Etude comparative des approches de traduction automatique

Imane TAGHBALOUT[1], Fadoua ATAA ALLAH[2], Mohamed ELMARRAKI[1]
[1]LRIT, Faculté des Sciences, Mohammed V-Agdal, Rabat, Maroc
[2]IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
Taghbalout.imane@gmail.com, ataaallah@ircam.ma, marraki@fsr.ac.ma

***Résumé*—Dans la perspective de réaliser un système de traduction automatique pour une langue peu dotée, cet article présente un état de l'art permettant d'introduire les concepts nécessaires et les notions de base sur la traduction automatique (TA). L'élaboration d'un système de TA peut être basée soit sur une approche statistique soit sur une approche à base des règles linguistiques ou bien sur une approche hybride qui mixe les deux.**

Mots clés : traduction automatique statistique, traduction automatique linguistique, langues peu dotées.

I. INTRODUCTION

Ces dernières années, les besoins en systèmes de traduction automatique augmentent de plus en plus. Les premières propositions pour la TA empirique ont été formulées en 1949 par Warren Weaver. La première démonstration d'un système de TA s'est tenue en 1954 et elle est connue sous le nom de l'«Expérience de Georgetown - IBM» [1]. Avant l'année 1990, les travaux menés sur la traduction automatique ont été typiquement focalisés sur une approche experte utilisant des analyseurs syntaxiques et sémantiques, aujourd'hui les approches statistiques fondées sur l'apprentissage automatique à partir de corpus bilingues, ainsi que les approches hybrides sont de plus en plus adoptées par de nombreux chercheurs. Dans cet article, nous allons aborder dans la deuxième section les principales approches statistiques et linguistiques utilisées pour la TA, après nous présenterons les avantages et les inconvénients qui caractérisent chacune de ces approches. La dernière partie sera consacrée aux méthodes d'extraction des corpus parallèles.

II. LES DIFFERENTES APPROCHES DE TA

A. *Approches linguistiques*

Le processus de TA linguistique se déroule en trois phases fondamentales [2] :

- L'analyse : analyser le texte source en des représentations intermédiaires en langue source ;
- Le transfert : transférer ces représentations intermédiaires vers des représentations intermédiaires en langue cible ;
- La génération : générer le nouveau texte en langue cible à partir des représentations intermédiaires en langue cible.

Le « triangle de Vauquois » proposé dans [2] décrit les différentes architectures linguistiques possibles d'un système de TA. Chaque chemin dans le triangle correspond à une architecture linguistique "Fig.1".

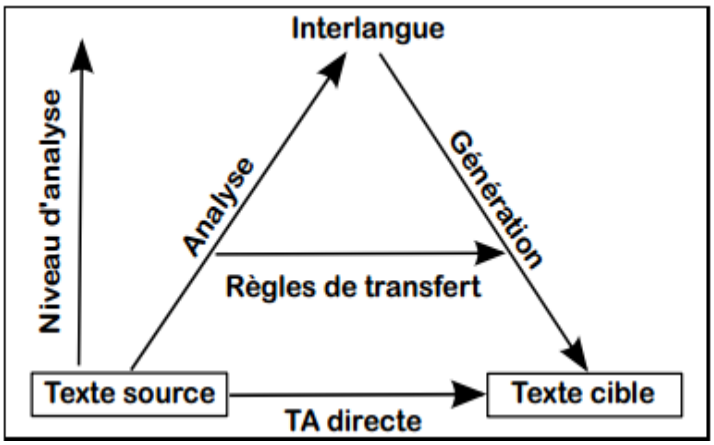

Figure 1 : Triangle de Vauquois, représentation des différentes architectures linguistiques

1) *Système de traduction directe :*

La phrase source est segmentée en des mots et le système exécute la traduction mot à mot. L'étape de transfert utilise une table bilingue qui associe pour chaque mot ou séquence de mots source un ensemble de règles de traduction et de réarrangement qui permettent de traduire et réordonner les mots dans la phrase cible.

