اختبار أداء نظام الترجمة الآلية الاحصائية Moses المكيف لدعم الثنائية اللغوية إنجليزي–عربي*

وفاء بن التركي، نصر الدين سمار

المـعـهـد العالي العربي للترجمة
01 حي الكاستور الفوج الثاني، شارع التبريزي (بيكاردي سابقا) – بئر مراد رايس 16013 الجزائر

obenterki@hotmail.com, n_semmar@hotmail.com

Abstract: *Statistical Machine Translation (SMT) is considered as sub-field of computational linguistics; and the latter is regarded as a branch of Artificial Intelligence (AI) dedicated to Natural Language Processing (NLP). The main purpose of this paper is shortening the distance between the Language and the most recent cutting edge technology dedicated to Machine Translation (MT). On the one hand, Statistical Machine Translation (SMT) considers the translation as a human craft, hence it uses linguistics monolingual and bilingual corpora translated by professional translators; the monolingual are used to train the Language Models (LM) and the bilingual are used to train the Translation Models (TM). On the other hand, it takes advantage of the processing high performance of computers by integrating Statistical Methods to select the best translation. This paper presents the basic concepts and approaches of Machine Translation (MT), and focuses on SMT, then introducing the features of the open source Moses Decoder. This system has been experimented and adapted to translate from English into Arabic. The English-Arabic prototype has been evaluated using MEDAR MT package and the obtained results were very encouraging.*

ملخص: تعد الترجمة الآلية الإحصائية فرعا من فروع الحوسبة اللغوية والتي تصنف بدورها ضمن فروع الذكاء الاصطناعي وقد انصب اهتمامنا في إطار هذا البحث على تقريب المسافة بين اللغة وأحدث طرق المعالجة المكرسة للترجمة الآلية. فمن جهة، تعترف هذه الأخيرة ضمنيا بأن الترجمة مهارة بشرية كونها تستعمل متونا لغوية أحادية وثنائية اللغة مترجمة من قبل مترجمين محترفين. إذ تستعمل المتون الأحادية اللغة لتدريب نماذج اللغة والمتون الثنائية اللغة لتدريب نماذج الترجمة. ومن جهة أخرى، تستغل الترجمة الآلية الاحصائية القدرات الفائقة للحاسوب للمعالجة من خلال ادماج الطرق الاحصائية لاختيار أحسن ترجمة. وقد تناول هذا البحث المفاهيم الأساسية، إضافة إلى مختلف مقاربات الترجمة الآلية مع التركيز على المقاربة الإحصائية التي شكلت المحور الرئيسي للبحث، مرورا بعرض نموذج موزس لفك التشفيرMoses decoder المصمم أصلا للترجمة من اللغة الفرنسية إلى اللغة الإنجليزية والذي قمنا بتكييفه ليتمكن من الترجمة من اللغة الانجليزية إلى اللغة العربية. ثم وصولا إلى تقييم نموذج الترجمة المكيف لدعم الثنائية اللغوية (إنجليزية– عربية). ومن ثم النتائج التطبيقية التي كانت مشجعة ومن ثم انتهى البحث إلى خلاصة تتضمن مجموعة من النتائج والاقتراحات.

Keywords: *machine translation, statistical machine translation, Arabic machine process, Moses Decoder, Giza alignment tool.*

كلمات مفتاحية: الترجمة الآلية، الترجمة الآلية الإحصائية، المعالجة الآلية للغة العربية، أداة فك التشفير موزس Moses decoder، أداة التصفيف جيزة++ Giza++ alignment tool.

** Evaluation of the performance of Moses statistical engine adapted to English-Arabic language combination*

1 مـقدمة

تعريف الترجمة الألية

الترجمة الآلية (MT) Machine Translation[1]، مصطلح معياري يشير إلى تقنية استخدام البرمجيات الحاسوبية (النظم الحاسوبية) لنقل مضمون نص في لغة طبيعية أولى يصطلح على تسميتها " **باللغة الأصل"S L** Source language إلى لغة طبيعية ثانية يصطلح على تسميتها "**بلغة الوصل**" Target language **"TL"**، كما يُصْطَلَحُ على تسمية النص الأصلي الذي يفترض معالجته بواسطة نظام الترجمة "**بالنص المُدخَل**" Input text حيث تتم معالجة النص حاسوبيا ومن ثم إنتاج نـص مترجم يصـطلح على تـسـمـيـتـه "**بالـنـص المُـخْرَجِ**" Output text، وتُجرى عملية الترجمة الآلية إما بمساعدة الإنسان أو من دونها.

2 مقاربات الترجمة الآلية Machine translation approaches[2]

يمكن تصنيفُ نُظُمِ الترجمة الآلية بعدة طرق. أَوَّلًا حسب اللغات التي تدعمها فإذا كانت مثلا تدعم ثنائية لغوية واحدة فإنه يُصطلح على تسميته بِـ **نظام ثنائي اللغة** لأنـه يـوفـر ترجمة من لغة طبيعية معينة (اللغة الأصل) إلى لغة طبيعية ثانية (لغة الوصل). أما إذا كان النظام يدعم أكثر من ثنائية لغات فيصطلح على تسميته بـ **نظام متعدد اللغات**.

كما يمكن تصنيف هذه النظم حسب استراتيجية الترجمة التي تنتهجها لإنتاج نصوص مترجمة ونميز ثلاثة أنواع أساسية للتصاميم العامة المعتمدة لبناء نظم الترجمة الآلية.

1.2 مقاربة الترجمة الآلـيـة المباشرة

تُعْرَفُ أقدم مقاربة للترجمة الآلية في التاريخ بمقاربة " **الترجمة المباشرة** " وقد اعتُمِدَتْ هذه المقاربة في كل نظم "**الجيل الأول**" وبرغم كل التحفظات والمآخذ التي تؤخذ عليها لا يمكن بأي حال من الأحوال إنكار فضلها، كونها أول لبنة في صرح نظم الترجمة الآلية. ولمعالجة الفشل الذريع الذي واجته هذه المنهجية تم تطوير نوعين آخرين من النظم، يعتمد كل منهما على "**الترجمة غير المباشرة**" وهي نظم "**الجيل الثاني**" والتي تنتهج إما **الترجمة التحويلية أو** الترجمة باستعمال **لغة وسيطة.**

1.1.2 نظم الترجمة الآلية المباشرة Direct systems (1950-1970)

يُصَنَّفُ هذا النوع ضمن الجيل الأول لنظم الترجمة الآلية وهي أقدم طريقة للترجمة عرفها تاريخ الترجمة الآلية. وقد صُمِّمَ هذا النظام بكل تفاصيله لثنائية لغات معينة. وتتم الترجمة بصفة مباشرة وتتضمن هذه الآلية أخذ النص اللغة الأصل والذي يكون على شكل سلسلة من الكلمات، ثم تُجَرَّدُ الكلمات من الأشكال الصرفية Morphological Inflections، من أجل الحصول على جذور الكلمات، وبعد ذلك يتم البحث عن جذوع الكلمات Lemmas في معجم ثنائي اللغات ويتعلق الأمر هنا باللغة الأصل ولغة الوصل وعند إيجاد جذر الكلمة المكافئة في لغة الوصل لكل كلمة في اللغة الأصل، يتم تعديل

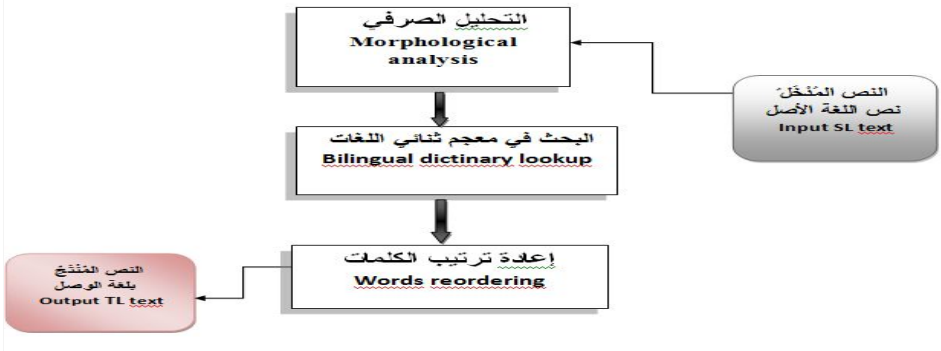
[1] Machine translation: A brief history, By W. John Hutchins

[2] The Oxford hand book of computational linguistics, By Ruslan Mitkov.

