



HAL
open science

Traduction automatique de documents manuscrits et typographiés en arabe par couplage étroit entre systèmes

Kamel Bouzidi

► To cite this version:

Kamel Bouzidi. Traduction automatique de documents manuscrits et typographiés en arabe par couplage étroit entre systèmes. Sciences de l'Homme et Société. 2016. dumas-01494465

HAL Id: dumas-01494465

<https://dumas.ccsd.cnrs.fr/dumas-01494465>

Submitted on 23 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Traduction automatique de documents manuscrits et typographiés en arabe par couplage étroit entre systèmes

Nom : BOUZIDI

Prénom : Kamel

Sous la direction de :

Monsieur Laurent Besacier

Monsieur Benjamin Lecouteux

Monsieur Olivier Kraif

Laboratoire : Laboratoire Informatique de Grenoble (LIG)

UFR Langage, lettres et arts du spectacle, information et communication

Mémoire de master 2 recherche - 30 **crédits** - **Mention** Sciences du Langage

Spécialité: Industries de la langue - Parcours : TALEP

Année universitaire 2015-2016



Traduction automatique de documents manuscrits et typographiés en arabe par couplage étroit entre systèmes

Nom : BOUZIDI
Prénom : Kamel

Sous la direction de :

Monsieur Laurent Besacier
Monsieur Benjamin Lecouteux
Monsieur Olivier Kraif

Laboratoire : Laboratoire Informatique de Grenoble (LIG)

UFR Langage, lettres et arts du spectacle, information et communication

Mémoire de master 2 recherche - 30 **crédits** - **Mention** Sciences du Langage

Spécialité: Industries de la langue - Parcours : TALEP

Année universitaire 2015-2016

Remerciements

Je tiens d'abord à exprimer ma plus profonde gratitude à mes encadrants, monsieur Laurent Besacier, monsieur Benjamin Lecouteux et monsieur Olivier Kraïf pour m'avoir guidé et pour toute l'attention, les conseils et les aides qu'ils m'ont apportés durant la réalisation de ce mémoire.

Je remercie tous mes collègues de l'équipe GETALP surtout Zied Elloumi à m'intégrer dans l'équipe et à me donner ses précieux conseils. Ainsi que tous mes amis qui m'ont encouragé tout au long de ce travail.

Je tiens également à remercier tous mes enseignants pour leurs nombreuses aides et pour la qualité de l'enseignement qu'ils m'ont disposé durant ce Master, surtout monsieur Georges ANTONIADIS le responsable de notre master.

Finalement, j'adresse un grand merci à ma famille, et en particulier mon père Mohamed, ma mère Zohra et ma chère sœur Hedia qui m'ont encouragé tout au long de mes études.

Kamel

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : Bouzidi

PRENOM : Kamel

DATE : 30/08/2016

SIGNATURE :



Sommaire

Introduction générale.....	1
1. Contexte du stage	3
1.1. Contexte et objectifs du projet TRIDAN	3
1.2. Description du projet TRIDAN	4
1.3. Présentation du porteur et des partenaires	6
2. Etat de l'art.....	7
2.1. Traduction Automatique Statistique (TAS).....	7
2.2. La traduction automatique neuronale:	14
2.3. Reconnaissance optique des caractères.....	18
3. Traduction de chaîne de mots.....	24
3.1. Création d'un système de traduction arabe/français	24
3.2. Traduction de la meilleure hypothèse	32
3.3. Traitement des mots hors vocabulaire	34
3.4. Nouveau modèle de langue (Interpolation linéaire).....	36
4. Traduction des graphes.....	38
4.1. Treillis des mots.....	38
4.2. Chaîne de traduction des graphes	40
4.3. Traitement des mots hors vocabulaire	44
5. Bilan	46
Conclusion et perspectives	49
Conclusion	49
Perspectives	50

Introduction générale

Ce mémoire de recherche est l'aboutissement d'un stage, qui s'inscrit dans le cadre de notre formation de deuxième année du Master sciences du langage, spécialité industrie de la langue, de l'Université Grenoble Alpes.

L'organisme accueillant est le Laboratoire Informatique de Grenoble LIG¹ et plus précisément l'équipe GETALP² (Groupe d'étude sur le Traitement Automatique de la langue Parlée).

Le stage a été co-encadré par M. Laurent Besacier, professeur à l'Université Joseph Fourier, M. Benjamin LECOUTEUX, Maître de conférences à l'Université Pierre-Mendès-France et M. Olivier Kraif, maître de conférences à l'Université Stendhal comme tuteur de stage.

Notre sujet s'intègre dans le projet TRIDAN que nous allons détailler dans la première partie qui comprend ainsi une description des partenaires de ce projet et de notre organisme d'accueil.

Notre deuxième partie présente un état de l'art. Nous présenterons d'abord le domaine de la traduction automatique statistique ainsi que l'utilisation de réseaux de neurones dans ce domaine. Ensuite, nous aborderons brièvement le domaine de la reconnaissance automatique des caractères.

Dans la troisième partie nous allons présenter une description des données utilisées pour la création de notre premier système de traduction ainsi que les différents prétraitements appliqués. Ensuite nous allons détailler notre chaîne de création du système de traduction avant de faire la traduction des sorties d'un système de reconnaissance optique de caractères ainsi que l'évaluation des résultats. En outre, nous détaillerons notre méthode de traitement des mots hors vocabulaire. Enfin, nous décrirons la création d'une nouvelle version de modèle de langue en utilisant la méthode d'interpolation linéaire entre plusieurs modèles de langue ainsi que l'évaluation des résultats de traduction en utilisant ce modèle interpolé.

¹ www.liglab.fr

² www.liglab.fr/fr/presentation/equipes/getalp

Dans la quatrième partie nous aborderons la traduction des graphes en commençant par une présentation générale des graphes de mots et l'adaptation du système de traduction pour traduire ce type de donnée d'entrée avant de décrire la chaîne de conversion des sorties OCR en graphes. La fin de cette partie portera sur l'ajout d'un traitement spécifique des mots hors vocabulaire sur les graphes ainsi que l'évaluation de notre résultat final.

Dans la cinquième partie nous décrirons la progression des performances de notre système pour chaque étape du projet.

1. Contexte du stage

1.1. Contexte et objectifs du projet TRIDAN

Le projet TRIDAN se place dans le contexte de la recherche d'informations dans les documents numérisés multilingues. L'application visée concerne un utilisateur souhaitant réaliser une requête de recherche d'information dans sa langue sur des documents écrits dans une langue qu'il connaît peu ou pas du tout.

Le projet vise à mettre au point des techniques permettant un couplage innovant entre la lecture automatique de document numérisés (LAD), leur traduction automatique et l'extraction d'information.

Chacune de ces techniques prise en isolation a fait récemment l'objet d'avancées technologiques importantes, mesurées par des évaluations internationales et par leurs utilisations de plus en plus fréquentes dans des applications industrielles. Cependant, leur utilisation conjointe dans un système de traitement de l'information intégré permettant l'extraction d'information nécessite une articulation innovante de différentes étapes. La transcription automatique de documents numérisés permet déjà, dans une certaine mesure, de réaliser une extraction d'information du type entités nommées (au sens large). Cependant, les systèmes actuels se limitent aux documents dont la structure présente peu de variation et pour lesquels le type et le format des informations à extraire sont connus à l'avance : montants dans des factures à format connus, noms dans des tables de recensement. Les systèmes actuels ne permettent donc pas une extraction d'information dans des documents à structure inconnue et très variable.

D'autre part, même si les systèmes de transcription automatique permettent la reconnaissance de l'écriture manuscrite, l'extraction d'information manuscrite reste problématique : en effet, cette reconnaissance nécessite généralement l'utilisation d'un lexique qui ne peut contenir toutes les formes d'information à extraire (nom propre, date, code, montant, etc.). L'information recherchée est la plupart du temps "hors vocabulaire" et ne peut donc être ni reconnue ni extraite.

Enfin, les techniques de traduction automatique et d'extraction d'information sont généralement développées pour s'appliquer sur du texte électronique qui présente un faible niveau de bruit (fautes d'orthographe, fautes de frappe). L'application de ces techniques sur

du texte bruité, issu de la reconnaissance automatique d'écriture, nécessite la prise en compte des erreurs et des incertitudes dans la suite de la chaîne de traitement.

1.2. Description du projet TRIDAN

Le projet TRIDAN vise à optimiser une chaîne de traitement permettant à un utilisateur de faire des requêtes de recherche d'information dans sa langue (langue cible) sur des documents numérisés contenant des informations dans une langue qu'il connaît peu ou pas du tout (langue source).

Afin de pouvoir se fonder sur des bases de données déjà disponibles, ils ont été choisis de traiter l'arabe comme langue source et le français comme langue cible. Une fois la preuve de concept faite sur le couple arabe-français, l'adaptation des techniques développées dans le projet à d'autres couples de langue devrait se résumer à une collecte de nouvelles ressources, indépendamment des techniques utilisées.