2) *Système de TA à base de règles de transfert*

La phrase source est analysée à l'aide d'un analyseur syntaxique et d'une grammaire. Cette phase d'analyse donne lieu à une représentation arborescente qui est ensuite convertie dans la langue cible. Pour assurer le passage de la représentation intermédiaire source à la représentation cible, une table bilingue, contenant les règles de transfert entre les représentations source et cible, est requise. La "Fig.2" ci-

dessous, extraite de [3], illustre un exemple de traduction à base de transfert.

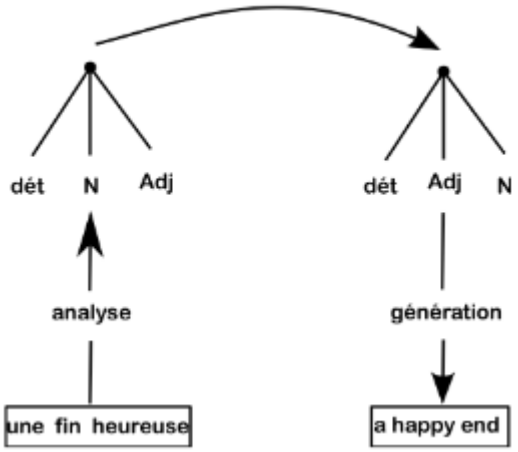

Figure 2. Exemple de traduction d'un groupe nominal en français vers une phrase en anglais

3) *Système de traduction par interlingue*

La traduction de n'importe quelle langue source vers n'importe quelle langue cible est le processus qui consiste à « enconvertir » la phrase source vers la représentation pivot puis à «déconvertir» la phrase cible à partir de cette représentation pivot. Le transfert des représentations de la langue source vers celles de la langue cible n'a plus lieu d'être. L'avantage de cette méthode est la possibilité de l'appliquer dans un environnent multilingue. Pour couvrir tous les sens de traduction entre *n* langues, nous n'avons besoin que de *n* modules d'enconvertion et de *n* modules de déconvertion.

Cependant, la complexité de cette méthode réside dans l'obligation de construire un vocabulaire pivot pour représenter tous les concepts possibles de toutes les langues et les liens des concepts entre deux langues. La construction peut être basée sur une langue artificielle «logique», sur une langue auxiliaire « naturelle » (comme l'anglais ou l'espéranto), sur un ensemble de concepts primitifs communs à toutes les langues, ou sur un vocabulaire « universel ». UNL (Universal Networking Language) est un exemple de langage pivot [4]

B. *Approches statistiques*

La traduction automatique statistique permet de traduire automatiquement un texte d'une langue source vers une langue cible en procédant à l'apprentissage automatique du système en appliquant des calculs mathématiques statistiques sur des données bilingues alignées. Un texte est traduit en fonction de la loi de probabilité P(*e*|*f*) qui calcule la probabilité pour que la phrase en langue cible *e* soit une traduction de la phrase *f*.

La TA statistique nécessite deux modèles : un modèle de langage, et un modèle de traduction. Ces modèles se basent sur la théorie mathématique de distribution et d'estimation probabiliste de Frederick Jelinek, développée à IBM T.J. Watson Research Center [5].

Le but de la traduction statistique est de trouver la phrase Ê, qui maximise P(*e*|*f*) étant donnée une phrase source *f*. Mathématiquement : Ê = argmaxE P(*e*|*f*).

D'après le théorème de Bayes : P(*e*|*f*) = P(*f*|*e*).P(*e*)/P(*f*)

Puisque le dénominateur P(*f*) est indépendant de P(*e*) la maximisation devient alors:

Ê = argmaxE P(*e*|*f*) = argmaxE P(*e*)* P(*f*|*e*).

P(*e*) est appelé le modèle de la langue cible, tandis que P(*f*|*e*) est appelé un modèle de traduction. Ainsi, la traduction statistique repose sur deux modèles:

- Modèle de traduction qui propose pour chacun des segments de la phrase source, des traductions candidates en langue cible ;
- Modèle de langue qui se charge de capter les contraintes imposées par la syntaxe de la langue cible par une ou plusieurs fonctions probabilistes.

La "Fig.3" ci-dessous illustre l'architecture d'un système de traduction statistique.