ترتيب الكلمات وفق ما تتطلبه اللغة. فمثلا يكون ترتيب الكلمات الجملة الفعلية في اللغة العربية كالتالي:

فعل - فاعل- مفعول به، خلافا للغة الانجليزية والفرنسية اللتان يكون الترتيب فيهما كما يلي: **فاعل- فعل- مفعول به**. وللقيام بوظيفته، يعتمد هذا النوع من النظم على معجم ثنائي اللغات، ويستعمل نفس البرنامج الحاسوبي لتحليل نص اللغة الأصل ولِتَوْلِيدِ to generate نص لغة الوصل، وتجدر الإشارة هنا إلى أن التحليل النحوي والمفرداتي Lexical and syntactic analysis لِنَصِّ اللغة الأصل لا يتم إلاَّ عند الضرورة وذلك عند وجود حالات اللَّبْس(الغموض) Ambiguities، التي تُعيقُ فهم نص اللغة الأصل، ويؤدي بالتالي إلى صعوبة في اختيار العبارات المكافئة والترتيب المناسب للكلمات في لغة الوصل. وينبغي أن نشير هنا إلى أن تحليل نص لغة الوصل موجه تحديدا لإنتاج تمثيلات Representations أُرْصِدَتْ للغة وصل معينة. وبالتالي، لا يتسنى لنا استعمال تلك التمثيلات لتوليد Generate لغة وصل Target language أخرى ، لذلك يسمى هذا النوع من النظم بالنظم الأحادية الاتجاه Systems Unidirectional.

رسم بياني يمثل آلية الترجمة المباشرة Direct Translation process



2.1.2 تقييم نظم الترجمة المباشرة

أرجع الاختصاصيون آنذاك سبب فشل هذه النظم لكونها محدودة على عدة مستويات :

- تفترض هذه النظم أن الجملة لا تعدو كونها سلسلة كلمات متتالية مستقلة عن بعضها البعض بدل أن تكون وحدة متكاملة ومترابطة نحويا صرفيا ودلاليا.
- تتجاهل هذه النظم الروابط النحوية والصرفية بين الكلمات و كذا العلاقات الدلالية التي تجمع الألفاظ مما يؤدي إلى ترجمة خاطئة أو ضعيفة لغويا.
- تتميز الترجمة بكونها ترجمة حرفية لا تعبأ بالسياق رغم استعمال موارد لغوية ثرية كالمعاجم ثنائية اللغة
- تتميز هذه النظم بكونها أحادية الاتجاه حيث أنها مصممة لثنائية لغوية واحدة إضافة إلى كون كل الوحدات النمطية مدمجة في وحدة واحدة مما يضفي عليها نوعا من الجمود

نظرا لصعوبة ادخال تغييرات على جزء من هذه البنية دون التأثير على الأجزاء الأخرى من المنظومة .

2.2 مقاربات الترجمة الآلية غير المباشرة (نظم الجيل الثاني)

يمكن تصنيف عدة أنواع من مقاربات الترجمة الآلية غير المباشرة، نوع يعتمد على **القواعد النحوية** ونميز صنفين نظم تعتمد **مقاربة تحويلية** وأخرى تعتمد **مقاربة محورية (لغة وسيطة**[3]) (Trujillo, 1999)، ونوع يتبنى **مقاربة تجريبية Empirical approach**، ونميز صنفين : المقاربة **الاحصائية** (2010 ,Kohen[4]) والمعتمدة أساسا على نماذج **IBM**، والمقاربة **المعتمدة على الأمثل** (Wu, 2007) ، وهناك نوع ثالث يتمثل في ما يسمى بالنظم **الهجينة Hybrid** المركبة من مقاربتين على الأقل ومثال ذلك النظم التي تتجه نحو دمج الحسابات الاحصائية مع مختلف مصادر المعلومات اللغوية (Shwenk et al., 2007) .

1.2.2 نظم الترجمة الآلية المعتمدة على القواعد Rule-based Machine Translation

تتبنى النظم - المعتمدة على القواعد- المعالجة اللغوية كركن أساسي في معالجتها للغات الطبيعية، وهي تُمَثِّلُ أول نظم " الجيل الثاني" لنظم الترجمة الآلية. ونميز ضمن هذا النوع من النظم مقاربتين **المقاربة التحويلية** و**مقاربة اللغة الاصطناعية (اللغة الوسيطة)**، وقد تم تصنيفها كنظم معتمدة على القواعد لتمييزها عن النظم الحديثة التي تنتمي إلى الجيل الثالث مثل النظم المعتمدة على الذخائر النصية والتي يصطلح على تسميتها بالنظم التجريبية.

- **مراحل الترجمة**

يمر النص "المُدْخَلْ" عبر ثلاث مراحل :

أ- **مرحلة التحليل:** وهي المرحلة التي يتم خلالها تقدير التمثيل النحوي لجملة لغة الوصل.

ب- **مرحلة التحويل (النقل):** وهي المرحلة التي يتم خلالها تحويل التمثيل النحوي لتمثيل مكافئ في لغة الوصل.

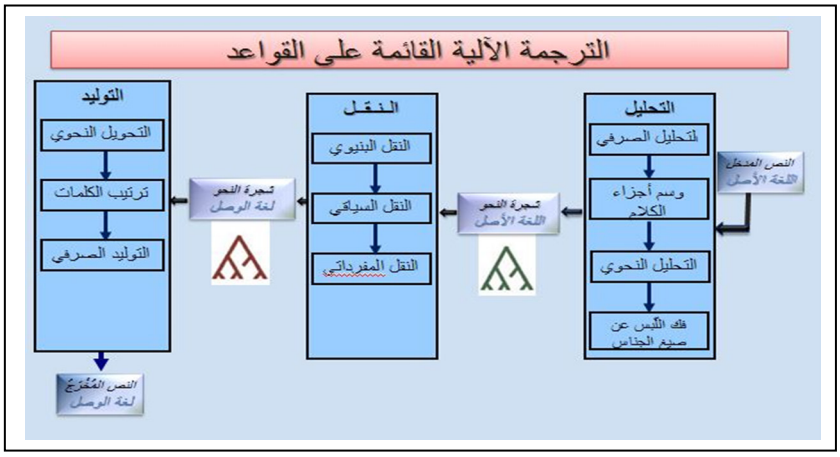
مرحلة التوليد : وهي المرحلة التي يتم فيها توليد جملة لغة الوصل انطلاقا من البنية النحوية الناتجة عن مرحلة التحويل.

1.1.2.2 مقاربة النظم التحويلية Transfer systems approach

تعتمد **مقاربة الترجمة الآلية التحويلية** على غرار مقاربة **اللغة الاصطناعية** لغة وسيطة للانتقال من اللغة الأصل إلى لغة الوصل . إذ تتم الترجمة في هذا المنهج على ثلاث مراحل، **التحليل** ثم **التحويل** وأخيرا **التوليد**، مرورا بتمثيلين وسيطيين للغتي الأصل والوصل.

[3] Translation Engines:Techniques for Machine Translation, By Trujillo, A. (1999).

[4] Statistical Machine Translation, By Koehn, P. (2010).



مراحـل الترجمة

أ- **مرحلة التحليل:** يتم تحليل النص الأصلي(المُدْخَلُ) ونقله إلى صيغة تحويلية مجردة من الكثير من الخصائص النحوية المميزة للنص الأصلي، علما بأن حل التباسات الازدواج المفرداتي أو الدلالي إن وُجِدَا، يتم قبل الانتقال إلى الصيغة المجردة، وهي عملية مستقـلة عن لغة أو لغات الوصل. والهدف من هذه البنية المجردة هو تسهيل الترجمة عن طريق صيغة نموذجية للجملة يمكن تكييفها لمختـلف البُنَى الممكنة للجملة.

ب- **مرحلة التحويل:** بعد هذا التحليل يتم تحويل الصيغة الجديدة للنص "المُدْخَلِ" إلى صيغة مكافئة في لغة الوصل. وهناك عدة أشكال للتحويل وتتوقف هذه الأشكال على مستوى التحويل الذي تُخضع إليه جملة اللغة الأصل، فكلما كانت الصيغة المحولة أكثر تجريدا كلما سَهُلَ تصميم وحدة نمطية للتحويل ملائمة للصيغة التحويلية.

ت- **مرحلة التوليد:** يتم في هذه المرحلة استخدام التمثيل الوسيط الناتج عن المرحلة السابقة لتوليد النص النهائي بلغة الوصل.

2.2.2 مقاربة اللغة الاصطناعية (لغة وسيطة) Interlingua approach

تنتمي هذه المقاربة على غرار مقاربة الترجمة الآلية القائمة علـى الطرق التحويلية إلـى ما اصْطُلِحَ على تسميته "بالجيل الثانـي" لنظم الترجمة الآلية، ويلائم هذا النوع من المقاربات النُّظُمَ المتعددة اللغات[5] بصفة خاصة.

تصنف "**كثاني أهم المقاربات**" من حيث التصميم[6]، ويَفْتَرِضُ هذا المنهج إمكانية إعادة صياغة نص اللغة المصدر على شكل تمثيل دلالي نحوي مُجَرَّدٍ معروف في عدة لغات، لكن ذلك لا يضفي عليه سمة اللغة العالمية. وانطلاقا من هذا التمثيل الوسيط (اللغة الاصطناعية) يمكن توليد نصوص بلغات متعددة.