La figure suivante représente une vue générale sur le projet TRIDAN :

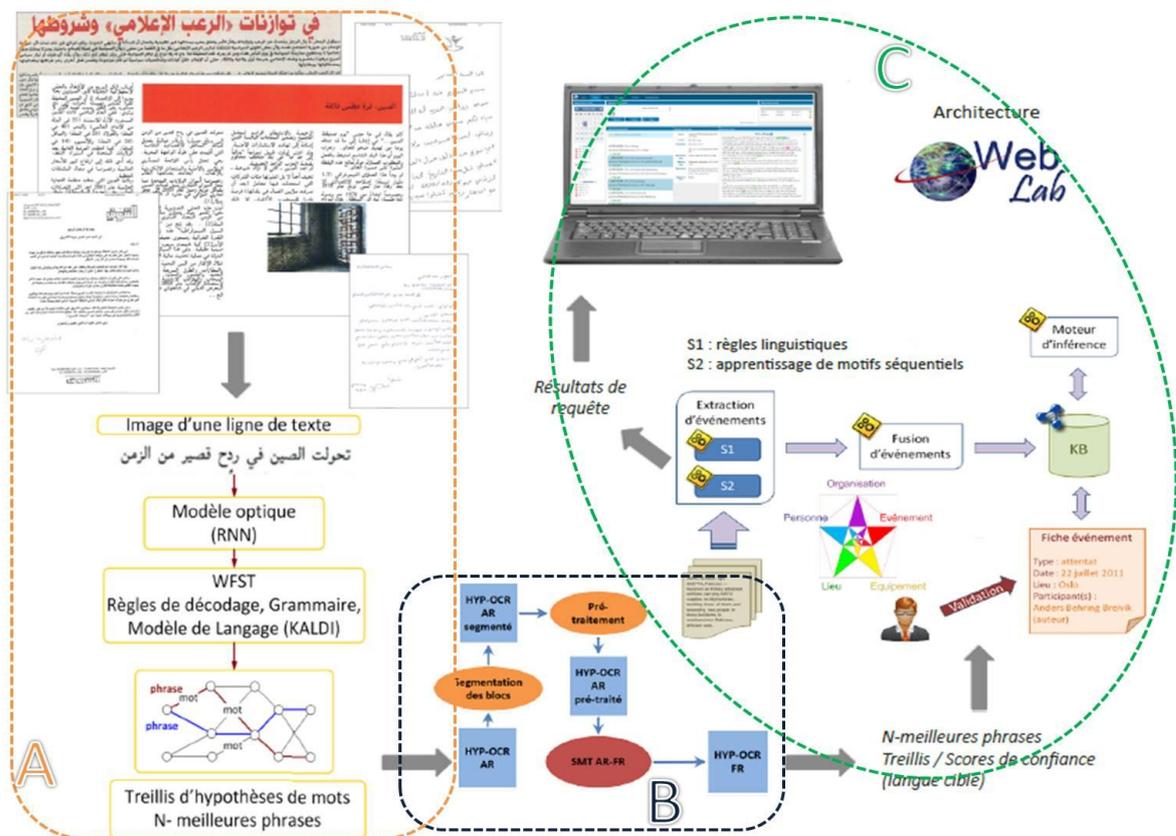


Figure 1 : projet TRIDAN, vue d'ensemble

Le projet décompose en trois sous-projets :

Sous-projet A Développement d'un système de reconnaissance d'écriture arabe mixte imprimé/manuscrite permettant une modélisation des zones de texte quel que soit le style d'écriture. Les trois principales problématiques adressées par ce sous-projet sont :

- le développement de méthodes d'analyse de documents robuste au bruit et permettant de détecter les zones de texte avec un rappel maximal. Contrairement aux approches standards qui vise une analyse fine des zones de texte (localisation, typage), nous voulons privilégier la détection des zones de texte les plus larges possible, quel que soit le type d'écriture, afin de garder les relations sémantiques entre les éléments textuels.
- la reconnaissance d'entités ou de mots hors vocabulaire, en se basant sur les séquences des caractères en parallèle d'une reconnaissance avec lexique et modèles de langue.
- l'extraction d'indices pertinents pour la détection d'entités nommées qui seront utilisés par le système d'extraction d'information.

Sous-projet B (notre rôle dans le projet) : Couplage innovant des modules transcription/traduction et traduction/recherche d'information.

- Création d'un système de traduction pour traduire les sorties OCR
- Adaptation du système de traduction automatique pour traduire des documents de type treillis (N-best).
- Nouvelle méthodes pour traiter les mots hors vocabulaire OOV.

Sous-projet C : Nouvelles méthodes d'extraction d'information à partir de texte bruité.

- adaptation faiblement supervisée des modèles d'extraction à des sorties bruitées de traduction.
- intégration des alternatives et des scores de confiance fournis par les modules de traduction pour l'extraction d'information et l'indexation.
- optimisation de l'interaction entre le module de traduction et ceux d'extraction et d'indexation.

1.3. Présentation du porteur et des partenaires

LIG (organisme d'accueil de notre stage): L'équipe GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) est née en 2007 lors de la création du Laboratoire d'Informatique de Grenoble (LIG). Issue de l'union vertueuse de chercheurs en traitement de l'écrit et de la parole, le GETALP est une équipe pluridisciplinaire dont l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale). La méthodologie de travail du GETALP s'appuie sur des allers-retours continus entre collectes de données, investigations fondamentales, développement de systèmes opérationnels, applications et évaluations expérimentales.

A2iA (Porteur) : fondée en 1991, A2iA S.A. est un éditeur de logiciels spécialisé dans la reconnaissance d'écriture manuscrite et imprimée. A2iA réalise aujourd'hui plus de 9 millions d'euros de chiffre d'affaire, dont la moitié à l'export. A2iA a développé des technologies de reconnaissance et de traitement de documents qui lui permettent d'occuper une position de leader mondial sur le domaine de la reconnaissance d'écriture manuscrite. Basé à Paris, le centre de R&D regroupe 30 ingénieurs et docteurs sur un total de 45 employés pour A2iA S.A.

Airbus Defence and Space : le département « Etudes amonts » de Airbus DS réalise des études amonts à caractère technique et/ou opérationnelles et des démonstrateurs appliqués pour des clients tels que le Ministère de la Défense, de l'Intérieur, la Commission Européenne ou l'Agence de Défense Européenne. Ce département rassemble de nombreuses compétences consacrées aux techniques informatiques de traitement, d'exploitation et de manipulation de l'information multi-sources, notamment des technologies WEB et analyse de données non structurées. Ce département a la charge des programmes d'études amont HERISSON sur l'évaluation des technologies de traitement de l'information non structurée, MAURDOR sur la reconnaissance de documents écrits (avec A2iA), TRAD sur la traduction automatique de texte et de la parole.

2. Etat de l'art

2.1. Traduction Automatique Statistique (TAS)

2.1.1. Principe de base

Les systèmes de TAS se basent sur un modèle mathématique de distribution et d'estimation probabiliste développée par les chercheurs d'IBM (Berger et al., 1994).

Un système de traduction statistique arrive à traduire des séquences des mots en se basant sur des exemples de traductions; ces exemples sont sous la forme des corpus parallèles, qui comportent un ensemble de paires de phrases qui sont des traductions les unes des autres.

Tout d'abord un processus d'alignement sera effectué sur ces données consistant à définir des correspondances entre chaque mot ou ensemble de mots en langue source et sa traduction en langue cible. Ces alignements font partie des données d'apprentissage. Une fois l'apprentissage fait, le système sera capable de traduire des nouvelles phrases qui ne sont pas parmi les exemples d'apprentissage en recombinaison des morceaux de ces exemples. (Gahbiche-Braham, 2013)

Un système de traduction automatique est basé sur trois composants :

- Un modèle de langage qui calcule $P(T)$.
- Un modèle de traduction qui calcule $P(S|T)$.
- Un décodeur qui prend une phrase S (dans la langue source) et produit la phrase la plus probable T (dans la langue cible).

Les deux modèles (de langage et de traduction) sont appris automatiquement en utilisant des corpus monolingues et bilingues.