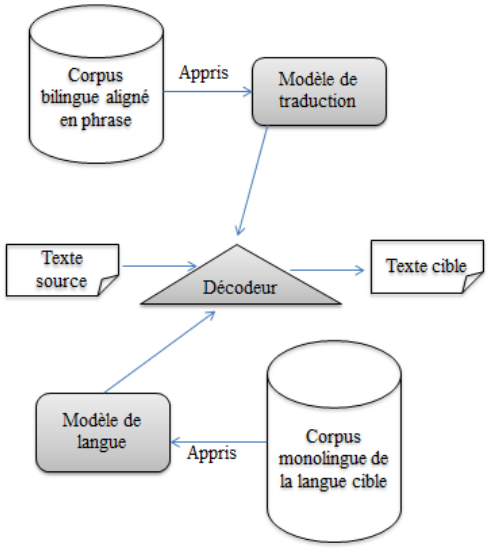

Figure 3: Exemple d'architecture d'un système statistique de traduction automatique

Les méthodes statistiques peuvent être basées soit sur des corpus ou bien sur des exemples. La table.1 définit les deux types des approches statistiques.

TABLE 1 METHODES STATISTIQUES

Méthodes statistiques	
Basées sur des corpus	***Basées sur des exemples***
Apprentissage automatique des "règles" de traduction à partir d'un corpus bilingue [6].	Réutilisation des exemples de traductions existantes comme base pour la nouvelle traduction [7].

Les modèles de traduction automatique statistique étaient à base de mots où les phrases sont traduites mot par mot. Ensuite, Koehn, Och, et Marcu ont amélioré ces modèles et ont fondé des systèmes de TA à base de séquences de mots.

Ces modèles représentent, aujourd'hui, l'état de l'art de la traduction automatique statistique [8].

C. Traduction automatique hybride (mixte)

Ces dernières années, les chercheurs s'intéressent de plus en plus aux approches mixtes qui pourraient tirer profit des forces de chacune des approches statistiques et linguistiques. Par exemple, des hypothèses issues d'un système de traduction empirique peuvent être triées à l'aide d'informations linguistiques sources ou cibles.

III. AVANTAGES ET INCONVENIENTS DE CHAQUE APPROCHES DE TRADUCTION

Dans cette section, nous allons mettre le point sur les avantages et les limitations des approches linguistiques et statistiques.

A. Approches à base des linguistiques

La table.2 regroupe certains avantages et inconvénients des approches linguistiques

TABLE 2 AVANTAGES ET INCOVENIENTS DES APPROCHES LINGUISTIQUES

Prérequis	Entrées de dictionnaire et des règles linguistiques
Avantages	La qualité de la traduction est très satisfaisante surtout lorsqu'il s'agit d'une traduction spécialisée.
Inconvénients	L'élaboration des règles linguistiques est couteuse en termes de temps et d'effort.

B. Approches à base des statistiques

La table.3 regroupe certaines forces et limites des approches statistiques

TABLE 3 AVANTAGES ET INCOVENIENTS DES APPROCHESSTATISTIQUES

Prérequis	Mémoires de traduction (des corpus parallèles)
Avantages	-La qualité de la traduction s'améliore automatiquement au fur et à mesure que la base d'apprentissage s'enrichie ; - Peu de ressources linguistiques nécessaires ; -Rapide à implémenter et facile à maintenir ; -Portabilité : Facilité à étendre vers une nouvelle langue ; -Possibilité d'évaluer de façon expérimentale des systèmes pratiques et des hypothèses scientifique.
Inconvénients	Difficulté de construire des ressources linguistiques suffisantes en quantité et en qualité.

Comme présenté dans la table.2, le prérequis pour l'élaboration d'un système de TA statistique est la disponibilité d'un corpus de taille importante de données parallèles pour l'apprentissage. Or ce type de corpus n'est pas toujours disponible en quantité suffisante pour une langue peu dotée. Dans la section suivante, nous citons un ensemble de méthodes de construction et d'extraction des corpus parallèles.

