوية، والترجمة الآلية الوسيطة التي تحاول نمذجة الدلالات. تتطلب حلول (RBMT) إنشاء قواميس لغوية متخصصة للترجمة تنمذج اللغات وترابطاتها معجمياً ونحوياً. وعادة ما تُنشأ هذه الموارد يدوياً أو شبه آلي [177].

وفي شكلها الخالص، تعتمد طريقة (SMT) على المدونات، أي تستفيد من أمثلة للترجمة موجودة فيما يسمى بالمدونات المتوازية (parallel corpora) أو المدونات ثنائية اللغة (bilingual corpora). فيما يلي تمثيل بسيط لما تقوم به أنظمة (SMT)، حيث يُحاذى تلقائياً كلا الطرفين (المصدر والهدف) للكلمة من النصوص المتوازية [178]، انظر الشكل رقم ٨.١. تستخدم محاذاة الكلمة لتعلم نماذج الترجمة التي تربط الكلمات وتسلسل الكلمات في اللغة المصدر بتلك الموجودة في اللغة الهدف [179]. عند ترجمة جملة في لغة المصدر (ويعرف أيضاً بفك التشفير)، يجمع المفكك الإحصائي (statistical decoder) المعلومات في نموذج الترجمة مع نموذج اللغة في اللغة المستهدفة لإنتاج قائمة مرتبة من الجمل المثلى في اللغة الهدف.

في العقدين الأخيرين، غيّر نجاح طرق الترجمة المعتمدة على (SMT) مجال الترجمة الآلية الذي كان يسيطر عليه سابقاً طريقة (RBMT). تجدر الإشارة إلى أن التمييز بين منهجية (SMT) و(RBMT) قد يكون خادعاً لأن القواعد اللغوية الواضحة يمكن أن تكون إحصائية ويمكن تعلمها بشكل تلقائي.



الشكل رقم (٨,١). زوج من الكلمات ومحاذاتها في جمل عربية وإنجليزية. والجملة هي "تحمل المعروضات الفنية الميزات الشخصية ل+ الفنانين".

وقد شهدت السنوات القليلة الماضية اهتماماً متزايداً في التهجين بين الطريقتين لإنشاء أنظمة تستغل مزايا كل من القواعد اللغوية والأساليب الإحصائية المستخدمة. وأنجح هذه المحاولات حتى الآن هي الحلول التي اعتمدت على النهج القائم على المدونات الإحصائية (SMT) واستخدمت إستراتيجياً القيود أو الميزات اللغوية.

٨,٢ مقارنة تعددية اللغات

بما أن الترجمة الآلية في جوهرها عبارة عن وصل لغتين ببعضهما، فإن التحديات التي تواجه الترجمة الآلية تكون مختلفة عندما تشترك اللغات ببعض الخصائص [180] أكثر مما لو كانت مختلفة. فاتجاه الترجمة، وطريقة الترجمة (صوت أو نص)، وتوافر الموارد أحادية وثنائية اللغة هي أيضاً من العوامل الهامة للأخذ في عين الاعتبار. في هذا الجزء، قارنا العربية من حيث التهجئة والنحو والصرف، مع ثلاث لغات أخرى ذات خصائص لغوية مختلفة إلى حد ما وهي: الصينية والإنجليزية والإسبانية[1]. راجع ملخص المقارنة في الشكل رقم (٨,٢).

(١) اللغات الأربع التي نناقشها هنا جميعها تتميز بغنى مواردها وكثافتها اللغوية العالية. من المهم أن نشير إلى أن لهجات اللغة العربية، التي ليست جزءا من هذا الكتاب، من الناحية التقنية فقيرة في مواردها أو اللغات ذات الكثافة المنخفضة (التي ليس لديها الكثير من المتحدثين بها). مسألة ثراء الموارد لن يتم مناقشتها هنا.