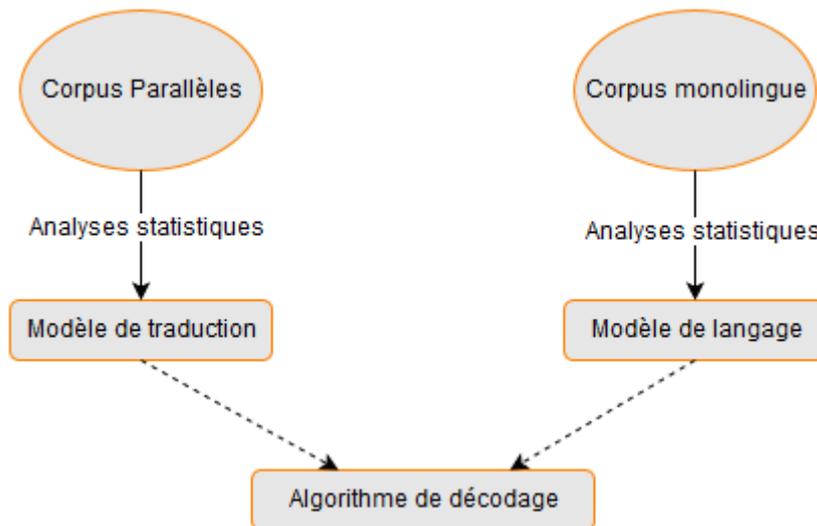


Figure 2 : Processus de la traduction automatique statistique

Le problème de traduction est reformulé par le théorème de bayes :

$$P(t|s) = \frac{P(s|t).P(t)}{P(s)}$$

Équation 1 : Théorème de bayes

Comme $P(s)$ est indépendante du texte cible t donc la traduction la plus probable t^* est obtenue en maximisant l'équation suivante

$$t^* = \operatorname{argmax}_t P(t|s) = \operatorname{argmax}_t P(s|t)P(t)$$

Équation 2 : Recherche de la traduction optimale

Généralement, plus la qualité de données augmente plus le système statistique devient performant.

2.1.2. Les modèles de traduction

Dans cette section on présente les deux modèles de traduction qui sont les modèles de traduction à base de mots et les modèles de traduction à base de segments.

2.1.2.1. Les modèles de traduction à base de mots

Les modèles de traduction à base des mots sont présentés par (Brown et al. 1990; Brown et al. 1993) et reposent sur le principe de la traduction mot par mot. L'unité de traduction pour ces modèles est le mot.

Parmi les problèmes de la traduction à base de mots il y a ce qu'on appelle le non déterminisme des appariements mot à mot, (ambiguïté lexicale) par exemple la traduction de mot « avocat » vers l'anglais peut être « avocado » ou « lawyer » selon le contexte.

Un autre problème est qu'un ensemble de mots peut être traduit par un seul mot et vice-versa. Les alignements consistent à faire correspondre les mots ou segments en langue source et leur traduction en langue cible, mais un mot en langue source peut s'aligner avec plusieurs mots dans la langue cible, ou à aucun mot.

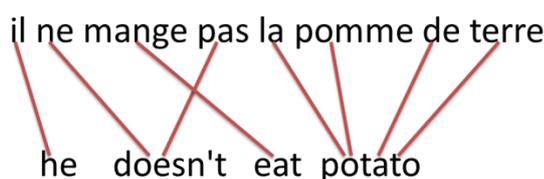


Figure 3 : Exemple d'alignement mot-à-mot entre une phrase en français et sa traduction en anglais

On note, sur la figure suivante que la phrase en arabe est de la forme Verbe-Sujet-Complément d'Objet par contre en français elle est de la forme Sujet-Verbe-Complément d'Objet donc l'ordre des mots n'est pas forcément le même pour une phrase et sa traduction, la tâche de réordonner les mots après la traduction est appelée le ré-ordonnement.

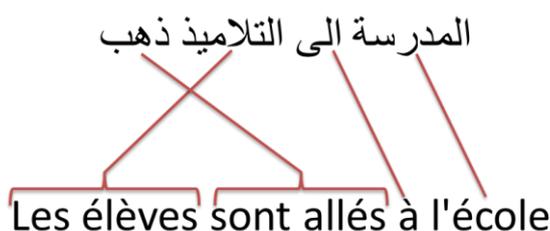


Figure 4 : Exemple d'alignement d'une phrase en arabe et sa traduction en français.

2.1.2.2. Les modèles de traduction à base de segments

Comme on a vu, les modèles à base de mots présentent quelques inconvénients tel qu'un mot peut être traduit par un ou plusieurs mots, donc le choix de mot comme unité de traduction n'est pas la meilleure solution, traduire un groupe des mots aide à la désambiguïsation et à l'amélioration de la traduction.

Le modèle (ou table) de traduction est construit à partir d'un corpus parallèle. Ce corpus bilingue doit être aligné à l'aide d'un outil d'alignement (Och et Ney, 2003; Gao et Vogel, 2008) pour avoir une traduction pour chaque segment de la phrase source. (Gahbiche-Braham, 2013)

L'alignement se fait dans les deux sens source-cible et cible-source, comme dans les modèles à base de mots les alignements utilisés sont mot-à-mot et mot-à-plusieurs en ajoutant des alignements plusieurs-à-plusieurs.

Une fois les alignements effectués, ils sont symétrisés afin de trouver les intersections et les unions de ces alignements. Cette étape de symétrisation représente l'alignement final qui sera utilisé pour l'apprentissage. (Gahbiche-Braham, 2013)

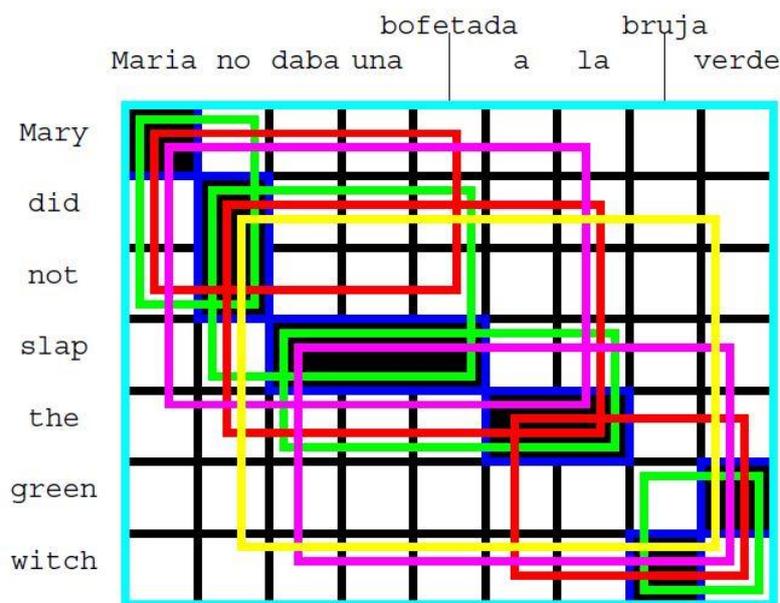


Figure 5 : Exemple d'alignement d'une phrase (Knight et Koehn, 2004).

Chaque segment en langue source peut avoir plusieurs hypothèses de traduction en langue cible. Et pour choisir le meilleur alignement, un score de probabilité est calculé pour chaque segment selon l'équation suivante :

$$P(t|s) = \frac{c(s, t)}{\sum_{t_i} c(s, t_i)}$$

Équation 3 : Probabilité des segments

Sachant que :

- **s** : le segment en langue source.
- **t** : le segment en langue cible.
- **c(s, t)** : le nombre de paires de segments dans lesquels apparaît un segment donné sur l'ensemble du corpus.

2.1.3. Le modèle de langue

Le modèle de langue est une composante essentielle du système TAS qui est en charge de la prise en compte des contraintes exigées par la syntaxe de la grammaire et le lexique de la langue cible permettant de donner un sens à la probabilité d'un mot dans son contexte et donc de trouver le mot le plus probable sachant ceux qui le précèdent.

Le but du modèle de langue est d'estimer la probabilité $P(t_1^I)$ d'une séquence de mots

$t_1^I = t_1 \dots t_i \dots t_l$. Cette probabilité est calculée par l'équation :

$$P(t_1^I) = P(t_1) \prod_{i=2}^I P(t_i | h_i)$$

Équation 4 : Probabilité d'une séquence de mots

Avec $h_i = t_1, \dots, t_{i-1}$ l'historique du mot t_i .

Les modèles de langue influent sur le choix des mots. Mais certaines fois un mot n'existe pas dans notre modèle de langue (mot hors vocabulaire) alors ce mot ne sera pas choisi par le système de traduction.

En général le modèle n-gramme est le plus utilisé avec **n** qui varie de 1 à 5. Un score est attribué à chaque séquence des mots est correspond à la probabilité d'apparition de cette

séquence dans un texte. Ce score représente la dépendance de chaque mot par rapport aux **n-1** mots qui le précèdent.

Fréquemment, on trouve des séquences de mots correctes mais elles n'existent pas dans notre modèle de langue donc une probabilité nulle est attribuée. C'est pour cela que des méthodes, appelées méthodes de lissage (smoothing) ont été développées afin de résoudre ce problème et parmi ces approches de lissage on cite le lissage Kneser-Ney (Kneser et Ney, 1995).

En pratique, les outils les plus utilisés pour la construction de modèles de langue sont SRILM (Stolcke, 2002) et IRSTLM (Federico et al. 2007).

2.1.4. Le décodage

Après que le modèle de langue et le modèle de traduction sont construits, ces deux modèles seront utilisés par le décodeur pour trouver pour chaque phrase source une phrase cible qui est la meilleure hypothèse de traduction.

Le décodage peut être considéré comme le processus le plus important et le plus compliqué dans un système de traduction, vu qu'il s'agit de sélectionner la meilleure hypothèse de traduction qui assure le meilleur transfert du sens à partir d'un grand nombre de possibilités de traduction.

