# Survol des méthodes des systèmes de traduction automatique des dialectes Arabes

Ridouane TACHICART & Karim BOUZOUBAA
Ecole Mohammadia d'Ingénieurs, Rabat, Maroc.

## I. INTRODUCTION

La traduction Automatique (TA) est un processus qui consiste à transcrire un texte d'une langue naturelle (par exemple, l'anglais) vers une autre (par exemple, le français). Ce processus apparemment simple est en réalité complexe car il ne se limite pas à une simple substitution mot à mot. Il doit analyser et interpréter le texte et comprendre les relations entre les mots qui peuvent en influencer le sens. Ceci requiert une connaissance de la grammaire, de la syntaxe (structure de la phrase) et de la sémantique (sens des mots), à la fois dans la langue source et dans la langue cible. Le recours à la traduction automatique existe dans le domaine scientifique et technique, commerciale, littéraire, militaire, médicale, audio-visuelle, etc. Son défi majeur est de produire des traductions comparables à des traductions humaines. C'est pour cela qu'il existe des méthodes pour la construction des systèmes de traduction qui produisent des qualités différentes selon l'approche adoptée et les ressources employées.

## II. MÉTHODES EXISTANTES DE LA TRADUCTION AUTOMATIQUE

Globalement, il existe deux stratégies fondamentales pour la réalisation d'un système de traduction. D'une part l'approche basée sur les règles repose sur une analyse morphologique en exploitant les informations linguistiques de la langue source et cible. Elle se déroule en trois phases (Figure 1) dont la première est l'analyse qui produit une série de découpages pour déterminer la structure grammaticale de chaque mot, sachant que souvent un analyseur morphologique est introduit dans cette phase. Ensuite, chaque analyse précédente est associée à une ou plusieurs analyses de la langue cible en utilisant un dictionnaire bilingue dans la phase de transfert. Enfin, un texte en langue cible est produit dans la phase génération en respectant un ordre approprié.

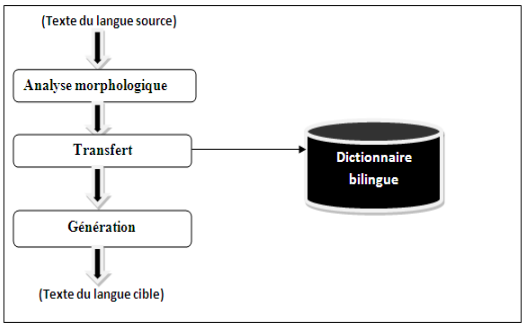


Figure 1. Traduction à base des règles

D'autre part, l'approche statistique considère que chaque phrase de la langue cible est une traduction d'une phrase de la langue source avec une certaine probabilité, et ce à partir de l'analyse d'un corpus parallèle par un modèle de traduction et l'analyse d'un corpus monolingue de la langue cible par un modèle de langue. Ce dernier introduit les contraintes imposées par la syntaxe de la langue cible et estime la probabilité d'une phrase de cette langue, tandis que le modèle de traduction modélise le processus de génération d'une phrase source à partir d'une phrase cible. Le résultat généré par ces deux modèles est utilisé par un décodeur qui calcule la probabilité de traduction de la phrase source vers la phrase cible en un temps acceptable (Figure 2).

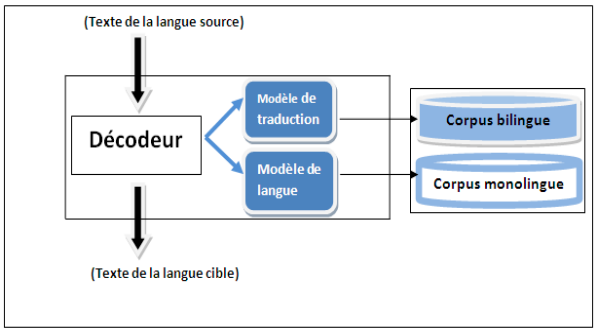


Figure 1. Traduction statistique

Nous identifions dans la partie suivante les étapes qui constituent le processus de la construction d'un système de traduction, et donnons des illustrations en termes de travaux, en se limitant aux dialectes arabes.