٨,٢,١ التهجئة

من ناحية التهجئة، تقع الألفباء العربية المقلصة بالتشكيل الاختياري، واللواصق الشائعة، بين الإسبانية والإنجليزية (لكلا الحروف الهجائية) من جهة، والصينية (كنظام معقد مكون من حوالي عشرة آلاف حرف يسمى كتابة صورية/إشارية logographic) من جهة أخرى. إن تقطيع النص (tokenization) في اللغة العربية هي عملية أسهل بكثير من التجزئة الصينية (Chinese segmentation). لكن كلتا اللغتين تظهران تحديات مماثلة عند ترجمة نص ممسوح بالماسح الضوئي (OCR). بصورة عامة، إن غياب التشكيل يضيف إلى غموض الترجمة من اللغة العربية، لكنه يمثل مشكلة فعلية خاصة عند النقل الكتابي لاسم العلم [21، 45، 50، 22]. والخبر السار هو أنه عند الترجمة إلى اللغة العربية، مقارنة بالترجمة من اللغة العربية، فإن التشكيل الغائب في مخرجات الترجمة قد يجعل بعض أخطاء الترجمة لا صلة لها بالموضوع.

	العربية	الإسبانية	الإنجليزية	الصينية
التهجئة **Orthography**	ألفباء مقلصة– اختيارياً optionally-reduced alphabet	ألفباء alphabet	ألفباء alphabet	أحرف صورية/إشارية logographic characters
الصرف **Morphology**	غنية جداً very rich	غنية rich	فقيرة poor	فقيرة جداً very poor
ترتيب الفعل والفاعل **Subject-Verb order**	V Subj V_{Subj} Subj V	V_{Subj} Subj V	Subj V	Subj V

الشكل رقم (٨,٢). مقارنة بين اللغة العربية والإسبانية والإنجليزية والصينية عبر ستة جوانب لغوية.

هوامش الجدول: V=Verb, Subj=Subject, V_{Subj}=Pro-dropped Verb, N=Noun, Adj=Adjective, Poss=Possessor, Rel=Relative Clause.

Adj 的 N	Adj N	N Adj	N Adj	معدل الصفة **Adjectival Modifier**
Poss 的 N	N of Poss Poss 's N Poss N	N de Poss	N Poss	معدل الملكية **Possessive Modifier**
Rel 的 N	N Rel	N Rel	N Rel	معدل الوصل **Relative Modifier**

تابع الشكل رقم (٨,٢).

٨,٢,٢ الصرف

تعتبر اللغة العربية الأكثر تعقيدًا من الناحية الصرفية مقارنة باللغات الأخرى، تليها الإسبانية والإنجليزية وأخيراً الصينية، التي هي لغة معزولة ولا وجود للصرف فيها. وقد أدى التعقيد الصرفي في اللغة العربية إلى وجود عدد كبير من الأشكال الممكنة للكلمة، وهذا يؤدي إلى مشكلة زيادة التناثر بالإضافة إلى زيادة عالية في المفردات غير المعروفة (OOV). في دراسة قام بها [50]، وجد أن حوالي ٦٠٪ من المفردات غير المعروفة (OOV) في نظام للترجمة الآلية من اللغة العربية إلى الإنجليزية عبارة عن أفعال وأسماء وصفات، وكثير منها عبارة عن أشكال لمتغيرات صرفية جديدة من كلمات نادرًا ما تشاهد.

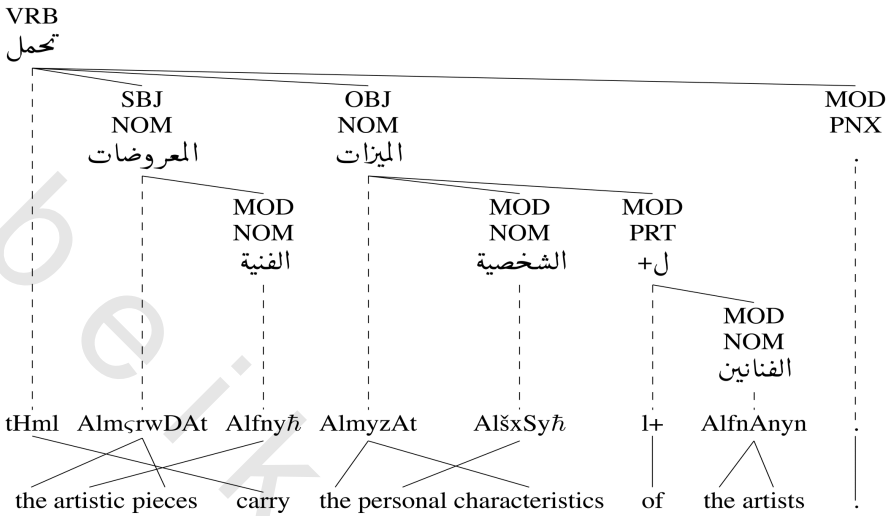
عادة ما يُتعامل مع التعقيد الصرفي للعربية ونتائجها من خلال التقطيع التلقائي لتفريق الكلمات إلى وحدات أصغر بتناثر أقل. وقد تم البحث في مسألة الطريقة المثلى للتقطيع من قبل مختلف الباحثين معظمهم يعمل على الترجمة الآلية من اللغة العربية إلى اللغة الإنجليزية. وقد قام لي (Lee) [108] بالتحقيق في استخدام المحاذاة التلقائية لأقسام الكلام الموسومة بالإنجليزية والعربية المجزأة إلى اللاصقة الجذعية لتحديد التقطيع المناسب للغة العربية. كما أجرى [83، 105] مجموعة كبيرة من التجارب بما في