Pour générer le document cible ayant la plus grande probabilité, le décodeur utilise la fonction de densité fournie par le modèle. Cette tâche peut être accomplie en résolvant l'équation suivante :

$$\hat{e} = \underset{e}{\operatorname{argmax}} P(e^I | f^J)$$

Équation 5 : Fonction de densité utilisée par le décodeur

Où f^J est le document source et e^I est l'ensemble des documents de la langue cible.

2.1.4.1. *Moses*

Moses est un outil très performant de traduction automatique statistique à base de segment, il permet d'implémenter automatiquement un système de traduction automatique pour n'importe quelle paire de langues. C'est un outil libre et gratuitement disponible sur le web et toujours en cours de développement.

Cette boîte à outils est constituée principalement de deux composants:

- **Processus d'entraînement:** permet de construire un modèle de traduction à partir des données parallèles.
- **Décodeur :** une application C++ permettant de déterminer la traduction la plus probable d'une phrase source selon le modèle de traduction.

MOSES est conçu en plusieurs modules :

- **Train-MOSES** permet de préparer les données d'entraînement et de réaliser l'apprentissage.
- **MERT-MOSES** permet d'ajuster les poids des différents modèles afin d'optimiser la performance de la traduction lors de ce qu'on appelle la phase de développement.
- **MOSES-cmd** contient les outils et l'exécutable de décodage des systèmes de type *Phrase-Based*.
- **MOSES_Chart-cmd** contient les outils et l'exécutable de décodage des systèmes de type *Hierarchical-Phrase-Based*.

2.1.5. Évaluation de la traduction automatique

Un texte généré par un processus de traduction automatique doit être évalué. Il existe deux types d'évaluation : évaluation automatique et évaluation manuelle.

2.1.5.1. *Evaluation manuelle*

L'évaluation manuelle (ou subjective) est effectuée par l'intervention de l'être humain, qui juge de la qualité de traduction en fonction de critères précis telle que la fluidité, la fidélité du texte traduit. Tous ces travaux demandent plus de travail manuel et beaucoup de temps.

Ce type d'évaluation donne la mesure la plus exacte des performances de système, mais il nécessite l'intervention d'experts bilingues.

2.1.5.2. *Evaluation automatique*

L'évaluation automatique est effectuée à l'aide d'une métrique automatique, elle nécessite une ou plusieurs traductions de référence (réalisée par des humains). Ceci permet de faire une comparaison entre la sortie de traduction automatique et une traduction référence et déterminer le degré de ressemblance.

Le score BLEU (Bilingual Evaluation Understudy) proposée par (Papineni et al. 2002) est l'unité de mesure la plus utilisée qui se calcule en comparant la sortie de traduction automatique à une ou plusieurs références. En effet le principe est basé sur une comparaison de courtes séquences de mots (n-grammes) pour chaque phrase. Le score BLEU varie de 0 à 1, ou il peut aussi être exprimé généralement en pourcentage.

Le score OOV (Out-Of-Vocabulary) : ce score représente le pourcentage des mots non-traduits dans la sortie d'un système de traduction, et dans la plupart des systèmes de traduction ces mots gardent leur forme initiale ou sont remplacés par des étiquettes spéciales (exemple : <unk>). Le score OOV dépend généralement de la quantité de données utilisées dans l'entraînement de la table de traduction.

2.2. La traduction automatique neuronale

Les réseaux de neurones artificiels ont retrouvé une place centrale dans le paysage de l'apprentissage automatique et ont récemment permis des avancées significatives pour de nombreux domaines applicatifs. Pour ce qui concerne le traitement automatique des langues (TAL), les applications sont également nombreuses et variées. L'importance des modèles neuronaux dans le domaine du traitement automatique des langues ne cesse de croître

2.2.1. Réseaux des neurones

Les réseaux de neurones sont basés sur un modèle simplifié de neurone. Ce modèle donne la possibilité d'effectuer certaines fonctions du cerveau, comme la mémorisation associative, l'apprentissage par l'exemple, le travail en parallèle, mais les réseaux biologiques restent plus compliqués par rapport aux modèles mathématiques et informatiques.

Le réseau de neurones est généralement constitué d'un ensemble "d'unités"(ou neurones). Chacune de ces unités possède une petite mémoire locale. Ces neurones sont connectés par des canaux de communication (les connexions ou synapse), qui servent à transporter les données numériques. Les unités ne peuvent agir que sur les entrées transmises par leurs connexions et sur leurs données locales.

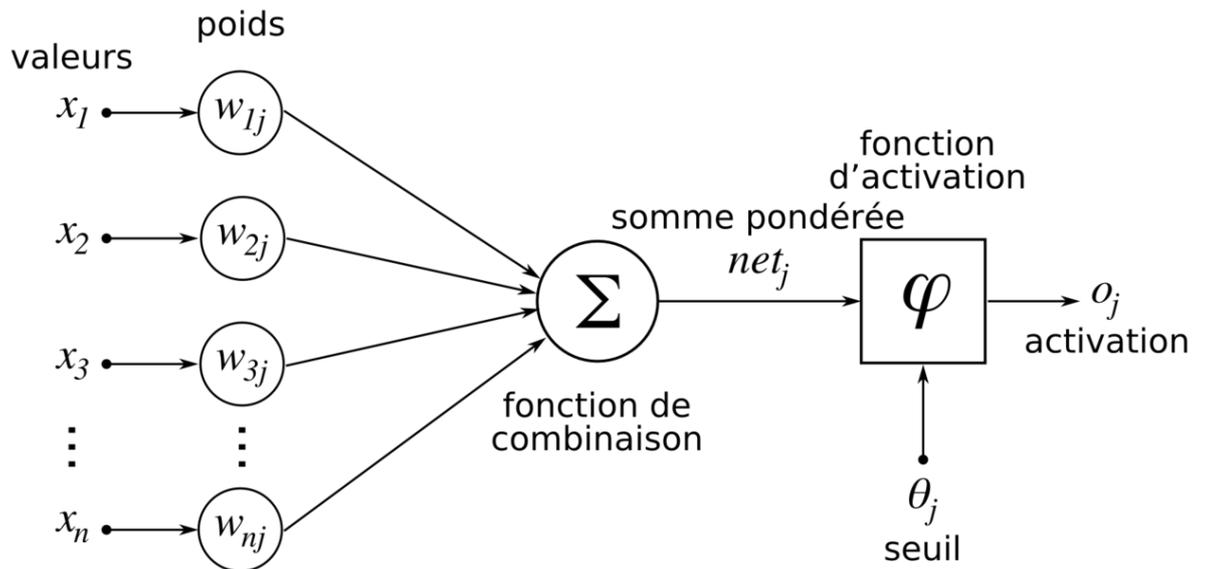


Figure 6 : Structure d'un neurone (Wikipédia)

Les réseaux de neurones sont caractérisés par leurs capacités d'apprentissage. Cela signifie que, comme les enfants apprennent à distinguer entre les chats et les chiens à travers des exemples de chats et de chiens, les réseaux de neurones apprennent également à partir des exemples.

2.2.2. Modélisation d'un réseau de neurone

Globalement, un réseau de neurones possède la même structure que chacun de ses neurones. Il doit pouvoir calculer des valeurs de sorties ($a_1, a_2, a_3, \dots, a_n$) en fonction de des entrées ($x_1, x_2, x_3, \dots, x_n$). Un premier ensemble de neurones applique aux entrées leur propre fonction d'activation, ce qui donne un certains nombres de résultats. Un second ensemble de neurones prend ces résultats en entrée et calculent de nouveau, avec leur propre fonction d'activation, des résultats qu'ils transmettent à l'ensemble de neurones suivant, etc., jusqu'à atteindre le dernier ensemble de neurones : les sorties de ces derniers neurones sont alors considérées comme les sorties du réseau.

Un réseau de neurones peut donc être représenté par les poids w des différents neurones. Ces poids peuvent varier au cours du temps, en fonction des entrées présentées X .

2.2.3. Modèles de langage neuronaux

Le modèle neuronal standard, tel qu'il est décrit dans l'article (Bengio et al., 2003) introduit l'usage des réseaux neuronaux pour la modélisation du langage. L'idée principale est d'associer (ou de projeter) chaque mot à un vecteur w de dimension $|V|$, où $|V|$ est la taille du vocabulaire ; w contient une seule valeur non-nulle (typiquement égale à 1) dans un espace continu de dimension m par multiplication avec la matrice R de dimension $|V| \times m$, puis d'utiliser le réseau neuronal pour apprendre la distribution comme une fonction des vecteurs.