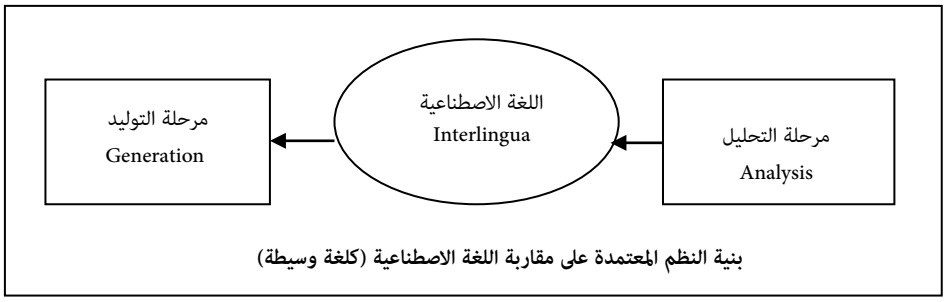
[5] Multiligual systems (basic strategies) : Machine translation : a brief history . By W.John Hutchins

[6] المصدر: Computational linguistics : Machine translation over view, By John Hutchins

وخلافا للترجمة التحويلية تتم الترجمة على مرحلتين: من نص اللغة الأصل إلى اللغة الاصطناعية ومن اللغة الاصطناعية إلى لغة الوصل.

مراحل الترجمة

أ- **مرحلة التحليل:** هي أول مرحلة في هذه العملية، ويتم فيها تحليل جملة اللغة الأصل، ثم استخلاص المعنى الدلالي، بعدها يصاغ المعنى على شكل تمثيل مجرد abstract representation موافق له.



بنية النظم المعتمدة على مقاربة اللغة الاصطناعية (كلغة وسيطة)

وتجدر الإشارة إلى أن آليات التحليل مكرسة للغة أصل (مصدر) معينة، وهي مستقلة عن لغة الوصل.

ب- **مرحلة التوليد:** وهي ثاني و أخر مراحل هذه المقاربة ويتم خلالها توليد جملة لغة الوصل انطلاقا من التمثيل المجرد الوسيط (اللغة الاصطناعية) وتُسْتَخْدَمُ وحدة نمطية مستقلة تماما عن اللغة المصدر.

— **الفرق بين الترجمة بمقاربة تحويلية والترجمة بمقاربة اللغة الاصطناعية :**

يكمن الفرق بين المقاربتين في كون **المقاربة التحويلية** تحتاج إلى تمثيلين مجردين، وتتم عبر المراحل التالية: مرحلة التحليل، تمثيل مجرد للغة الأصل، مرحلة النقل، تمثيل مجرد للغة الوصل، مرحلة التوليد.

أما بالنسبة لمقاربة اللغة الاصطناعية فهي تتم على ثلاث مراحل فقط : التحليل، تمثيل وسيط بلغة اصطناعية (مثل لغة الاسبرنتو) ثم مرحلة التوليد.

3.2.- نظم الترجمة الآلية المعتمدة على مقاربة تجريبية

3.2 .1. نظم الترجمة القائمة على المقاربة الاحصائية[7]

Machine translation systems based on a Statistical approach[8]

بناء على ما سبق، وإذا حاولنا استخلاص قاسم مشترك بين مناهج المقاربات المذكورة آنفا، سنجد أن جلها تعتمد على نص "**مُدْخَلِ**" في **اللغة الأصل** وبصرف النظر عن مراحل المعالجة التي يخضع لها سطحية كانت أو عميقة والمرتبطة أساسا بنوع المقاربة، فإننا سنحصل في النهاية على نص "**مُخْرَجٍ**" في **لغة الوصل.**

[7] Statistical MachineTranslation, By Philip Koehn.

[8] A statidtical Approach to Machine Translation, By peter F .Brown, and Al.

في المقابل، ثمة طريقة أخرى في التعاطي مع مسألة الترجمة، حيث أن المقاربة التي نحن بصدد تعريفها، تتناول الترجمة من زاوية مغايرة لسابقاتها، لاسيما وأنها تركز على النتيجة بدل المعالجة. وفي هذا السياق، سنتناول رأي بعض فلاسفة اللغة الذين يرون أنه لا يمكن لجملة بلغة معينة أن تكون ترجمة لجملة أخرى بلغة ثانية بأي حال من الأحوال، ولا شك أن هذا الرأي متشدد إلى حد ما، بالرغم من أنهم قد يُرْجِعُونَ ذلك لِكَوْنِ الترجمة حسب رأيهم غير وفية للنص الأصلي، لأنها قد تعتمد في صياغتها على بنية مختلفة عن جملة اللغة الأصل وحتى السياق قد يختلف، ومثال ذلك العبارة التالية " **رهن الاشارة** " باللغة العربية تقابلها عبارة "At one's back and call" باللغة الانجليزية، أو عبارة " **سبق السيف العدل**" تقابلها عبارة "Lock the barn door after the horse is stolen". إذا تأملنا بنية الثنائيتين فإننا نجد بنيتين مختلفتين في سياقين مختلفين تماما، وفي المقابل نجد أن كلاهما تؤديان المعنى لكن في سياقين متوافقين مع ثقافة كل منهما.

في هذه الحالة كان التركيز على المعنى على حساب بنية الجملة، وباستعمال ألفاظ مغايرة تماما لتلك التي وردت في نص اللغة الأصل. ومن ناحية أخرى، نجد أنه في حالة استخدام ألفاظ مكافئة لتلك الواردة في النص الأصلي مع مراعاة احترام البنية التركيبية للنص الأصلي، فإننا دون شك، سنحصل على جملة مكافئة من حيث الألفاظ ومن حيث البنية إلا أنها ستكون بعيدة كل البعد عن كونها ترجمة لجملة اللغة الأصل SL، لأنها جملة لا معنى لها، وإن وُجِدَ المعنى ليس بديهيا أن يؤدي نفس معنى جملة اللغة الأصل. وبالتالي، يجب أن نعي أن مصادفة هذا النوع من المشاكل شيء وارد ليس فقط ضمن مفاهيم معينة في ثقافات محددة، وإنما يكفي أن تُستعمل استعارات Methaphors أو تراكيب أو كلمة أو زمن لا مقابل له في لغة الوصل ليحدث هذا اللَّبس Ambiguity.

فيما يلي سنتناول مفهوم البرمجة الخطية لحل معضلة الترجمة الاحصائية ولهذا يجدر التذكير بالمفهوم العام للبرمجة الخطية قبل تناول كيفية تطبيقها على نماذج اللغة والترجمة:

- **تعريف البرمجة الخطية**

البرمجة الخطية Linear programming أسلوب أساسي ومهم يساعد متخذي القرار على اتخاذ قرارات صحيحة وبطريقة علمية. وتعد مسائل البرمجة الخطية جزءاً من مسائل البرمجة الرياضية التي تشمل الخطية منها واللاخطية[9].إن البرمجة الرياضية الخطية هي مسألة تفضيل، ويُقصَد هنا بمسائل التفضيل تلك المسائل الرياضية التي تبحث عن تعظيم Maximize أو تقليل Minimize لدالَّة (تابع) خطية مرتبطة بمقيدات رياضية خطية أيضاً.

Subject to : Ax=b

x>=0

حيث أن المتغير X هو شعاع المتغيرات التي نريد حلها، علما بأن A هي مصفوفة تتضمن معاملات coefficients

معروفة، وتسمى cx "الدالة الهدف" object function، أما Ax=b فتسمى "القيود" constraints، علما بأنه يفترض بأن تكون كل هذه المتغيرات ذات أبعاد متناسقة، بمعنى أن هناك علاقة خطية تجمع بينها.

[9] Linear programming : http://ar.wikipedia.org/wiki/%D8%A8%D8%B1%D9%85%D8%AC%D8%A9_%D8%AE%D8%B7%D9%8A%D8%A9

- **إسقاط مفهوم البرمجة الخطية على الترجمة الآلية الإحصائية**

لا غضاضة من الاعتراف أنه من الصعب وأحيانا مستحيل توفير ترجمة أمينة ووفية وبلغة طبيعية صحيحة (بلغة الوصل) في الوقت ذاته، لأن جمع هذه المقومات معا يمثل حلم كل مترجم. وغالبا وبرغم اجتهاد المترجمين يتحتم عليهم التنازل عن فكرة تحقيق كل الشروط مجتمعة للحصول على ترجمة ملائمة ويكتفون بتحقيق بعضها. ومن هنا، يمكن أن نُكَوِّنَ فكرة أولية عن الطريقة التي يجب أن تنتهجها نظم الترجمة الآلية لإتمام عملية الترجمة. بداية، يتم تحديد **الهدف** الذي تصبو إليه نظم الترجمة الآلية وهو الحصول على نص "مُخْرجٍ" بأفضل جودة ممكنة، وفي هذا الإطار وإذا استعملنا البرمجة الخطية، حيث ان الدالة التي تمثل "الهدف" و التي تبحث عن الحل الأفضل (الترجمة الأفضل بالنسبة لنا) وكانت متغيراتها (x,y)، واللذان يمثلان على الترتيب (الوفاء، الفصاحة) فالدالة "الهدف" يجب أن تحقق أقصى قيم لِـ (x) و(y) للحصول على أحسن ترجمة. ومن أجل ذلك، صُمِّمت النظم القائمة على المعلومات الاحصائية، حيث أنها تحقق الفكرة الموضحة آنفا، عن طريق بناء نماذج محتملة للمتغيرات :

(x): الوفاء Faithfulness،

(y): الفصاحة Fluency) ثم يتم دمج هذه النماذج لاختيار الترجمة الأكثر احتمالا. إذا اعتبرنا الجداء (**x*y**) مؤشرا للجودة حيث: (x): متغير يمثل قيمة الوفاء، (y): متغير يمثل قيمة الفصاحة.