## III. TRAVAUX DANS LA TRADUCTION AUTOMATIQUE DES DIALECTES ARABES

### *A. Construction des dictionnaires*

Un dictionnaire bilingue est employé dans la phase de transfert de la méthode basée sur les règles pour produire l'équivalent d'une entrée source dans la langue cible. Dans ce cadre, AlSabbagh R. et Girju R. (2010) ont construit un dictionnaire bilingue pour le dialecte Egyptien en utilisant une méthode associative basée sur la corrélation de la cooccurrence des mots dans quatre dialectes arabes: l'Egyptien, le Marocain, l'Iraquien et le Golfe, et ce à partir des requêtes dans le web. Cette approche (circulaire) a permis de surmonter le problème de rareté des corpus en dialectes Arabes puisque l'acquisition d'une occurrence des mots pour un dialecte est conditionnée juste par l'acquisition de l'autre occurrence dans l'autre dialecte. L'évaluation de performance de cette méthode a atteint 70,9% en F-mesure.

K. Meftouh, et K. Smaili (2013) ont présenté une méthode pour la construction d'un corpus parallèle (Arabe Moderne Standard -MSA- versus dialecte Algérien) afin d'être utilisé dans le développement d'un système de traduction pour ces deux langues. En effet, ils ont commencé d'abord par l'enregistrement audio des discussions réelles en dialecte algérien entre les gens dans des situations différentes (Hôpital, Ecole, Administrations,

etc.) puis, ils ont enlevé les segments bruités dont le son n'est pas clair. Le résultat de ce traitement leur a permis de construire l'équivalent de dix heures des discussions. Ensuite, ils ont réalisé une transcription manuelle en lettres arabes de ces enregistrements pour construire un corpus monolingue, et ont fait une extraction de tous les mots qui seront affiliés à leurs correspondants en MSA pour construire un dictionnaire bilingue. Enfin ils ont remarqué que 65% du vocabulaire algérien est d'origine arabe standard, tandis que 19% est d'origine Française. Le reste de ce vocabulaire (16%) est partagé entre l'origine Turque et Amazighe.

Boujelbane, R. et al (2013) ont développé un dictionnaire bilingue annoté en vue de créer un corpus du dialecte tunisien et un modèle de langue pour un système de reconnaissance de la voie. En effet, ils ont exploité le corpus Penn Arabic Treebank[1] en MSA pour extraire les correspondances en dialecte tunisien des verbes qui existent dans ce corpus. Ils ont constaté que 60% des verbes changent totalement dans leurs morphologies durant la traduction. Enfin, pour évaluer la qualité de ce dictionnaire ils ont calculé la corrélation entre leurs traductions avec une traduction des spécialistes, celle-ci a atteint 74,79%.

### *B. Construction de corpus*

Cette étape constitue un élément basique pour la méthode statistique car les systèmes de traduction statistique se basent sur l'analyse des corpus monolingue et bilingue afin de construire des paramètres pour un décodeur qui produit la probabilité de la traduction finale. Boujelbane, R. et al (2013) ont travaillé sur la génération automatique d'un corpus en dialecte Tunisien à l'aide d'un outil développé spécialement pour ce fait. En effet, le texte source en MSA est analysé d'abord par l'analyseur morphologique MADA[2] qui ajoute à chaque mot source des informations morphologiques et lexicales. Ensuite, cet outil exploite un dictionnaire bilingue annoté pour faire la correspondance entre le mot source et le mot cible.

(Graja M, Jaoua M, et Belghith, L 2010) ont produit un premier corpus tunisien 'TuDicol' à partir des dialogues enregistrés dans les stations de train entre les voyageurs et agents, puis ont présenté une étude sur le lexique de ce corpus qui contient 127 dialogues combinant 893 discours dont le nombre total des mots est 3403. Le résultat de l'analyse du corpus 'TuDicol' montre que 11,81% des mots du corpus ont une origine française.

Chiang et al. (2006) ont étudié le problème d'analyse de la transcription du dialecte levantin parlé dans les pays du moyen orient (Liban, Syrie et Jordanie). Ils affirment qu'il est possible de construire des corpus annotés à partir des lexiques qui relient des lexèmes levantins avec leurs équivalents en MSA Modern Standard Arabic, et cela en connaissant les différences morphologiques et syntaxiques qui existent entre ces deux langues. Les lexiques crées sont ensuite utilisés pour tester un texte en dialecte Levantin transcrit.