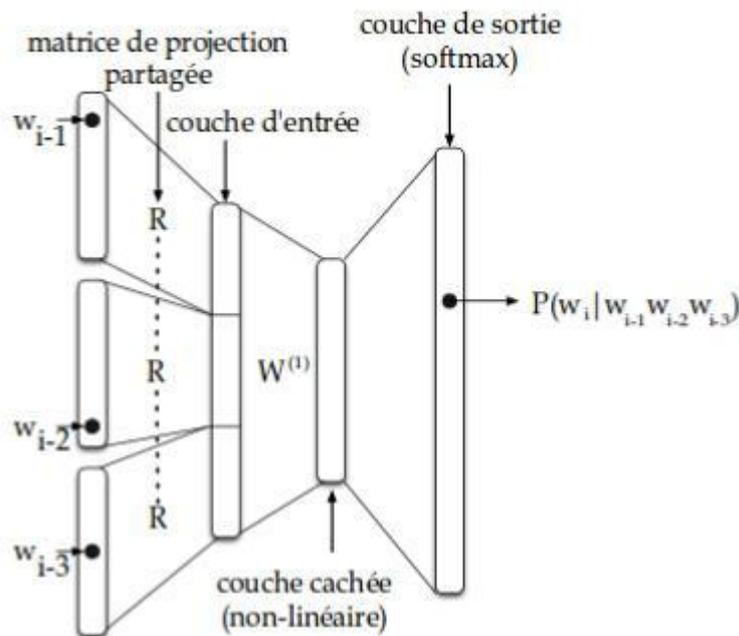


Figure 7 : Modèles de langage neuronale

2.2.4. Modèle neuronal pour la Traduction automatique (Séquence vers séquence) :

Une approche actuelle basé sur un Encodeur et un décodeur introduit par Bengio dans son article (Bengio et al. 2014, Learning phrase representations using Recurrent Neural Network (RNN) encoder-decoder for statistical machine translation) ou Encodeur/décodeur sont des réseaux de neurones récurrents (bidirectionnels) composés par des unités à portes

(Long Short Term Memory LSTM ou Gated Recurrent Unit GRU) qui permette d'apprendre à mémoriser/oublier.

La séquence d'entrée est encodée dans un vecteur de faible dimension (quelques centaines) et à partir de cette séquence, la séquence de sortie est générée. En effet, l'encodeur permet de compresser l'entrée dans un vecteur de taille fixe, et le décodeur génère une séquence à partir de la représentation compressée.

La figure ci-dessous représente une architecture neuronale permettant la mise en œuvre d'un tel modèle. Le réseau comporte un encodeur et un décodeur:

1. Encodage des mots dans un vecteur 1-hot (w_i) : (encodage one-hot consiste à représenter les valeurs par une représentation binaire).
2. Projection du vecteur 1-hot dans l'espace continu : embedding (c_i)
3. Mise-à-jour incrémentale de l'état caché de l'unité (h_i) récurrente de l'encodeur
4. On obtient une représentation de la phrase
5. Mise-à-jour incrémentale de l'état caché de l'unité (z_i) récurrente du décodeur
6. Calcul de la distribution de probabilité (p_i) pour tous les mots suivants
7. Détermination du mot suivant (le plus probable)

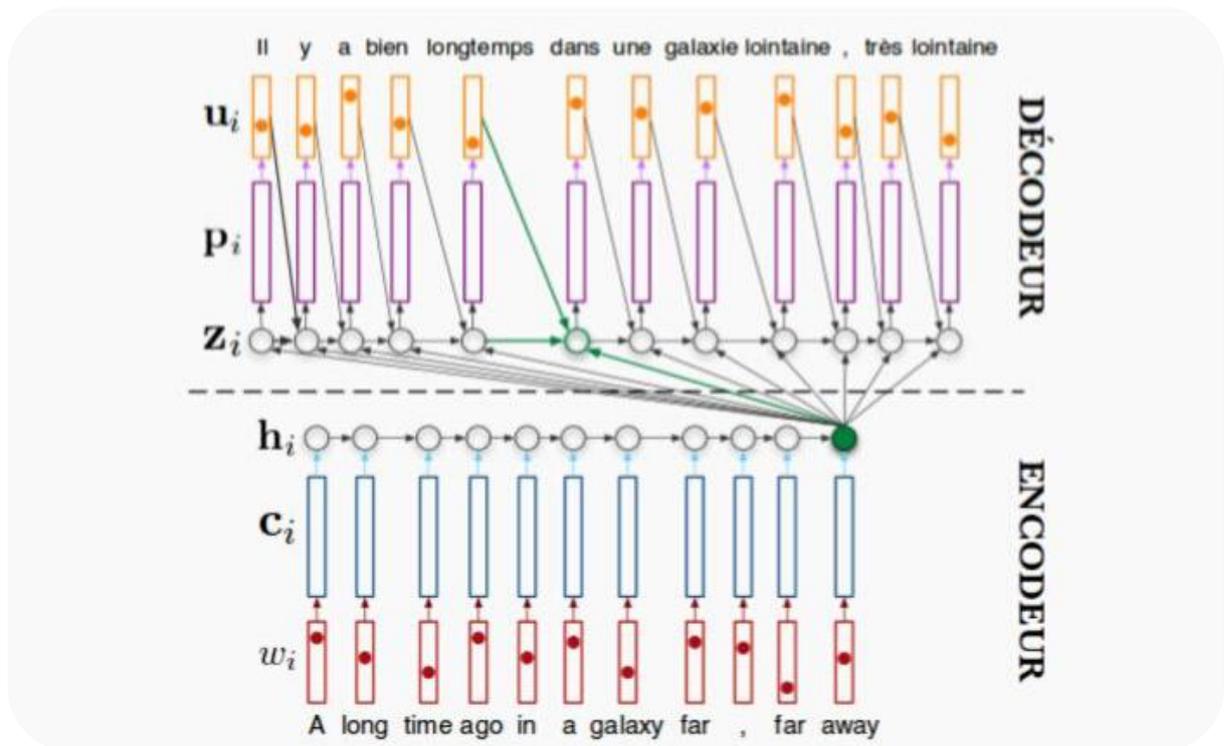


Figure 8 : exemple d'architecture neuronale pour la TA

2.3. Reconnaissance optique des caractères

2.3.1. Introduction

La reconnaissance optique de caractères OCR (Optical Character Recognition) est une tâche qui consiste à réaliser une transformation en représentation symbolique d'un texte sur un support papier en un texte sous format d'un fichier informatique.

La reconnaissance de l'écriture manuscrite est plus compliquée et générale que l'écriture typographique et cette partie insiste plus sur la reconnaissance de l'écriture manuscrite.

2.3.2. Production et reconnaissance

La reconnaissance de documents est le fait d'obtenir la forme logique à partir de la forme papier.

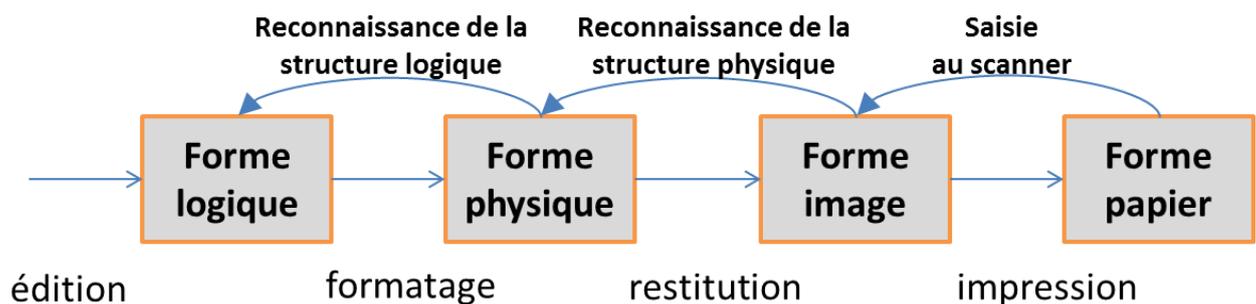


Figure 9 : Processus de production et de reconnaissance de documents

Les étapes de la reconnaissance des documents:

Le document papier est acquis à l'aide d'un scanner de manière à obtenir une image sous format informatique. Une image bruitée et biaisée appelée image brute. Cette image est ensuite prétraitée on obtient alors une image épurée avec une représentation plus claire.

La reconnaissance de la structure physique se fait en deux étapes la détection des zones de l'image et le découpage. En outre la détection consiste à classifier les différentes zones de l'image en graphique, texte, table, formule, dessin ou bien photo par contre le

découpage sert à ranger les zones du texte en colonnes, lignes, mots et signes (Robadey, 2001).

La reconnaissance de la structure logique a pour but d'étiqueter logiquement les différents objets de la structure physique et aussi de réorganiser ces objets identiquement au flux de lecture (Robadey, 2001).