يمكن بناء نموذج لترجمة جملة في **اللغة الأصل** SL إلى جملة في **لغة الوصل** TL، وتكون الصيغة كالتالي:

$$\textbf{أحسن ترجمة A} = Arg_x max_{\ TL}\ \mathbf{P}(SL|TL)\ .\ \mathbf{P}(TL)$$

علما بأن :

$$Arg_x max\, f(x)\ \in \{x|\ \forall\, y: f(y) \leq\ f(x)\}$$

$Arg_x max$ هو وسيط أعظمي[10]

P(SL|TL) : هو احتمال الترجمة ويجسد أيضا ميزة الوفاء faithfulness للنص الأصلي

(TL) **: هو** احتمال اللغة ويجسد أيضا ميزة **الفصاحة** في النص المُتَرْجَمْ.

وللتذكير : يعرف **الوسيط الأعظمي** (بالإنجليزية $Arg_x max$) في الرياضيات بأنه مجموعة قيم x التي من أجلها تأخذ الدالة f(x) أعظم قيمة لها، فمثلا إذا اعتبرنا x تنتمي إلى مجموعة الأعداد الصحيحة وكانت الدالة f(x) معرفة كما يلي:

$$f(x)=1-|x|،$$

فإن أعظم قيمة لِـ الدالة f(x) ستكون عند القيمة x =0

وعليه فإن $Arg_x max\ f(x)= 0$

1.1.3.2 نموذج نظام ترجمة الآلية بمقاربة إحصائية

تعتمد المقاربة الاحصائية على الطرق الاحصائية في معالجتها وبالتالي سنستعمل بعض الصيغ الرياضية، ولتبسيط العبارات الرياضية سنستعمل بعض الرموز عوض التسميات التالية:

[10] http://en.wikipedia.org/wiki/Arg_max

- اللغة الأصل : (SL) وهي اللغة الانجليزية.
- لغة الوصل : (TL) وهي اللغة الفرنسية بالنسبة لنماذج IBM
- جملة اللغة الأصل (اللغة المصدرsource language): (e)
- جملة لغة الوصل (اللغة الهدف target language): (f)
- كلمة من اللغة الأصل: (e_i)
- كلمة من لغة الوصل : (f_i)

بهدف ترجمة e (جملة اللغة الأصل) :

1. نفترض أن كل الجمل في لغة الوصل(f) هي احتمالات ترجمة لجمل اللغة الأصل.
2. يتم اختيار الترجمة المرفقة بأعلى احتمال.

$$\mathbf{Arg_e max\ P(e|f) = arg_e max} \frac{P(f|e) * P(e)}{P(f)}$$

$Arg_e max\ P(e|f)$: أحسن ترجمة أو الترجمة الأكثر احتمالا في لغة الوصل **الوسيط الأعظمي لجملة لغة الوصل ($Arg_e max$)** يمثل أعلى قيمة من قيم احتمالات ترجمة لغة الوصل P(e|f) : الاحتمال المشروط لورود جملة اللغة الأصل **(e)** شريطة معرفة (f)، بمعنى احتمال ورود (f) كترجمة لِـ **(e).**

التعريف الرياضي للاحتمال المشروط

يسمى **P(e|f)** الاحتمال المشروط، ويتمثل في احتمال تقاطع حدثين وبالتالي يساوي جداء احتمالي الحدثين علماً بأن الأول (الحدث الأول) قد وقع.

P(e|f): احتمال مشروط[11] للحدث e بمعرفة الحدث f.

وإذا طبقنا نظرية بايز The Bayes' theorem نحصل على الصيغة التالية:

$$\textbf{(1)} \leftarrow \quad \textbf{Best T= Argmax}_{e} \quad \frac{P(e|f) * P(e)}{P(f)}$$

علما أن : « Best T » ترمز لأحسن ترجمة وهي مرفقة بأعلى احتمال

الوسيط الأعظمي $Argmax_e$ يمثل أعلى قيمة تحققها الصيغة (1)

علما بأن:**P(e)**: هو نموذج اللغة بالنسبة لنظام الترجمة بمقاربة إحصائية وهو أيضا احتمال الحدث **e** حسب مبرهنة بايز.

P(f|e): هو نموذج الترجمة بالنسبة لنظام الترجمة بمقاربة إحصائية وهو أيضا احتمال مشروط للحدث **f** بمعرفة الحدث **e** : احتمال وقوع الحدث **e**(الجملة الأصل).

نحن نبحث عن أكبر قيمة لاحتمال الجملة الأصل من بين كل احتمالات الترجمة المتوفرة **لجملة اللغة الأصل المحددة** وعليه فإن القاسم Denominator p(f)(احتمال جملة للغة الوصل) قيمة ثابتة أيضا

[11] الاحتمال المشروط : http://www.arab-ency.com/index.php?module=pnEncyclopedia&func=display_term&id=69

(لأن لدينا جملة لغة وصل محددة) وبناء عليه يمكننا تجاهل هذه القيمة الثابتة و بالتالي اختصار العبارة السابقة لتصبح كما يلي:

$$Best\ T = Argmax_e P(e|f) * P(e) \quad \textbf{(2)} \leftarrow$$

$P(e|f)$: نموذج الترجمة.

$P(e)$: نموذج اللغة.

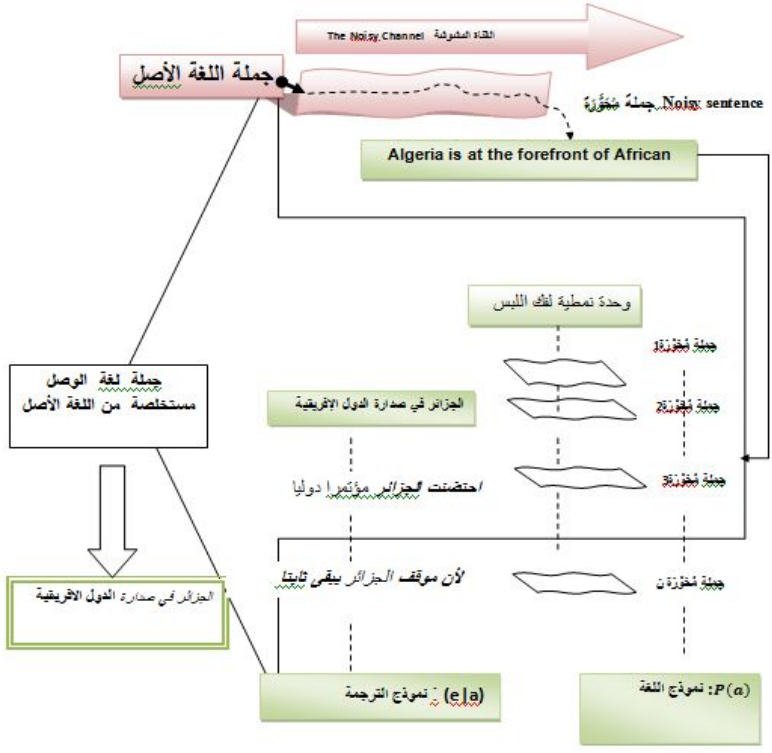
2.1.3.2 القناة المشوشة Noisy Channel

تسمى المعادلة (2): القناة المشوشة Noisy channel وتعتمد على جزئين:

← P(f|e) **نموذج الترجمة.**

← **P(e) نموذج اللغة.**

يتعين علينا في هذا السياق، أن ندرك بأن إسقاط نموذج القناة المشوشة على نماذج الترجمة الآلية بمقاربة إحصائية يتطلب منا تفكيرا بأثر رجعي يمكننا من خلاله التعرف على أصل الأشياء، وفي هذه الحالة سنفترض أن جملة اللغة الأصل **e** لا تعدو كونها نسخة مُحَوَّرة (مشوهة) مستخلصة من مجموعة نماذج من الترجمات (جمل لغة الوصل) وعليه فالمهمة المنوطة بنا هي التعرف على جملة لغة الوصل (الترجمة) المجهولة التي أنتجت الجملة f.