ذلك استخدام مخططات متعددة للمعالجة القبلية (preprocessing) تعكس مستويات مختلفة من التمثيل الصرفي وتقنيات متعددة لفك اللبس والتقطيع. وهناك نتائج أخرى تم ذكرها تستخدم مخططات محددة للمعالجة القبلية وتقنيات مختلفة، كما في [181، 182، 183، 98]. وقد تبين أيضاً أن التحسينات في محاذاة الكلمة تمت باستخدام مختلف طرق التقطيع الصرفي [107]. من حيث المبدأ، يمكن استخدام مختلف طرق التقطيع المثلى مع مختلف أجزاء نظام الترجمة الآلية مادام أنه يتم التنسيق فيما بينها. على سبيل المثال، يمكن استخدام المواد المعجمية (lemmas) للمحاذاة التلقائية، ولكن يمكن استخدام بعض أشكال الكلمات المصرفة المنزوعة الملحقات (inflected decliticized form) في نموذج الترجمة. في قسم ٥,٣ تم مناقشة مختلف مخططات التقطيع.

تواجه الترجمة من لغات أخرى إلى العربية مشكلة إضافية وهي: أن المخرجات يجب أن تكون بشكل معقد صرفياً حتى لو استُخدمت بعض الأشكال المبسطة في نماذج الترجمة أو القواميس. وقد تجلى بنجاح إعادة دمج أو تركيب (recombination) وتجميع النص (detokenization) للعربية بواسطة الأبحاث التالية [99، 13، 106].

٨,٢,٣ النحو

من المعروف أن اللغة العربية هي لغة معقدة من الناحية الصرف نحوية (morphosyntactically) مع وجود العديد من الاختلافات عن الإسبانية والإنجليزية والصينية. نصف هنا أربع ظواهر نحوية هي: ترتيب الفعل والفاعل (subject-verb order)، والمخصص/المعدل النعتي (adjectival modification)، والتعديل في صيغة الملكية أو ما يطلق عليها مخصص/معدل الإضافة (possessive modification)، والمخصص/المعدل الوصلي (relative modification). يوضح الشكل رقم ٨,٣ بعض هذه الظواهر في سياق اللغة العربية إلى اللغة الإنجليزية.



الشكل رقم (٨,٣). زوج من الكلمات المتحاذية في جملة عربية وإنجليزية. وقد استخدم التمثيل النحوي للعربية الذي أخذ من طريقة نظام كاتب (CATiB) للتحشية، لأغراض التوضيح.

قد يأتي الفاعل في العربية: (أ) مستتراً (الفعل المتصرف)، أو (ب) متقدماً على الفعل (pre-verbal)، أو (ج) متأخراً عن الفعل (postverbal). تأتي كل حالة مع قيودها الصرف- نحوية الخاصة. كما تسمح اللغة الإسبانية باستتار الفاعل في سياقات مماثلة للعربية، ولكن على عكس اللغة العربية فالإسبانية ليس لديها خيار يتعلق بترتيب الفعل والفاعل. عموماً، كل من اللغة الإنجليزية والصينية عبارة عن لغات تحتوي على الفعل والفاعل. وبالنظر إلى الاحتمالات الثلاثة لبيان موضع الفاعل عند الترجمة من اللغة العربية، يتمثل التحدي هنا في تحديد ما إذا كان هناك فاعل واضح، وإذا كان الأمر كذلك، هل هو قبل أو بعد الفعل. وبما أن الفاعل في العربية يتبع الفعل، فإن سلسلة من الفعل والعبارة الاسمية قد تعتبر إما فعلاً وفاعلاً أو فعلاً مستتراً (pro-dropped Verb) ومفعولاً به. وتتفاقم المشكلة مع الفاعل إذا كان طويلاً جداً بحيث