IV. EXTRACTION DES CORPUS PARALLELES

Les méthodes les plus communes pour construire des corpus parallèles consistent en :

- des méthodes automatiques qui collectent des paires de phrases parallèles à partir du Web [9], [10] ;
- des méthodes d'alignement qui extraient des documents/phrases parallèles à partir de deux corpus monolingues [10].

Il y a aussi les méthodes d'extraction de paires de phrases parallèles à partir d'un corpus comparable [1][11].

Aussi, DO [1] a proposé trois méthodes pour l'extraction d'un corpus d'apprentissage parallèle à partir d'un corpus comparable : La première méthode suit l'approche de recherche classique qui utilise des caractéristiques générales des documents ainsi que des informations lexicales du document pour extraire à la fois les documents comparables et les phrases parallèles. Or, cette méthode requiert des données supplémentaires sur la paire de langues. Cependant, la deuxième méthode est une méthode entièrement non supervisée. Elle peut être appliquée pour toute paire de langues, pour un corpus comparable en entrée seulement en utilisant la technique RIT (Recherche d'Information Translingue). La dernière méthode est une extension de la deuxième méthode, elle utilise une troisième langue pour améliorer le processus d'extraction de deux paires de langues.

V. CONCLUSION ET PERSPECTIVES

Cet article décrit les approches principales de la traduction automatique à savoir : l'approche linguistique et statistique. En outre, il met en évidence les prérequis, les forces et les faiblesses de chacune d'elles.

Suite à cette étude, il s'est avéré que la TA à base de l'approche statistique "Table.3" présente plus d'avantages par rapport à la TA à base des règles linguistiques, qui reste une méthode couteuse en termes d'effort et du temps. Or, pour une langue peu dotée, telle que l'amazighe, qui souffre de la rareté des ressources linguistiques surtout en matière de textes parallèles et comparables, un système de TA linguistique s'avère le plus adéquat. Et dans ce contexte, nous optons, dans une première étude, pour le choix d'une approche de TA par interlangue, suite aux avantages qu'elle présente, principalement, dans le support d'environnement multilingue.

[1]Un corpus comparable contient des données qui ne sont pas parallèles c'est-à-dire les phrases des deux côtés ne sont pas alignées, mais «étroitement liés par les mêmes contenus » [1].

REFERENCES

[1] T. DO, "Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée ". Thèse de doctorat. l'École Doctorale Mathématiques, Scienceset Technologies de l'Information. France. 2011.

[2] B. Vauquois, "A Survey of Formal Grammars andAlgorithms for Reconition andTranslation", FIP Congress-68.Edinburg, 254-260, 1968.

[3] C. Lavecchia, "Les triggers inter-langues pour la traduction automatique statistique". Thèse de doctorat, Université de Nancy 2. 2010.

H. Uchida, M. Zhu, "The universal networking language beyond machine translation". UNDL Foundation. Japan. 2001.

[4] P. F. Brown, V. J. Pietra, S.A.D Pietra, R.L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation". IBM T.J. Watson Research Center. 264-311.1993.

[5] P. F. Brown, V. J. Pietra, S.A.D Pietra, R.L. Mercer, "A statistical approach to machine translation". IBM Thomas J. Watson Research Center Yorktown Heights, NY , 79-85.1990.

[6] M. Nagao, "A framework of a mechanical translation between Japanese and English by analogy principle". In proceedings of the Intl. NATO symposium on Artificial and human intelligence, Lyon, France , 173-180. 1984.

[7] P. Koehn, F.Josef Och, D. Marcu, "Statistical phrase-based translation". In NAACL ″03 : Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Morristown, NJ, USA. Association for Computational Linguistics. 48-54. 2003.

[8] P. Resnik, "Parallel Strands: a preliminary investigation into mining the web for bilingual text". In proceedings of AMTA. 1998.

[9] P. Resnik, "Mining the Web for bilingual text". In proceedings of ACL. 1999.

[10] A.Patry, P.Langlais, "Paradocs: un système d'identification automatique de documents parallèles". In proceedings of TALN. 2005.

[11] B. Zhao, S. Vogel, "Adaptive parallel sentences mining from Web bilingual news Collection". In proceedings of IEEE International Conference on Data Mining. 2002.