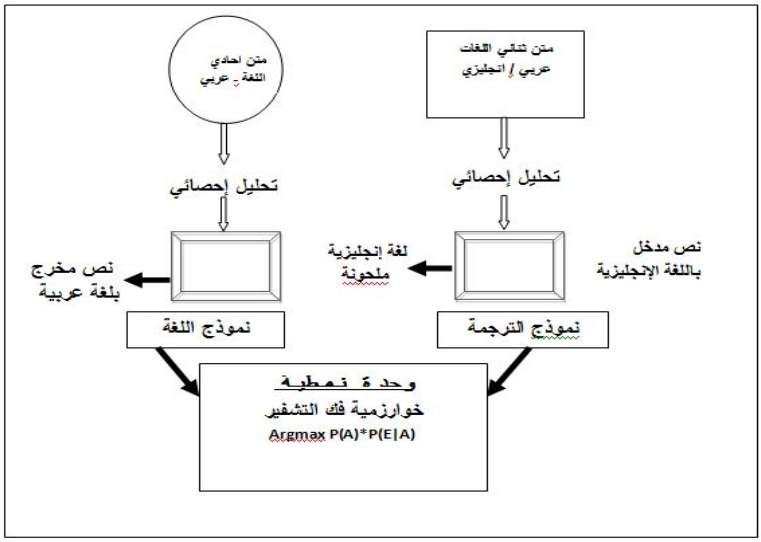


رسم بياني يمثل نموذج القناة المشوشة

تـوضيح

يمثل الشكل البياني آلية عمل نموذج القناة المشوشة Noisy channel لنظام ترجمة آلية بمقاربة إحصائية. وبغرض ترجمة جملة باللغة الانجليزية وهي **اللغة الأصل** إلى اللغة العربية وهي **لغة الوصل**، سنتعامل مع الجمل المصدر والجمل الهدف على حد سواء، بانتهاج طريقة البحث بأثر رجعي لتحديد الجمل المولدة لكل منها. واعتمادا على هذا المنهج يتم بناء نموذج لآلية التوليد من جملة باللغة العربية إلى جملة باللغة الانجليزية مرورا بقناة مشوشة علما بأن لغة الوصل هي اللغة العربية، إلا أننا سنذهب في الاتجاه المعاكس:

لدينا جملة في **اللغة الانجليزية** (اللغة الأصل) ومهمتنا ترجمتها إلى **اللغة العربية (لغة الوصل) a** سنفترض جدلا أن الجملة **e** ليست **إلا مُخْرَجَةً Output** لجملة عربية **a** تمر عبر القناة المشوشة وسنتحرى الدقة في البحث عن أفضل جملة عربية **a** تكون جملة مصدرا للجملة الانجليزية **e**.



الشكل البياني الموضح أدناه يبين آلية عمل القناة المشوشة ضمن نظام الترجمة الإحصائية

3.1.3.2 المعضلات الثلاثة التي تواجه نظم الترجمة الآلية الاحصائية

تقوم هذه النظم على فكرة استعمال الترجمات البشرية المتوفرة لترجمة نصوص جديدة. وقد أفضت هذه الفكرة
إلى مقاربتين :الترجمة الإحصائية أو القائمة على الاحتمالات والترجمة القائمة على الأمثلة.
إذ تستعمل الترجمة الإحصائية متنا لغويا ثنائي اللغات لتدريب منظومة الترجمة بهدف استخلاص نموذج الترجمة، كما تستعمل متنا أحادي اللغة لتدريب منظومة الترجمة بهدف استخلاص نموذج اللغة. بعد التدريب على النموذجين السابقين، تعتمد منظومة الترجمة الاحصائية على المعرفة المكتسبة من خلال التدريب لحساب احتمال أن تكون جملة معينة بلغة الوصل ترجمة لجملة اللغة الأصل.و توجد عدة مناهج لإتمام عملية النمذجة : هناك مناهج قائمة على **الكلمات**، ومناهج قائمة على **مقاطع من الكلمات** وأخيرا مناهج قائمة على **تركيب الجملة**.

وتكون هذه النظم أكثر فعالية عندما تكون الثنائية اللغوية المعالجة متقاربة على مستوى البنية الصرفية على غرار اللغتين الفرنسية والإنجليزية، وعند استعمال متون لغوية ثرية وذات حجم كبير للتدريب.
وتجدر الإشارة إلى كون نقطة القوة التي تميز نظم الترجمة الإحصائية والتي تتمثل في قدرتها على معالجة أي ثنائية لغوية، هي التي قد تحد من فعاليتها وتؤدي إلى ترجمات خاطئة وملحونة لغويا.

- ورغم أدائها الجيد، يبقى مستوى الترجمة في نظم الترجمة الإحصائية مرتبط بِتَوَفُّرِ متون ذات حجم كبير، الشيء الذي يصعب توفيره بالنسبة لكل الثنائيات اللغوية وفي كل الاختصاصات، إضافة إلى كونها عملية جد مكلفة، وهذا ما يحد من فعالية هذه النظم.
- **P(A):** احتمال وقوع الحدث A
- **P(B):** احتمال وقوع الحدث B
- **P(A|B)**: يسمى الاحتمال الشرطي للمتغير العشوائي (الحدث) A، ويتمثل في احتمال وقوع الحدث A شريطة معرفة
  الحدث B .

أ- نموذج اللغة: لدينا سلسلة هجائية متتالية باللغة العربية(جملة) ولتكن (a) وعن طريق وحدة نمطية يتم اسناد قيمة P(a) (احتمال الجملة العربية) لكل جملة مقترحة:

- إذا كان الجملة a (سلسلة هجائية متتالية) تتسم بلغة عربية صحيحة فذلك يستلزم أن P(a) ⟸ احتمال كبير.
- إذا كانت الجملة a (سلسلة هجائية متتالية) تتسم بلغة ركيكة (ملحونة) فذلك يستلزم أن P(a) ⟸ احتمال ضعيف.

ب- نموذج الترجمة : لدينا ثنائية من السلاسل الهجائية (a،e) وعن طريق وحدة نمطية يتم اسناد قيمة P(e|a) وهو احتمال الحدث **e** (الجملة الإنجليزية **e** بالنسبة لنا) والمشترط بمعرفة الحدث **a** (الجملة العربية **a**).
-1 في حالة ما إذا كان الزوج (a،e) يمثل كُلُّ ترجمة للآخر (كل جملة هي ترجمة للأخرى) فإن ذلك يستلزم أن P(e|a) ⟸ احتمال كبير.

4.1.3.2 وحـدة فك التشفير: خوارزمية فك التشفير

لدينا **نموذج للغة** و**نموذج للترجمة** و**جملة جديدة(جملة مدخلة) ج،(e)** والمطلوب هو إيجاد الجملة العربية **(a)** التي من أجلها تـأخذ الصيغة التالية أعظم قيمة ممكنة لها: **P(a).P(e|a)**

وتُقْرَأُ: احتمال الجملة العربية جداء احتمال الجملة الانجليزية المشروط بمعرفة الجملة العربية .

2.3.2 الترجمة المعتمدة على الأمثلة Example-based machine translation

تم استحداثها سنة 1984 من طرف ناجاو ماكوتو NAGAO Makoto[12]، وتعتمد نظريته أساسا على مبدأ التماثل. وتعتبر أن اللغات تملك انتظـــامـــا كافيـــا يـــتـــيـــح استنتاج الترجمات انطلاقا من أمثلة الترجمة وبالتالي تعتمد هـــذه المـــقـــاربة عـــلى التناظر بين أزواج من الترجمات المصفوفة.

[12] : Example-based Machine Translation : http://en.wikipedia.org/wiki/Example-based_machine_translation

وتفترض هذه النظم أن عملية الترجمة تخضع لنفس المنطق. وقد أعلن ناجاو NAGAO عن الفكرتين التاليتين معتبرا إياهما مبدأين توجيهيين لمقاربة الترجمة المعتمدة على الأمثلة:

- لا يقوم المترجم البشري بترجمة جملة عن طريق تحليل لغوي عميق
- يقوم المترجم البشري بتقسيم الجملة إلى عدة أجزاء ثم ترجمة كل جزء منها وفي النهاية تجمع الترجمات الجزئية لتشكل جملة لغة الوصل. وتُنْجَزُ كل ترجمة جزئية باستعمال مبدأ التماثل بإجراء مقارنة واعية أو لاشعورية مع ترجمات سابقة الانجاز. وفي هذا الإطار، تهدف الترجمة المعتمدة على الأمثلة أساسا إلى استعمال التماثل الذي تتسم به اللغـة لترجمة النص. وقد أثمرت هذه المقاربة نتائج جيدة خـاصة لثنائية لغات مثل الانجليزية واليابانية، بالإضافة إلى ذلك كان لها دورا في تطوير أداة أحدثت ثورة في مجال الترجمة بمساعدة الحاسوب وتتمثل في ذاكرة الترجمة.

4.2 نظم الترجمة الآلية الهجينة Hybrid Machine Translation

لم تستطع النظم المعتمدة على الأمثلة ولا تلك التي تتبنى مقاربة إحصائية أن تثبت أن نتائجها أحسن من تلك التي تنتجها نظم الترجمة المعتمدة على القواعد إلا أن كل منها اثبت نجوعه في مجال محدد. وسرعان ما برزت عدة نظم هجينة نتيجة لذلك، فمثلا هناك بعض المشاكل لا تجد حلا لها إلا من خلال نظـم الترجمة الآلية المعتمدة على الأمثلة، وبسبب نجاعة هذه الأخيرة في هذه الحالات، تلجأ النظم الهجينة التي المزاوجة بين النظم القائمة على الأمثلة ونظم أخرى إلى تفعيل وحدة نمطية مكرسة للنظام القائم على الأمثلة، للتعرف تلقائيا على هذه الحالات وبالتالي معالجتها.