Dans le même sens, Belgasem M. (2009) a construit un corpus de différents dialectes arabes à partir des émissions TV. En effet, il s'agit de télécharger et enregistrer les émissions TV de différents dialectes arabes qui consistent essentiellement en débats politiques, puis de transcrire ces enregistrements à l'aide de l'outil Transcriber[3]. Les différentes étapes du travail de transcription sont : la segmentation de la bande son, l'identification des tours de paroles et des locuteurs, l'identification des sections thématiques, la transcription orthographique et la vérification.

### *C. Convention d'orthographe des dialectes*

En raison de difficultés trouvées lors du traitement du contenu en dialectes arabes, un seul mot peut être transcrit de différentes manières. Par exemple, en dialecte Marocain, le mot **غناكل** /je vais manger/ peut être aussi écrit comme **غاناكل**. Des chercheurs ont proposé des normes et standards pour l'écriture de ces dialectes dans le but de produire des systèmes de traduction plus robustes. En effet, Nizar Habash et al. (2012) ont proposé CODA comme méthode pour une écriture standard des dialectes arabes et ce dans le but de faciliter le traitement automatique de ces langues. En effet, chaque mot en dialecte a une unique forme orthographique dans CODA qui représente sa morphologie et phonologie. Cette méthode implémente les règles phonétiques de l'arabe standard. Par exemple en dialecte Marocain le mot كراج 'garage' se prononce **'**garage**'** mais son équivalent en CODA est **جراج** . De même pour le mot راسك /ta tête/ qui se prononce **'**rasek**'** mais son équivalent en CODA est **رأسك**.

### *D. Analyseurs morphologiques*

Dans les systèmes de traduction basés sur les règles, un analyseur morphologique est souvent intégré et intervient dans la phase Analyse pour décomposer le mot source en une succession de préfixe, radical et suffixes possibles qui constituent des entrées pour la phase suivante.

Habash N. et Rambow (2006) ont développé MAGEAD un système à base de règles qui permet de décrire les systèmes morphologiques des différentes variétés de l'arabe (dialecte Levantin et MSA) et de les compiler sous la forme d'un transducteur fini. MAGEAD effectue une analyse morphologique profonde. Partant d'une forme verbale ou nominale de l'arabe, il en fait l'analyse sous la forme d'une racine, d'une classe et de traits morphologiques. La précision de cet analyseur a atteint 94.2%.

Habash N. (2011) ont construit ADAM un analyseur morphologique des dialectes Egyptien et Levantin sur la base de l'analyseur morphologique de l'Arabe standard Buckwalter[4], En effet ils ont étendu sa liste des préfixes, suffixes et schèmes par ceux de ces deux dialectes. ADAM a atteint une précision de 84,4 % sur un ensemble des mots qui contient les dialectes déjà cités.

[1] http://www.ircs.upenn.edu/arabic/

[2] http://www1.cs.columbia.edu/~rambow/software-downloads/MADA_Distribution.html

[3] http://trans.sourceforge.net/

[4] http://catalog.ldc.upenn.edu/LDC2002L49

Khalid Almeman et Mark Lee (2012) ont travaillé sur un analyseur morphologique multi-dialectes de l'arabe qui utilise des ressources différentes. D'abord, ils ont utilisé l'analyseur morphologique Alkhalil après son adaptation avec les dialectes arabes pour décomposer les mots et identifier les suffixes et préfixes puis les analyser. Ensuite, si le résultat n'est pas satisfaisant, le web est employé pour tirer des statistiques sur les segments du mot déjà analysé. Ainsi le segment qui a le pourcentage le plus élevé est le schème. Ils ont abouti à 94% de précision sur un corpus de différents dialectes arabes.

### *E. Systèmes de traduction automatique des dialectes Arabes*

C'est le processus qui permet de construire un système de traduction avec une architecture qui rassemble plusieurs étapes en se basant sur une approche par règles, statistique ou hybride. Plusieurs chercheurs ont exploité des outils existants du traitement automatique de l'arabe standard pour développer leurs systèmes de traduction des dialectes arabes.