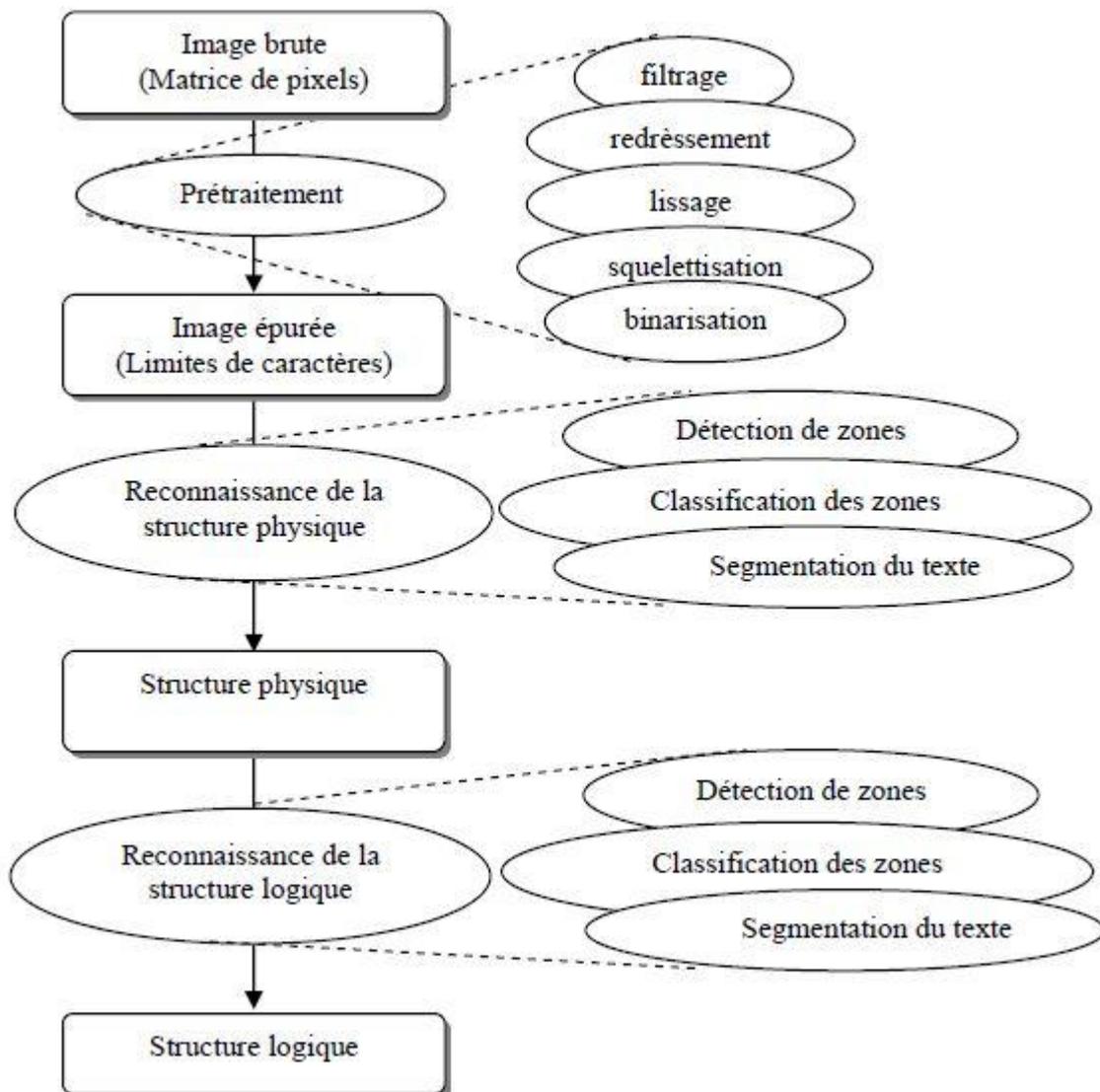


Figure 10 : Etapes de la reconnaissance de documents (Robadey, 2001)

2.3.3. Caractéristique de l'OCR

La création d'un système OCR est plus ou moins difficile car il dépend de l'application à implémenter et des données à traiter.

Le système OCR est composé de 3 outils :

- Outil d'acquisition : dont la reconnaissance dépend essentiellement du type de données à acquérir, on distingue deux types d'acquisition une est dite en ligne et l'autre est dite hors ligne.
- Approches de reconnaissance : il existe deux approches :
 - Globale qui a pour rôle de reconnaître la représentation intégrale de mots et le décrire indépendamment des caractères qui le constituent.
 - Analytique qui a pour rôle de segmenter l'image des mots en entrée en caractères ou en fragment morphologique.
- La nature des traits (caractéristiques) change d'une approche à une autre, on trouve généralement 5 groupes de caractéristiques :
 - Caractéristiques topologiques.
 - Caractéristiques structurelles.
 - Caractéristiques statistiques.
 - Globales ou locales.
 - Superposition des modèles et corrélation.

2.3.4. Problèmes liés à l'OCR

Le système OCR devient compliqué à cause de :

Disposition spatiale du texte :

La classification de Tappert (Belaïd, A.) indique que la présentation du texte peut subir deux types de contraintes : *externes* conduisant à une écriture *pré casée, zonée, guidée* ou *générale*; et *internes* provenant des habitudes propres à chaque scripteur et conduisant à une écriture *détachée, groupée, script* (bâton), purement *cursive* ou *mixte*. Il est évident que l'écriture détachée reste la plus facile à réaliser du fait de la séparation quasi immédiate des lettres. Au contraire, l'écriture cursive nécessite plus d'efforts du fait de l'ambiguïté des limites entre les lettres (Belaïd, A.).

Nombre de scripteurs :

La difficulté de reconnaissance dépend de nombre de scripteurs et augmente avec ce nombre. On trouve trois types de scripteur : mono, multi et omni.

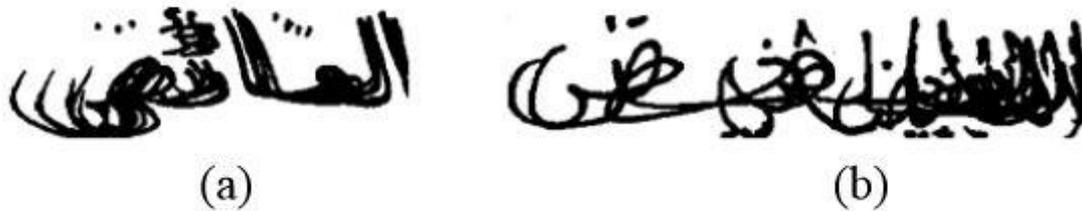


Figure 11 : Scripteur

Dans la figure 11, nous considérons la superposition d'un mot écrit six fois par le même scripteur (figure 11a) et celle produite par six scripteurs différents (figure 11b)

Taille du vocabulaire :

On a deux types d'applications, celles à vocabulaire limité (<100 mots) et celles à vocabulaire très étendu (>10 000 mots).

Pour le premier type (limité) la complexité est moindre car la réduction du nombre limite l'encombrement mémoire et la complexité la reconnaissance.

2.3.5. Organisation générale d'un système de reconnaissance

Le système de reconnaissance de l'écrite manuscrite se base essentiellement sur le fait d'identifier de façon exacte l'entrée d'une image d'un texte sur papier ou bien photographié (Al-Rashaidh, 2006) et de la convertir en un texte sous le format informatique, par exemple le format HTML ou bien LATEX.

Alors pour ce faire on fait appel aux étapes suivantes (Al-Rashaidh, 2006) :

- Acquisition (Scanning, Numérisation),
- Prétraitement,
- Segmentation en caractères séparés ou segments reliés à un caractère,
- Extraction des caractéristiques,
- Classification, suivie éventuellement d'une phase de post-traitement.

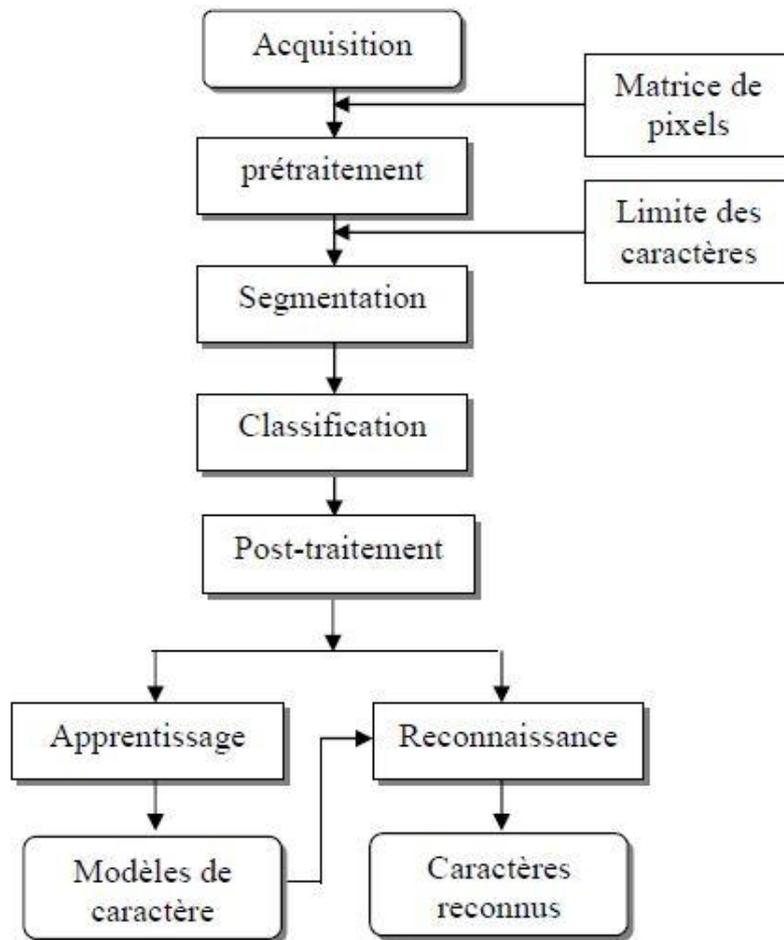


Figure 12 : Schéma général d'un système de reconnaissance de caractères (Zaiz, 2010)

2.3.6. Caractéristiques de l'écriture arabe

Le système d'écriture de la langue arabe est standard dans 27 pays arabe, dans la langue arabe on remarque l'absence des voyelles alors que ces dernières sont marquées par ce qu'on appelle diacritiques. La reconnaissance de texte arabe reste un défi vu qu'il n'y a pas beaucoup de recherches sur ce domaine.