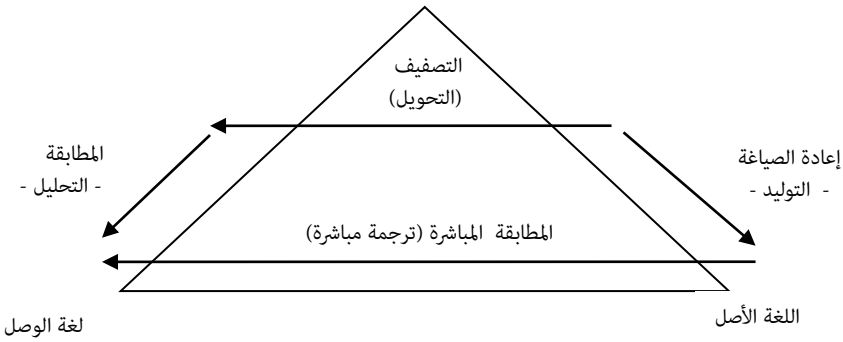
وتعجز النظم القائمة على القواعد عن تحديد هذه الحالات، نظرا لطبيعة ومراحل المعالجة المنوطة بها. ومثال ذلك العبارات الاص idiomatic expressions التي تشكل تحديا بالنسبة للنظم القائمة على القواعد. فيما يزاوج نوع أخر من نظم الترجمة الآلية القائمة على مقاربة هجينة بين مرحلتي التحليل والتوليد لنظم معتمدة على القواعد ومرحلة التحليل لنظام معتمد على الأمثلة، بينما تعتمد **النظم الهجينة** المكرسة لمعالجة الخطاب والكلام المنطوق في مرحلة التحليل على التحليل الإحصائي بصفة خاصة، بينما تُنَفَّذُ مرحلتي التحويل والتوليد باستعمال النظام القائم على القواعد.

هناك نوع آخر من النظم الهجينة وهي **النظم الهجينة ذات المحركات المتعددة** يخضع نص اللغة الأصل إلى مختلف نظم التوجيه الآلية التي تُكَوِّنُ مُجْتَمِعَةً **النظام الهجين ذو المحركات المتعددة Multi-engines hybrid system.**

ومهما اختلفت التقنيات المُعْتَمَدَةُ ضمن هذه النظم، والتي تراوح الخطى بين نظم قائمة على المعاجم وأخرى معتمدة على القواعد في التحليل والتوليد، ونوع آخر يعتمد على الأمثلة أو طرق إحصائية بحتة، إلا أنها في النهاية تشترك في طريقة التقييم، حيث يقوم كل منها بإسناد درجات للنص المُخْرَجِ، وذلك بواسطة **وحدات نمطية Modules** مدمجة في كل محركات نظم الترجمة المتوفرة ضمن منظومة الترجمة الهجينة، وبفضل خوارزمية تتولى مهمة إنجاز التقييم وإسناد درجات تحدد مدى وثاقة النص المُخْرَجِ. علما أن لكل نظام معايير خاصة للتقييم، تلائم طريقة معالجته للنص وتسمح بإسناد درجات على هذا الأساس.

في نهاية هذه المعالجة يتم عرض النصوص المخرجة على وحدة نمطية تمثل مرحلة وسيطة، تتم خلالها مقارنة مختلف النصوص المخرجة (مختلف الترجمات الناتجة عن النظم المدمجة في هذا النظام الهجين)

ثم يُختار النص المُخْرَجُ المرفق بأعلى درجة تلقائيا في حالة إنفراده بهذه الدرجة، أو تُعتمد ترجمة أدرجت من قِبَلِ كل نظم الترجمة، أو قد يلجأ النظام إلى إثراء هذه الترجمة عن طريق تركيب أحسن الأجزاء من أفضل الترجمات المقترحة.



- **مراحل الترجمة**

1- المطابقة : البحث عن مقاطع من نص اللغة الأصل في المتن المرجعي.
2- **التصفيف** : تحديد الترجمات المقابلة لهذه المقاطع.
3- **إعادة الصياغة** : صياغة مقاطع الترجمة للحصول على النص الملائم بلغة الوصل.

3 الترجمة الآلية الاحصائية الداعمة للثنائية اللغوية إنجليزية - عربية باستخدام الأداة مـوزس لفك التشفير Moses decoder[13]

1.3 أداة فك التشفير موزس Moses decoder في نظم الترجمة الإحصائية

اعتمدت النظم الأولى للترجمة الآلية بمقاربة إحصائية على مناهج قائمة على الكلمة كوحدة للترجمة، إلا أن جودة النصوص المخرجة لم تكن بالمستوى المرجو، ويعود سبب هذا القصور إلى كون استخدام الكلمة كوحدة للترجمة يؤدي إلى تصفيف الكلمات كل على حدة، وبالتالي تصبح الكلمة معزولة وتفتقر إلى السياق الذي تستمد منه معناها. وقد حثت هذه النتائج المتواضعة الباحثين على التفكير في مناهج جديدة تفضي إلى نتائج أكثر فعالية وبمستوى مقبول، ومن ثم جاءت فكرة استعمال نظم إحصائية تعتمد الجملة كوحدة للترجمة بدل الكلمة.

وقد كان كل من (Och and Ney, 2004) و(Marcu and Wung, 2002) و(Kohen and al., 2003) روادا لهذا المنهج، ويسمح هذا الأخير بتصفيف متعدد للكلمات Multiple word alignement، خلافا للمنهج المعتمد على الكلمة الذي يسمح بتصفيف كلمة مع كلمة أو اكثر، تسمح النظم القائمة على الجملة بتصفيف سلسلة من الكلمات في اللغة الأصل مع سلسلة من الكلمات في لغة الوصل، حيث تترجم سلسلة الكلمات كوحدة واحدة (single unit).

ويعتبر نظام موزس Moses علامة فارقة في مجال الترجمة الآلية بمقاربة إحصائية إذ أحدث قفزة نوعية في مجال الترجمة الآلية، وكان ذلك في سنة 2007 ضمن بحث (Kohen et al., 2007) وذلك في

[13] Design of Moses decoder for Statistical Machine Translation: By Hieu Hoang, Philip Koehn.

إطار مشروع Euro Matrix project، وهو مشروع مصنف ضمن المصادر المفتوحة open source، ومتاح للجميع على شبكة الإنترنت publically available. علما أن مهمة الأداة Moses هو استرجاع الجمل المصفوفة من المتون اللغوية التي خضعت للتصفيف على مستوى الكلمات ضمن أداة Giza++ واستعمالها (أي المتون المصفوفة) في عملية الترجمة.

وقد أحدثت نظم الترجمة الآلية بمقاربة إحصائية والمعتمدة على الجملة كوحدة للترجمة قفزة نوعية في مجال الترجمة الآلية بمقاربة إحصائية، مجسدة بذلك التطور الطبيعي (Brown et al., 1990) للترجمة الإحصائية المعتمدة على الكلمة كوحدة للترجمة. وقد وُجِدَتْ عدة أدوات لفك التشفير قبل Moses ، على غرار **ATS** لِـ (Och and Ney, 2004) و **PHARAOH** لِـ (Kohen, 2004). تقوم الأداة موزس بتحديد الجملة المرفقة بأعلى احتمال والموافقة لجملة نص اللغة الأصل، وبإمكان أداة التشفير انتاج قائمة مرتبة تتضمن الجمل المرشحة بان تكون ترجمة صحيحة لجملة اللغة الأصل، إضافة إلى توفير معلومات مختلفة تبرر اختياراتها (أي الأداة) لا سيما كيفية اختيار الجمل المتطابقة في الثنائية.

2.3 بنية أداة فك التشفير موزس Structure of Moses decoder

في الفقرة التالية سنتناول بنية الأداة موزس **Moses** والوحدات النمطية المكونة لـها : لقد تم تصميم أداة فك التشفير **موزس Moses** بدقة لا متناهية في إطار بنية تتألف من عدة وحدات نمطية (modules) مستقلة عن بعضها البعض علما أن بيئة البرمجة المستعملة هي بيئة غرضية التوجه object oriented.

وتقوم هذه البنية على وحدات نمطية مستقلة تجعل من موزس أداة مرنة، تسمح بإضافة وحدات نمطية جديدة من جهة، ومن جهة أخرى فهي سهلة الصيانة بسبب استقلال وحداتها عن بعضها البعض.

3.3 المكونات الوظيفية الأساسية لـموزس Moses

حرص مصمموا الأداة موزس لفك التشفير Moses decoder على جعل هذه الأداة مصدرا مفتوحا، ومن ناحية أخرى عملوا على تزويدها بمكتبة موارد برمجية موجهة للبحث research-oriented software libraries، علما أن الاختيار لم يكن جزافيا بل لكونها (أي مكتبة الموارد البرمجية المستعملة) تمثل معيارا صناعيا في مجالها، حيث روعي أن يكون محتوى المكتبة مستقلا عن مجالي المعالجة الآلية للغات NLP أو الترجمة الآلية بمقاربة إحصائية SMT. وقد تمت برمجتها بلغةِ C++، كما تمت صياغة موزس وتكييفها مع مكتبة سي.جي.آل CGAL library[14] التي تُسْتَعْمَلُ في الهندسة الحاسوبية Computational Geometry، بالإضافة إلى مكتبة دي.سي.أم.تي.كاي DCMTK والمستعملة في نظم التصوير الإشعاعي Medical imaging المكرسة للمجال الطبي.