Yahya alAmlahi (2007) a présenté un algorithme pour un système de traduction du dialecte Yemenien vers le MSA sans utilisation d'outils mais à l'aide d'un algorithme qui analyse les mots de ce dialecte sur la base de la liste de ses affixes. Cet algorithme est basé sur les règles morphologiques. Il fait la Tokenization du texte source, analyse chaque Token et vérifie si la racine du mot a conservé sa forme d'origine de l'arabe standard pour appliquer certaines règles de transfert.

Shaalan et al (2007) étaient concentrés sur le problème de traitement automatique du texte en arabe standard qui contient quelques segments du dialecte égyptien. Pour résoudre ce problème ils ont d'abord traduit les mots en dialecte égyptien vers l'arabe standard en utilisant des dictionnaires bilingues et les règles de transfert puis ont procédé au traitement de ce texte en utilisant des outils du traitement de l'arabe standard. Ils ont noté une correspondance intéressante entre le dialecte égyptien et l'arabe standard.

Khaled Shaalan et Hitham M. Abo Bakr (2008) ont construit leurs système de traduction du dialecte égyptien vers l'arabe standard sur la base de l'analyseur morphologique Buckwalter. En effet, ce dernier a été alimenté par des éléments annotés du dialecte égyptien, puis un nouveau fichier a été introduit pour coder les règles de correspondance entre le dialecte égyptien et le MSA.

(Hitham M. Abo Bakr et al. 2008) ont présenté une approche hybride pour générer l'arabe standard avec les diacritiques à partir du dialecte égyptien sans diacritiques. Ils ont introduit cette fois une nouvelle annotation de la nature de mot (POS) pour annoter les données du dialecte égyptien. En effet le texte source est transformé en translitération de Buckwlater. Enfin un processus génère le texte traduit en arabe standard avec les diacritiques.

(Wael Salloum et Nizar Habash 2011) ont présenté une méthode basée sur les règles pour produire la traduction des dialectes arabes vers l'arabe standard en se limitant aux mots hors vocabulaire et les mots de faible fréquence. Ils ont enrichit ADAM avec les préfixes et les suffixes des deux dialectes égyptien et Levantin puis ils ont utilisé les règles de transfert pour générer les traductions.

Ahmed Hamdi, et al. (2012) ont réutilisé MAGEAD pour traduire les verbes du dialecte tunisien. Ils ont alimenté MAGEAD par les préfixes et suffixes du dialecte tunisien pour ce fait, et ont abouti à une précision de 75%.

Emad Mohamed, e al. (2012) ont présenté une méthode basée sur les règles pour générer automatiquement la traduction des mots en arabe standard vers le dialecte égyptien. Leur hypothèse se base sur l'observation de la morphologie du dialecte dont le schème est généralement importé de l'arabe standard, tandis que les préfixes et suffixes changent. Le pourcentage des mots inconnus a atteint 16.66 %.

Wael Salloum et Nizar Habash (2012) ont utilisé ADAM, analyseur morphologique des dialectes arabes, pour créer le système ELISSA, basé sur les règles, qui traduit les dialectes arabes (Levantin, Iraquien et Egyptien) vers le MSA. Dans une première étape (Analyse) le texte introduit en dialecte est analysé par ADAM, puis dans l'étape de transfert qui peut être réalisée de deux manières : traduction surfacique utilisant des dictionnaires bilingues ou traduction profonde utilisant l'outil MADA + TOKAN pour produire des mots en arabe standard qui constituent une combinaison de mots candidats à former une phrase. Enfin, l'outil SRILM[5] qui aide à construire un modèle de langue est introduit pour sélectionner la combinaison qui a une forte probabilité et qui va constituer ainsi la traduction finale de la phrase source.

### *F. Métriques d'évaluation*

Pour tester la qualité de la traduction produite par le système de traduction développé, il existe des métriques qui assurent leurs évaluations par rapport à une traduction humaine de référence.

Le score BLEU (Bilingual Evaluation Understudy) est proposé par (Papineni et al 2001). L'idée principale est la comparaison de la sortie du traducteur avec une/des traductions de référence. Les statistiques de cooccurrence et de n-grammes, basées sur les ensembles de n-grammes pour les segments de traduction et de référence, sont calculées pour chacun de ces segments et sommées sur tous les segments. Cette moyenne est multipliée par une pénalité de brièveté, destinée à pénaliser les systèmes qui essaieraient d'augmenter artificiellement leurs scores en produisant des phrases délibérément courtes. Le score BLEU varie de 0 à 1 et il est d'autant meilleur qu'il est grand. BLEU a gagné le statut de mesure automatique de référence au sein de la communauté de traduction automatique.