L'écriture arabe possède les caractéristiques suivantes (Gheithet-Anssari, 2009):

- Par nature, le script arabe est cursif
- Le texte arabe s'écrit de droite à gauche
- L'alphabet arabe comporte 28 lettres de base

- La plupart des lettres arabes sont composées de deux composantes, l'une est principale et l'autre est secondaire (on ne parle pas des diacritiques) : par exemple le caractère Noun (ن) se compose d'un corps et un point au-dessus, le lette Teh (ت) a le même principe mais avec deux points et la lettre Kef (ك) qui possède au milieu de son corps un petit zigzag (s).

3. Traduction de chaîne de mots

Dans cette partie nous allons présenter les étapes de création de notre premier système de traduction ainsi que les données d'apprentissages et les différents prétraitements appliqués. Ensuite nous allons détailler notre chaîne de traduction des sorties d'un système de OCR ainsi que l'évaluation des résultats. En outre, nous décrirons notre méthode de traitement des mots hors vocabulaire. Enfin, nous utilisons une nouvelle version de modèle de langue en utilisant la méthode d'interpolation linéaire entre plusieurs modèles.

3.1. Création d'un système de traduction arabe/français

Dans cette section, nous allons commencer par la description des corpus utilisés pour la création de notre système de traduction, par la suite nous abordons les différents prétraitements appliqués sur ces corpus. De plus, nous décrivons la chaîne de création de notre système et son évaluation.

3.1.1. Description des corpus

Europarl :

Le corpus Europarl³ est un corpus parallèle de transcriptions de séances parlementaires correspondant à la période 1996-2011 et il comprend des versions en 21 langues européennes.

L'extraction et le traitement de ce corpus a pour but de générer des textes alignés pour les systèmes de traduction automatique statistique. Nous allons utiliser seulement la partie française de la version 7 pour la création de notre modèle de langue.

Corpus	Nombre de phrase	Nombre de mots
Français	2 007 723	52 525 000

Tableau 1 Corpus Europarl

MultiUN :

Le corpus parallèle MultiUN⁴ est extrait du site Web des Nations Unies, après être nettoyé et converti en XML par le Laboratoire de langage et technologie DFKI GmbH (LT-DFKI)⁵, Allemagne. Les documents ont été publiés par l'ONU de 2000 à 2009.

³ www.statmt.org/europarl

⁴ www.euromatrixplus.net/multi-un/

⁵ www.dfki.de/lt/publication_show.php?id=4790

Corpus	Nombre de phrase	Nombre de mots
Arabe	9 929 567	222 387 310
Français	9 929 567	285 520 384

Tableau 2 Copus MultiUN

News Commentary :

Le corpus News Commentary est un corpus parallèle aligné au niveau des phrases.

Ce corpus contient des extraits de diverses publications de presse et de commentaires du projet Syndicate⁶, il existe en 5 langues (anglaises, françaises, espagnoles, allemandes et tchèques).

Corpus	Nombre de phrase	Nombre de mots
Arabe	90 753	2 180 814
Français	90 753	2 372 649

Tableau 3 Corpus News-Commentary

Opensub :

Le corpus opensub est extrait du site Web opensubtitles⁷ il s'agit des sous-titres des films multilingues.

Corpus	Nombre de phrase	Nombre de mots
Arabe	4 381 835	27 739 977
Français	4 381 835	32 269 908

Tableau 4 Corpus Opensub

Trame :

Le corpus parallèle Trame correspond à environ 90 heures des discours radio et télévisés arabe enregistrés, transcrits et ensuite traduits en français.

Corpus	Nombre de phrase	Nombre de mots
Arabe	20 539	546 257
Français	20 539	758 030

Tableau 5 Corpus Trame

⁶ www.project-syndicate.org/

⁷ <http://opensubtitles.org>

Wit3 :

Le corpus parallèle WIT3 (Web Inventory of Transcribed and Translated Talks) est une collection de séminaires transcrits et traduits. Le noyau de ce corpus est le corpus TED (Talks), redistribué essentiellement par le site web conférence TED⁸.

Corpus	Nombre de phrase	Nombre de mots
Arabe	87 732	1 946 275
Français	87 732	2 436 720

Tableau 6 Corpus Wit3

Corpus de développement et test TRIDAN :

Notre corpus Dev et Test se compose de 3 types de documents des journaux (Corpus C2), des courriers typographiés (Corpus C3) et des courriers manuscrits (Corpus C4). Ces corpus sont en arabe et nous avons pour chaque corpus 2 références françaises traduites par des humains.

Corpus	DEV	T EST
C2 journaux	250	267
C3 Courrier	155	146
C4 Courrier	390	350

Tableau 7 nombre de lignes Copus dev/test

3.1.2. Prétraitement des corpus

Avant la création de notre système de traduction nous avons appliqué certains prétraitements sur notre corpus d'entraînement.

Normalisation des diacritiques

En arabe les voyelles sont différentes par rapport aux autres langues, en effet les signes de voyellisation sont des signes diacritiques placés au-dessous ou au-dessus des caractères.

Les voyelles en arabe ne sont pas obligatoires, elles sont utilisées pour faciliter la lecture ou pour enlever l'ambiguïté des textes, c'est pour cette raison que les textes juridiques et les ouvrages pédagogiques sont entièrement voyellés.

⁸ <http://www.ted.com>

Les diacritiques obligatoires :

En arabe on a des diacritiques obligatoires pour distinguer des lettres qui ont des formes très proches, ces diacritiques sont les points, deux points et trois points et aussi la position des points.

→		
ح	ح	ح
H	KH	J

Tableau 8 Exemple de diacritique obligatoire (position des points)

ٲ		
ن	ن	ن
N	T	ٲ

Tableau 9 Exemple de diacritique obligatoire (nombre de points)

A partir de la transcription ci-dessus on peut constater que le nombre de points et leurs positions sont importants dans certaines lettres arabes.

Les diacritiques facultatifs (ou les diacritiques de désambiguïsation) :

Pour les diacritiques facultatifs on peut les classer en 3 groupes :

Diacritiques simples :

Diacritique	Prononciation	Unicode Hex
ا	[a]	u064e
ا	[u]	u064f
ا	[i]	u0650
ا	[o]	u0652

Tableau 10 diacritiques simples

Diacritiques doubles :

Diacritique	Prononciation	Unicode Hex
اِ	[an]	u064b
اَ	[on]	u064c
اِ	[in]	u064d

Tableau 11 Diacritiques doubles

Le diacritique « chadda » :

Diacritique	Prononciation	Unicode Hex
آَ	[CD]	u0651

Tableau 12 Diacritique chadda

Le diacritique « chadda » a comme effet le doublement de la lettre à laquelle il est associé.

Suppression des diacritiques :

Bien que les diacritiques soient destinés à lever les ambiguïtés lors d'un traitement automatique, la majorité des systèmes de traitement automatique de langue comme Google Traduction, Buckwalter, MADA ou Xerox ne traitent pas les textes voyellés à cause du manque de ressources arabes voyellées. Donc, même si l'entrée est voyellée, ces systèmes commencent par la suppression de tous les diacritiques facultatifs, et ce sera le cas pour notre système aussi.

Suppression de Tatweel :

Le **Tatweel** (-) est un phonème particulier de la langue arabe. Il n'appartient ni à l'alphabet arabe ni au diacritique arabe, il est utilisé juste pour l'esthétique dans les textes arabes.

Sens	Mot normal	Mot avec Tatweel
Un livre	كتاب	كتــــــــــــاب
Un crayon	قلم	قــــــــــــلم

Tableau 13 Exemple d'utilisation de Tatweel

Tant qu'il ne change pas le sens de mots où il est utilisé, nous avons éliminé tous les caractères « tatweel ».

Normalisation de caractère (Hamza) :

Le "hamza" (ء) est un caractère un peu particulier de la langue arabe. Il est considéré comme une lettre ou un diacritique, et il a plusieurs règles d'écriture selon sa place dans le mot. Cette lettre peut disparaître dans certains cas comme quand il est sous forme dite « instable », par exemple quand il est précédé d'un autre mot. Dans ce cas nous allons normaliser cette lettre.