لقد تمت برمجة الأداة موزس بلغةِ C++، على غرار مكتبات الموارد البرمجية السالفة الذكر. حيث يتضمن البرنامج 20.000 سطر مقسمةَ على مكتبتين. وتستعمل الأداة موزس Moses المتون اللغوية النظيرة (المتوازية) parallel corpora والمصفوفة على مستوى الكلمات بواسطة الأداة Giza++ لتدريب نماذج الترجمة.

[14] CGAL library (Fabi et al., 2000).

وتجدر الإشارة إلى أن البنية الوظيفية لمكونات Giza++ غير واضحة المعالم مما يصعب مهمة الباحثين المهتمين بتطوير Giza++. على سبيل المثال مشكلة تنفيذ Giza++ على مترجم GCC (GCC compiler) ذو نسخة حديثة.

4. 3 خوارزمية فك التشفير decoding في نموذج موزس[15] Moses

تتميز نماذج الترجمة المحللة إلى عوامل بكثرة مراحلها فهي معقدة مقارنة بالمقاربة المعتمدة علة الجملة كوحدة للترجمة. ففي هذه المقاربة يمكننا استرجاع الجمل التي تمثل إمكانية للترجمة من جدول الجمل المدخلة المقترحة والتي يمكن أن تستعمل إحداها كجملة مدخلة تكافئ كجملة مدخلة input sentence نظيره كجملة مخرجة out put sentence وعادة يتم اعتبار 20 جملة الأولى الأفضل في القائمة أن هذا العدد يمثل اختيار مصمم النظام موزس و ليست قاعدة إذ يمكن استعمال عدد أكبر أو أقل حسب الحاجة.

وتعتمد خوارزمية حزمة البحث لفك التشفير the beam search decoding algorithm على الفرضيات، وبداية لا ترفق أي قيمة بالمتغير hypotheisis وفي هذه الأثناء تكون الخوارزمية قد قامت باسترجاع احتمالات الترجمة من جداول الترجمة .

الخطوة التالية هي إرفاق متغير الفرضيات بكل خيارات الترجمة المتاحة translation option طالما بقيت كلمات من اللغة الأصل دون احتمال للترجمة. ثم تستعمل هذه الفرضيات لتوليد فرضيات جديدة تمثل خيارات جديدة للترجمة وتكرر هذه العملية وستكون الفرضية المكتملة (أي التي تغطي كل الكلمات المدخلة) والمرفقة بأعلى درجة، هي التي تشير إلى أحسن ترجمة وفقا للنموذج. إذ يتم تكييف هذه الخوارزمية كنماذج الترجمة المحللة إلى عوامل، عن طريق المعالجة المسبقة لهذه العوامل وذلك قبل مرحلة البحث الاستكشافي heuristic للخوارزمية كخيارات للترجمة حيث يتم إحصاء كل خيارات الترجمة لجملة مدخلة معينة قبل مرحلة فك التشفير (Decoding) وبالتالي فإن البنية الأساسية لخوارزمية البحث لا تتغير. وتجدر الإشارة إلى ان مقاربة الترجمة المحللة إلى عوامل تجعل عدد خيارات الترجمة مضاعفا مقارنة بالترجمة القائمة على الجملة كوحدة للترجمة لأن توسيع قائمة الكلمات المدخلة عبر إدماج عوامل مدخلة وبالتالي توسيع قائمة الكلمات المخرجة و إدماج عوامل إضافية من شأنه أن يضاعف عدد خيارات الترجمة و يكون عبئا إضافيا على نظام المعالجة وبالتالي فإن هذا الحجم الهائل من الترجمات قد يشكل صعوبات للمعالجة. ومن اجل السيطرة على هذه المشكلة لجا الباحثون إلى استخدام تنفيذ المفكر لعملية التوسيع الناتجة عن النماذج المحللة وذلك بتعيين حد أقصى لعدد الجمل المدخلة والذي حدد بـ 50 جملة مدخلة. وبذلك يتم التحكم نسبيا في عدد الكلمات والعوامل المخرجة.

يبقى هذا الحل مؤقت وجاري البحث عن طرق أكثر فعالية للبحث عن أفضل 50 خيار للترجمة لتقوم مقام الحل الحالي الذي يعتبه الباحثين حلا حدا. نظرا يقصي جملا دون جمل أخرى.

[15] Moses Statistical Machine Translation system , Philip Koehn.

4 تقييم النتائج التطبيقية

4. 1 تقييم نموذج الترجمة المكيف لدعم الثنائية اللغوية (إنجليزية - عربية)

قمنا باختبار نجاعة نموذجنا المُكَيَّفِ لدعم الثنائية اللغوية (إنجليزية - عربية) والذي يعتمد على الأداة Moses لفك التشفير Moses decoder، كما استخدمنا بيانات و حزمة برمجيات (بروتوكول) تابعة لهيئة مِدَار[16] Medar.

وقد اعتمدنا على متن لغوي للتدريب الآلي automatic traing corpus، يتضمن 96063 جملة مصفوفة، أما المتن اللغوي الموجه للاختبار test corpus، فيتضمن 500 جملة، تمثل الخمسمائة (500) جملة الأولى الواردة ضمن متن التدريب.

ولتقييم جودة الترجمة آليا لجأنا إلى استعمال مقياس BLEU وهو مقياس مرجعي في مجال التقييم الآلي للنظم الإحصائية، حيث بلغت درجة BLEU التي تُقَيِّمُ جودة نتائج ترجمة نموذجنا نسبة **31.35%** وهي نتيجة جدّ مشجعة علما بأن متن التدريب المستعمل يتضمن 96063 جملة فقط، وهي جمل مسترجعة من متن لغوي مكرس للأخبار العامة وبديهي أن تكون اللغة المستعملة في هذا النوع من النصوص لغة بسيطة وليست غنية بالأساليب البلاغية لدرجة أن تعكس ثراء لغتين بمستوى العربية والإنجليزية.

وتجدر الإشارة إلى أن درجة BLEU[17] لتقييم نتائج ترجمة Moses الداعم للثنائية اللغوية (إنجليزي – فرنسي) بلغت نسبة **31.10 %** علما بأن المتن الذي اسْتُعْمِلَ لتدريب نماذج الترجمة يحتوي على 1.4 مليون جملة إنجليزية مصفوفة مع ترجمتها باللغة الفرنسية وهي مأخوذة من محاضر جلسات البرلمان الأوروبي.

وبالتالي، إذا كانت نتائج ترجمة نموذجنا قد بلغت 31.35 % بالرغم من استعمالنا لمتن محدود من حيث المضمون ومن حيث الحجم، فإننا نتوقع أن تكون نتائج الترجمة أفضل بكثير إذا توفر لدينا متن لغوي نظير غني ويعكس بحق ثراء الثنائية اللغوية (إنجليزية-عربية) ويكون بمستوى وتنوع متن محاضر جلسات البرلمان الأوروبي.

5 الخاتمة

1- الخلاصة

تجسد المعالجة الآلية للغات الطبيعية خيارا استراتيجيا للدول التي تراهن على ثقافاتها وتصبو لإيجاد مكانة تليق بها في عصر المعلومات information age . وقد كان الهدف من بحثنا هو استقراء وفهم حيثيات مجال الترجمة الآلية وهو مجال يصنف ضمن التقنيات المتقدمة في مجال الذكاء الاصطناعي، وتكريسه لخدمة اللغة العربية، عَلَّنَا نُوَفَّقُ في إضافة لبنة في صرح مجتمع المعرفة العربي، الذي يحتاج لإسهامات كل الكفاءات لتسخير التقنيات الحديثة لخدمة ثقافتنا وإثرائها.

وقد تميز البحث ببيئة عمل ثرية بكل المقاييس إذ سمحت لنا باكتساب خبرة في مجالين متباينين وهما الترجمة والذكاء الاصطناعي، وينكب البحث على دراسة وتكييف محرك الترجمة

[16] http://www.medar.info/index.php

[17] Fundamental and New Approaches to Statistical Machine Translation, By Lucia Specia.

Mosesالدَّاعِـمِ فـي الأصـل للثنائية اللغوية (فرنسية- إنجليزية) بهدف تمكينه من دعم الثنائية اللغوية (إنجليزية- عربية) وبالتالي مـن تـرجـمة النـصوص المُدْخَلَةِ بـاللُّغـَةِ الإنجليزية إلى اللغة العربية بصفة آلية صرفة.

وقد تمكنا من دراسة آلية عمل محرك الترجمة الإحصائية والمتمثل في أداة فك التشفير Moses، ومن ثم القدرة على استكشاف أسباب بعض الأخطاء التي تعود غالبا للثراء الصرفي للغة العربية، وبالتالي اقتراح حلول من شأنها تضييق نطاق الأخطاء الواردة ضمن النصوص المترجمة آليا.

ومن جهة أخرى، أتاح لنا هذا البحث فرصة ثمينة لمواكبة أحدث التطورات في مجال الترجمة الآلية وتقنيات المعالجة الآلية للغات الطبيعية على غرار التحليل الصرفي- النحوي Morpho-syntactic analysis والتحليل النحوي syntactic-analysis. وقد انصب اهتمامنا خاصة بالتقنيات المكرسة لمعالجة اللغة العربية.