Lavie et al. (2004) proposent la méthode METEOR qui est désignée à l'amélioration de la corrélation entre la traduction des systèmes de traduction et la traduction humaine au niveau des segments. La mesure est basée sur la moyenne harmonique des unigrammes de Précision (p) et de Rappel (r). Le score METEOR est calculé par la formule :

[5] http://www.speech.sri.com/projects/srilm/

$$METEOR = 10 \times p \times r \div (r + 9p) \times (1 - p). \quad (1)$$

*Où*

$$p = 0{,}5 \times (c \div um)^3. \quad (2)$$

Sachant que um est le nombre des unigrams en correspondance et c le nombre total des phrases.

## CONCLUSION

Nous avons présenté un aperçu sur les systèmes de traduction automatique ainsi que les méthodes utilisées pour les construire. Dans notre étude nous nous sommes limités aux systèmes qui traitent les dialectes arabes. Parmi nos travaux futurs, nous allons travailler sur le développement d'un système de traduction de l'arabe Marocain vers l'arabe standard.

## REFERENCES


Habash N., Diab M., Rambow O. (2011). Conventional Orthography for Dialectal Arabic. In: Proc. Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 21-27.

Habash N., Rambow O. (2011). MAGEAD: a morphological analyzer and generator for the Arabic dialects (2006). In: Proc. 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Stroudsburg, USA, July 2006.

Abo Bakr H., Shaalan K., Ziedan I. (2007). A Transferring Egyptian Colloquial Dialect into Modern Standard Arabic. In: Proc. International Conference on Recent Advances in Natural Language Processing, RANLP 2007, Borovets, Bulgaria, September 27-29.

Salloum W., Habash N. (2012). Elissa: A Dialectal to Standard Arabic Machine Translation System. In: Proc. 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, December 2012.

Almalahi Y., Fateh, A., (2007). Sana'ani Dialect to Modern Standard Arabic :Rule-based Direct Machine Translation.

Abo Bakr H., Shaalan K., Ziedan I. (2008). A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In: Proc. 6th International Conference on Informatics and Systems, INFOS 2008, Cairo, Egypt, Mars 27-28.

Mohamed E., Mohit B., and Oflazer K., (2012) Transforming Standard Arabic to Colloquial Arabic. In: Proc. 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012, Jeju, Korea, July 8-14.

Zbib R., Malchiod E., Devlin J., Stallard D., Matsoukas S. Schwartz R., Makhoul J., Zaidan O., Callison-Burch C. (2012) Machine translation of Arabic dialects. In: Proc. 12th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2012, Montreal, Canada, June 3-8.

Salloum W., Habash N. Plamondon, L., (2011). Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In: Proc. Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland, UK, July 27-31.

Boudlal A., Lakhouaja A., Mazroui A., Meziane A., Ould abdallahi ould bebah A., Shoul M. (2010).Alkhalil Morpho Sys1: A Morphosyntactic analysis system for Arabic texts. In: Proc. The International Arab Conference on Information Technology, ACIT 2010, Benghazi, Libya, December 14-15.

Stolcke A. (2002) SRILM An Extensible language modeling toolkit. In Proc. 7th International Conference on Spoken Language Processing, ICSLP 2002, Denver, Colorado, USA, September 16-20.

Denkowski M., Lavie A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In: Proc. 6th Workshop on Statistical Machine Translation, EMNLP 2011, Edinburgh, Scotland, UK, July 30–31, 2011.

Papenini K., Roukos S., Ward T., Zhu W. (2002). BLEU: a method for automatic evaluation of machine translation. In: Proc. 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, Pennsylvania, Philadelphia, USA, July 7-12.

Hamdi A., Boujelbane R., Habash N., Nasr A. (2013). Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. In: Proc. 20eme conférence sur le Traitement Automatique des Langues Naturelles, TALAN 2013, Sables d'Olonne, France, June 17-21.

Tim Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0. ", Tim Buckwalter ed Linguistic Data Consortium: University of Pennsylvania, 2002.