يمكن تقسيمه عن طريق الخطأ إلى عبارات اسمية صغيرة. ويشكل هذا تحدياً لأنظمة (SMT) (مع احتمالية استخدام حجم نافذة محدودة للعبارات) ونظم (RBMT) على حد سواء، التي قد تؤدي إلى أخطاء في التحليل النحوي. انظر الشكل رقم ٨,٤ لمثال يوضح فيها جملة فاعلها مركب من عشر كلمات.

[V اعلن] [NP-SBJ المنسق العام لمشروع السكة الحديد بين دول مجلس التعاون الخليجي] [*SUB* ان ...]

[V Aς1n] [NP-SBJ Almnsq Alς Am lmšrwς Alskh AlHdyd byn dwl mjls Altς Awn Alxlyjy] *[SUB An ...]*
[NP-SBJ The general coordinator of the railroad project among the countries of the Gulf Cooperation Council]
[V announced] *[SUB that ...]*

الشكل رقم (٨,٤). مثال على إعادة ترتيب الفعل والفاعل والمفعول به (VSO) في العربية لنظام اللغة الإنجليزية الذي يعتمد ترتيب الفاعل والفعل والمفعول به (SVO).

إن الترجمة من أي لغة من اللغات إلى اللغة العربية قد تكون أسهل من حيث المبدأ، لأن الحفاظ على ترتيب الكلمات الأصلي في اللغة المصدر مقبول في اللغة العربية. قد يكون هذا صحيحا نحوياً، ولكن سيكون لها بعض العواقب على طلاقة وانسيابية النص في اللغة العربية.

المعدلات/المكملات الفعلية (nominal modifiers) في العربية والإسبانية من جميع الأنواع (النعتية، والملكية، والنسبية) تتبع الاسم الذي يتم تعديله. وكذلك بالنسبة للصينية ولكن بالترتيب العكسي. فالصينية تستخدم أداة الوظيفة (*de* 的) لبيان جميع الأبنية المعدلة. وتستخدم الأسبانية أيضاً الكلمات الدالة مثل حرف الجر (*de*) لبيان تراكيب الملكية والضمائر ذات العلاقة. على سبيل المثال استخدام كلمة (que) لتعديل العلاقة. ومع ذلك فإن اللغة العربية تعتمد بصورة كبيرة على تنسيق دقيق لأل التعريف لتمييز تعديلات الإضافة والنعت. أما بالنسبة لأبنية الصِلَة، فإن

المخصصات الوصلية للنكرة في اللغة العربية تمنع وجود الضمير المتصل، مما يؤدي إلى غموض في بنية الجملة مقارنة بالإنجليزية: *the man wanted (by Mary)/(to go)*. في حين أن ترتيب الفعل والفاعل في اللغة الإنجليزية بسيط، إلا أن ظاهرة تعديل المخصصات الاسمية (nominal modification) في الإنجليزية منتشرة. في بعض الحالات، تكون اللغة الإنجليزية أقرب إلى اللغة العربية أو الإسبانية، وفي حالات أخرى تكون أقرب إلى الصينية. على وجه التحديد، لدى الإنجليزية الكثير من التنوع في بنى الملكية (possessive construction). على سبيل المثال، العبارات التالية في الإنجليزية (*the car keys*) و(*the car's keys*) و(*the keys of the car*) جميعها تترجم إلى (مفاتيح السيارة *mfAtyH AlsyArħ*) في العربية. في المقابل، نجد أن لدى اللغة العربية الكثير من التنوع في ترتيب الفعل والفاعل، ولكن ليس في تعديل ترتيب المكملات الفعلية.

عموماً، هناك الكثير من العمل الذي يجري على النمذجة النحوية للترجمة الآلية، وعلى وجه الخصوص من اللغة العربية للإنجليزية [184، 130، 185، 186، 156، 187] ومن الإنجليزية للعربية [188، 189].