وقد تحرينا الدقة في البحث لاستيعاب هذه التقنيات، وخَلُصْنَا إلى أنَّ اعتماد المعالجة اللغوية العميقة كمرحلة تسبق عملية إخضاع النصوص المُدخلة للمعالجة ضمن محرك الترجمة من شأنه أن يُحَسِّنَ جودة النصوص المترجمة آليا، وقد تأكدنا فعلا من ذلك عمليا. إذ لاحظنا أن النصوص التي خضعت لمعالجة لغوية مسبقة، أفضت إلى نصوص مترجمة بجودة أفضل مقارنة مع نصوص التي لم تخضع لمعالجة مسبقة، وهناك أشواط طويلة لا تزال تفصل بيننا وبين مستوى الترجمة المنشودة من طرف المستخدمين، غير أن طريق الألف ميل يبدأ بخطوة. وقد خلصنا في إطار هذا البحث إلى نتيجة مفادها أن تقييم محرك للترجمة الآلية لا يقتصر على استعمال مقياس BLEU أو غيره من المقاييس المتداولة لتقيم جودة الترجمة، بل هناك خطوات يجب انجازها قبل الوصول إلى مرحلة التقييم.وقد تبنينا هذا النهج خلال بحثنا، إذ ارتأينا أنه من الأهمية بمكان، دراسة بيئة تطوير وآلية عمل محرك الترجمة Moses قبل اللجوء إلى التقييم الآلي. وقد كان تنفيذ خطوات التقييم في إطار برمجيات مرجعية معروفة على غرار NIST وCESTA... وغيرها، وقد استقر اختيارنا في النهاية على مقياس BLEU.

كما تَطَلَّبَ تنفيذ هذا البرنامج خطوات دقيقة إضافة إلى برمجيات إضافية تُمَثِّلُ بيئة العمل التي يجب أن يُنَفَّذَ التقييم في إطارها، وقد تساءلنا عن مدى فعالية هذا المقياس في توفير تقييم يعكس المستوى الحقيقي لجودة الترجمة وجدير بالثقة.

وقد أفضت نتائج بحثنا إلى استنتاج وجود صلة وثيقة بين جودة الترجمة ونوع النصوص المترجمة، وذلك ما أكدته النتائج العملية، حيث لاحظنا بأن مستوى الترجمة لا تشوبه شائبة حين يتعلق الأمر بنص تقني أو علمي، وذلك لأن هذه الفئة من النصوص لا تحتمل أي نوع من اللَّبْسِ. وفي المقابل، لا تتطلب ترجمة محتوى البريد الإلكتروني أو المواقع الإلكترونية نفس المستوى من الدقة، وبالتالي لا يمكن اعتماد نفس المعايير في كلتا الحالتين لتقييم مستوى الترجمة.

ولقد توصلنا إلى نتائج جد مشجعة، إذ حقق نموذجنا رغم شح الموارد درجة تقييم بلغت نسبة **31.35%** الثنائية اللغوية (إنجليزية- عربية) باستعمال متن لغوي نظيرٍمتوازي يتضمن 96063 ثنائية جمل مصفوفة فقط علما بأن عملية الترجمة تمت في بضع دقائق ، مقابل **31.10 %** حققها النموذج المصمم أصلا لدعم الثنائية اللغوية (فرنسية - إنجليزية) والذي تم تدريبه بمتن يتضمن 1.4 مليون جملة مصفوفة، علما أنه بإمكاننا إثراء نتائج هذا البحث من خلال تحسين نتائج التحليل اللغوي وعن طريق إدماج معاجم الثنائية اللغوية ومتون لغوية موسومة annotated تم إنشائها بصفة آلية، وباستعمال تقنيات تصفيف المتون النظيرة التي يتبناها الخبراء في نفس الإطار. يبقى أن نشير في النهاية،

إلى أن نتائج ترجمة المحرك الإحصائيMoses والذي قمنا بتكييفه لدعم الثنائية اللغوية (إنجليزية- عربية) مقبولة إلى حَدٍّ ما. غير أن طموحنا أكبر وقناعتنا أن وجود متون غنية وذات حجم كبير من شأنه إحداث قفزة نوعية على مستوى جودة الترجمة، وتجدر الإشارة إلى أنه (أي المحرك الإحصائي Moses) قابل للتَّكَيُّفِ مع أي ثنائية لغوية ويكفي فقط توفير المتون اللغوية الداعمة للثنائية التي نريد إنشاء النموذج من أجلها.

2- مشاريع حالية ومستقبلية

1- تصميم واجهة المستخدم الرسومية graphical user interface لنموذج الترجمة المُكَيَّفِ لدعم الثنائية اللغوية (إنجليزية-عربية).

2- إدماج نتائج تصفيف متون لغوية مختصة specialized corpora ضمن جدول الترجمة للنموذج بهدف تحسين جودة الترجمة مثل المتون المختصة في مجال الطب، الاقتصاد وغيرها.

3- تكييف النموذج الإحصائي لثنائيات لغات أخرى (على غرار الثنائية اللغوية : فرنسية-عربية، عربية - فرنسية).

6 قائمة المراجع

[01] What's new in Statistical MachineTranslation, By Kevin Knight and Philipp Koehn Information Sciences Institute,University of Southern California (pdf Paper).

[02] Machine translation : a brief history . By W.John Hutchins. By W.John Hutchins. From : Concise history of the language sciences: from the Sumerians to the cognitivists. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press, 1995. Pages 431-445]

[03] The Oxford hand book of computational linguistics,By Ruslan Mitkov ,Oxford University Press, 2003 : Chapter 27 : pages 500-511, Machine Translation :General overview , By John Hutchins, Chapter 28 : pages 512-527, Machine Translation :Latest developements, By Harold Somers.

[04] Translation Engines: Techniques for Machine Translation, By Arturo Trujillo. Edition: Springer-Verlag London, Limited 1999.

[05] Statistical Machine Translation, Philipp Koehn Part 1: Morning Session, University of Edinburgh, September,10, 2007.(pdf Paper)

[06] Multiligual systems (basic strategies) : Machine translation : a brief history . By W.John Hutchins

[07] A Statistical approach to machine translation By Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. IBM Thomas J. Watson Research Center Yorktown Heights,Computational linguistics revue, volume 16 N Hutchins.

[08] Moses Statistical Machine Translation System User Manual and Code Guide, By Philipp Koehn, May 23, 2011, page 141 to 157.

[09] A Statistical approach to machine translation By Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. IBM Thomas J. Watson Research Center Yorktown Heights, Computational linguistics revue, volume 16

[10] Linear programming :

http://ar.wikipedia.org/wiki/%D8%A8%D8%B1%D9%85%D8%AC%D8%A9_%D8%AE%D8%B7%D9%8A%D8%A9

[11] http://en.wikipedia.org/wiki/Arg_max

[12] : الاحتمال المشروط

http://www.arab-ency.com/index.php?module=pnEncyclopedia&func=display_term&id=69

[13] Example–Based Machine Translation: http://en.wikipedia.org/wiki/Example-based_machine_translation

[14] Design of the Moses Decoder for Statistical Machine Translation, By Hieu Hoang, Philipp Koehn, University of Edinburgh , Software Engineering, Testing, and Quality Assurance for Natural Language Processing , workshop ACL-08: HLT, June 20, 2008 The Ohio State University Columbus, Ohio, USA. (pdf Paper).

[15] Moses Statistical Machine Translation system, Philip Koehn.

[16] http://www.medar.info/index.php.

[17] Fundamental and New Approaches to Statistical Machine Translation, By Lucia Specia.

مراجع باللغة العربية

[01] لسان العرب لابن منظور- الاصدار 1.03 .pdf book.

[02] المعجم الوجيز، مجمع اللغة العربية، جمهورية مصر العربية الطبعة الأولى،1980.

[03] معجم القاموس المحيط، الطبعة الثانية. مجد الدين محمد بن يعقوب الفيروز آبادي، توثيق: خليل مأمون شيحا. دار المعرفة للطباعة والنشر والتوزيع، بيروت – لبنان، 2007.

[04] معجم إلكتروني : الباحث العربي- قاموس عربي-عربي www.baheth.info الموقع الإلكتروني:

[05] قاموس المورد عربي- إنجليزي لِلدكتور : روحي البعلبك Al-Mawrid: A Modern English – Arabic Dictionary,
Munir Ba'albki, Dar El-Ilm Lil-Malayen, Beirut – Lebanon 2007. 41st ed,

[06] قاموس المبرق: للدكتور محمد إبراقن قاموس موسوعي للإعلام والاتصال فرنسي – عربي طبعة ثانية منقحة، منشورات ثالثة- الأبيار، الجزائر، 2007.

[07] The Dictionary English-Arabic General & Scientific dictionary of language and terms By research and studies center. Dar el Kitab Al Hadith Caire-Koweit-Algerie, 2nd Edition -2005-Dar Al Kotob Al Ilmiyah.Beyrouth- Liban.

[08] Longman Advanced American Dictionary, Pearson Education Limited, Essex – England, 2003.

[09] Merriam-Webster's Collegiate Dictionary, 11th ed. Merriam-Webster's, Incorporated, Massachusetts – U.S.A, 2004.

[10] Oxford Advanced Learner's Dictionary, 6th ed. Editor: Wehmeier, Sally, Michael Ashby. Oxford University Press, Oxford – England, 2000.