٨,٣ أحدث التطورات في الترجمة الآلية

في السنوات الأخيرة، تلقت الترجمة الآلية للعربية الكثير من الاهتمام. وقد أدى ذلك إلى إحراز تقدم كبير من ناحية الموارد المنشأة والأنظمة المبنية. فهناك العديد من المدونات المتوازية (parallel corpora) الكبيرة جداً والعديد من القواميس للغة الإنجليزية والعربية، وغيرها من اللغات، على سبيل المثال، تملك مدونة الأمم المتحدة وثائق متوازية في اللغة العربية والإنجليزية والصينية والإسبانية والفرنسية والروسية (انظر الملحق ج).

هناك العديد من الحملات التنافسية لتقييم أنظمة الترجمة الآلية ومن ضمنها اللغة العربية باعتبارها واحدة من لغاتها. من أبرزها تقييم NIST MT (للعربية- الإنجليزية)، ومؤخراً افتتح تقييم مدار MEDAR MT (للإنجليزية- العربية). وتركز بعض البرامج الحكومية الممولة في الولايات المتحدة الأمريكية مثل برنامج GALE (لترجمة النص والكلام للنص)، MADCAT (من الماسح الضوئي للنص) وTRANSTAC (لترجمة الكلام على الكلام) بشكل كبير أيضاً على تقييم الترجمة الآلية من العربية للغة الإنجليزية.

تركز غالبية البحوث في الترجمة الآلية للغة العربية على ترجمة اللغة العربية للغة الإنجليزية، إلا أن هناك بعض الجهود المنشورة في الترجمة من اللغة الإنجليزية إلى العربية [99، 188، 13، 106] وترجمة اللغة العربية للفرنسية [190] وحتى من اللغة العربية للصينية [191] وترجمة الدانماركية للعربية [192] وترجمة العبرية للعربية [193]. كما تستخدم العديد من الشركات عدة أنظمة للترجمة الآلية للترجمة من لغة لأخرى، من أبرز هذه الأنظمة هو مترجم قوقل الذي يسمح بترجمة ثنائية الاتجاه لأكثر من خمسين لغة بما فيها اللغة العربية. كما أن هناك أنظمة ترجمة آلية أخرى هامة للاستخدام العام تشمل مترجم مايكروسوفت بنج (Bing) ومترجم صخر (Sakhr's Tarjim).

وأخيراً، على الرغم من أن الغالبية العظمى من الأبحاث المنشورة في مجال الترجمة الآلية في العربية تستخدم طريقة (SMT)، فإننا على علم ببعض البحوث التي نشرت وتستخدم طريقة (RBMT) للترجمة الآلية للعربية: فأبحاث [194، 195] (تستخدم طريقة النقل في الترجمة) وأبحاث مثل [196، 197، 198] (تستخدم طريقة

اللغة الوسيطة في الترجمة). ويوجد اثنتين من الشركات العربية الرائدة في الترجمة الآلية التي تستخدم طريقة (RBMT) أو طريقة هجينة في الترجمة هما شركتا Apptek وصخر.

٨,٤ المزيد من القراءات

على الرغم من أن هذا الفصل ركز على الترجمة الآلية، إلا أن هذا التطبيق مهم وذو صلة بموضوع معالجة اللغة الذي يستحق الحديث عنه. إن استرجاع المعلومات عبر اللغات (Cross language information retrieval - CLIR) هو نوع من طرق استرجاع المعلومات التي تختلف فيها لغة الاستعلام عن لغة النص المبحوث، على سبيل المثال، البحث في نص عربي باستخدام استعلام إنجليزي. ونظراً للكمية المتزايدة من النص الرقمي، تتيح أنظمة CLIR للمستخدم القيام ببعض الفرز لمجموعة مختارة من الوثائق للترجمة الآلية أو الترجمة البشرية. وفي عام ٢٠٠١م، كانت اللغة العربية واحدة من اللغات التي أُخذت في الاعتبار في أحد فروع مؤتمر (TREC) [199] وكذلك في عام ٢٠٠٢م [200]. وقد استخدمت مدونة TREC العربية بمثابة اختبار للعديد من الباحثين [201، 202، 203].