On peut dire qu'en arabe le rôle du "hamza" se limite à la prosodie (comme une ponctuation) il indique un arrêt ou une pause (aucun son). Ainsi, nous avons eu l'idée de normaliser cette lettre/diacritique en rendant chaque $\bar{ا}$ ou $ا$ ou $\bar{ا}$ en $ا$.

Caractère	Unicode Hex
$\bar{ا}$	u0622
$ا$	u0623
$ا$	u0625
$ا$	u0627

Tableau 14 les différents "hamza"

Autre prétraitement :

- Correction des caractères mal-encodés.
- Tokenisation côté arabe avec l'outil Opennlp⁹
- Tokenisation côté français avec le script *tokenizer.pl*¹⁰ qui fait partie de la boîte à outils Moses.
- Normalisation de la ponctuation et des caractères spéciaux (points, virgules, guillemets, apostrophes, tirets, crochets, etc.) en utilisant le script Perl *normalize-punctuation.perl*¹¹ qui fait partie de la boîte à outils Moses.
- Conversion de toutes les données en UTF-8.
- Transformation de toutes les données en minuscule *lowercase.perl*¹² qui fait partie de la boîte à outils Moses.
- Suppression des phrases plus longues que 100 mots.

⁹ opennlp.apache.org

¹⁰ www.statmt.org/wmt08/scripts.tgz (scripts/tokenizer.perl)

¹¹ www.statmt.org/wmt11/normalize-punctuation.perl

¹² www.statmt.org/wmt08/scripts.tgz (scripts/lowercase.perl)

3.1.3. Création de notre système

Nous commençons cette partie par la figure 13 qui représente les différentes étapes de création de notre système de traduction :

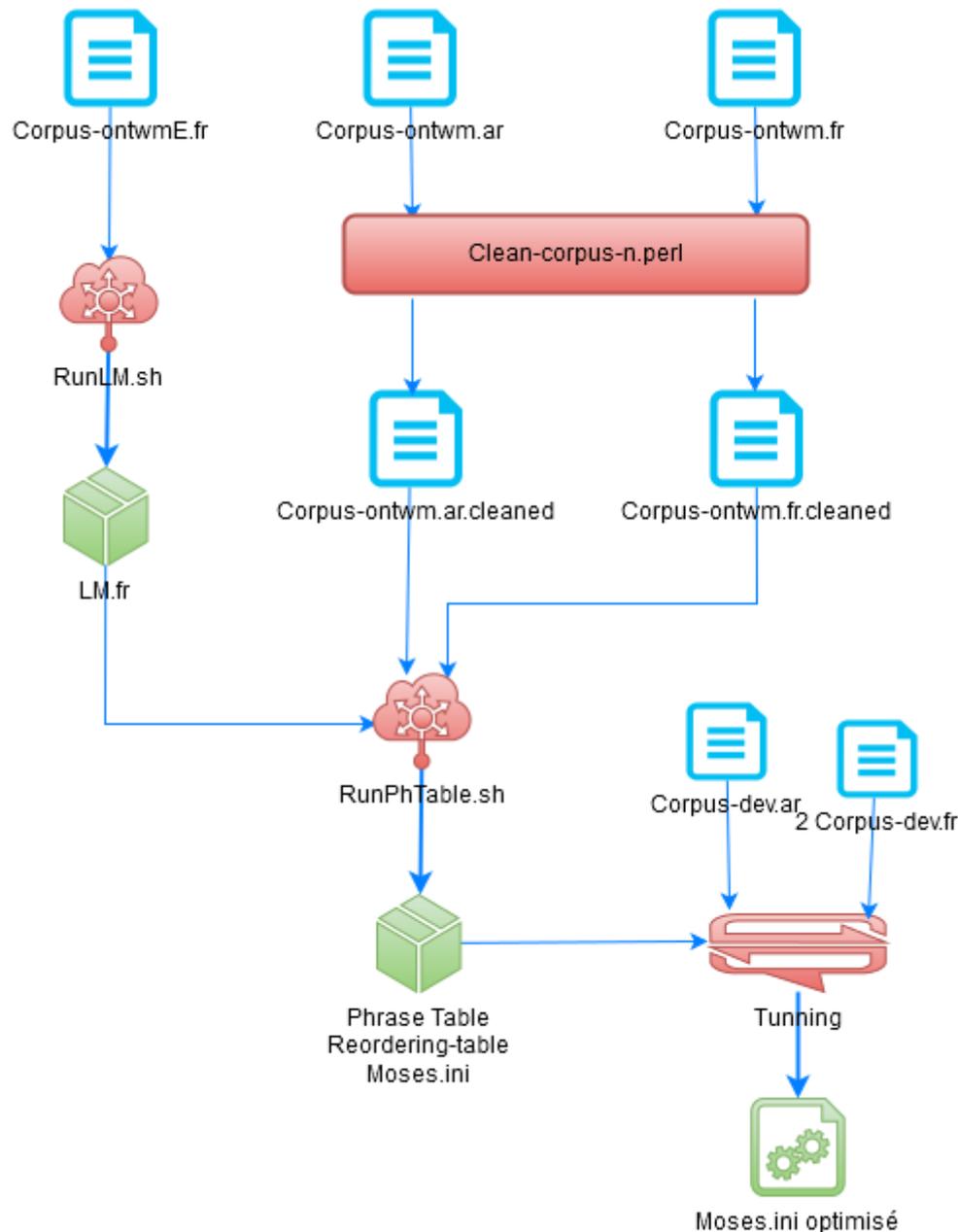


Figure 13 : Processus de création de notre système de traduction

1^{ère} étape :

Nous avons concaténé les corpus prétraités Opensub, News, Trame, Wit3, et Multi-UN pour obtenir un corpus parallèle (Corpus-ontwm.Fr/Ar) de 13,5 M lignes, et pour le corpus monolingue qu'il sera utilisé pour la création du modèle de langage nous avons ajouté aussi la partie française de corpus Europarl pour obtenir le corpus (Corpus-ontwmE.fr) de 16,5 M lignes.

2^{ème} étape :

Nous avons créé un modèle de langage 5-gramme en utilisant le script RunLm.sh qui utilise l'outil IRSTLM (M. Federico et al 2008).

```
lang="fr"
input=$1
echo "etape 1/7"
grep -v /^\W\+$/ $input | sort | uniq | $IRSTLM/bin/add-start-end.sh > $1.$lang.lm-train
echo "etape 2/7"
$IRSTLM/bin/build-lm.sh -t=/tmp -i $1.$lang.lm-train -o $1.$lang.iarpa.lm.gz -k 3 -n 5 -s improved-kneser-ney
echo "etape 3/7"
$IRSTLM/bin/compile-lm $1.$lang.iarpa.lm.gz /dev/stdout --text yes | gzip -c > $1.$lang.arpa.lm.gz
echo "etape 4/7"
zcat $1.$lang.arpa.lm.gz > $1.$lang.arpa.lm
echo "etape 5/7"
$IRSTLM/bin/sort-lm.pl -inv -ilm $1.$lang.arpa.lm -olm $1.$lang.arpa.inv.lm
echo "etape 6/7"
gzip $1.$lang.arpa.inv.lm
echo "etape 7/7"
$SMT_HOME/Tools/Moses-Decoder/bin/build_binary -i $1.$lang.arpa.inv.lm.gz $1.$lang.kenlm
```

Figure 14 : script RunLm.sh

3^{ème} étape

L'apprentissage de notre modèle de traduction et modèle d'alignement, nous avons utilisé pour cette tâche le script RunPhTable.sh

```
$MOSES/train-model.perl -external-bin-dir /SMT-Engine/Tools/Giza/ -root-dir ./ -
corpus CORPUS -f ar -e fr -alignment grow-diag-final-and -reordering msd-
bidirectional-fe -lm 0:5: $LM:8 -mgiza -mgiza-cpus 8 -parallel -cores 16 --first-step 1 --
last-step 9
```

Figure 15 : script RunPhTable.sh

4^{ème} étape :

Après avoir créé notre système de traduction, nous passons à l'étape d'ajustement des poids en utilisant le programme Minimum Error Rate Training (MERT) (Och and Ney, 2003) qui permet d'ajuster les poids du modèle de modèle de langage, du modèle de traduction et du modèle de distorsion à l'aide de notre corpus de développement arabe avec deux références français.

Evaluation

Les premiers résultats sur les corpus de référence OCR TRIDAN sont donnés dans le tableau suivant, ici l'OCR est supposée parfait (sans erreurs).

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	29.00	21.22	21.58
Test	26.95	20.31	19.90

Tableau 15 Evaluation de notre premier système en score BLEU

3.2. Traduction de la meilleure hypothèse d'OCR

Nous décrivons dans cette section la chaîne de traduction des sorties OCR envoyées par A2IA, ici nous allons devoir traduire une entrée bruitée contenant des erreurs d'OCR. La figure ci-dessous représente les différents composants de notre chaîne de traduction :

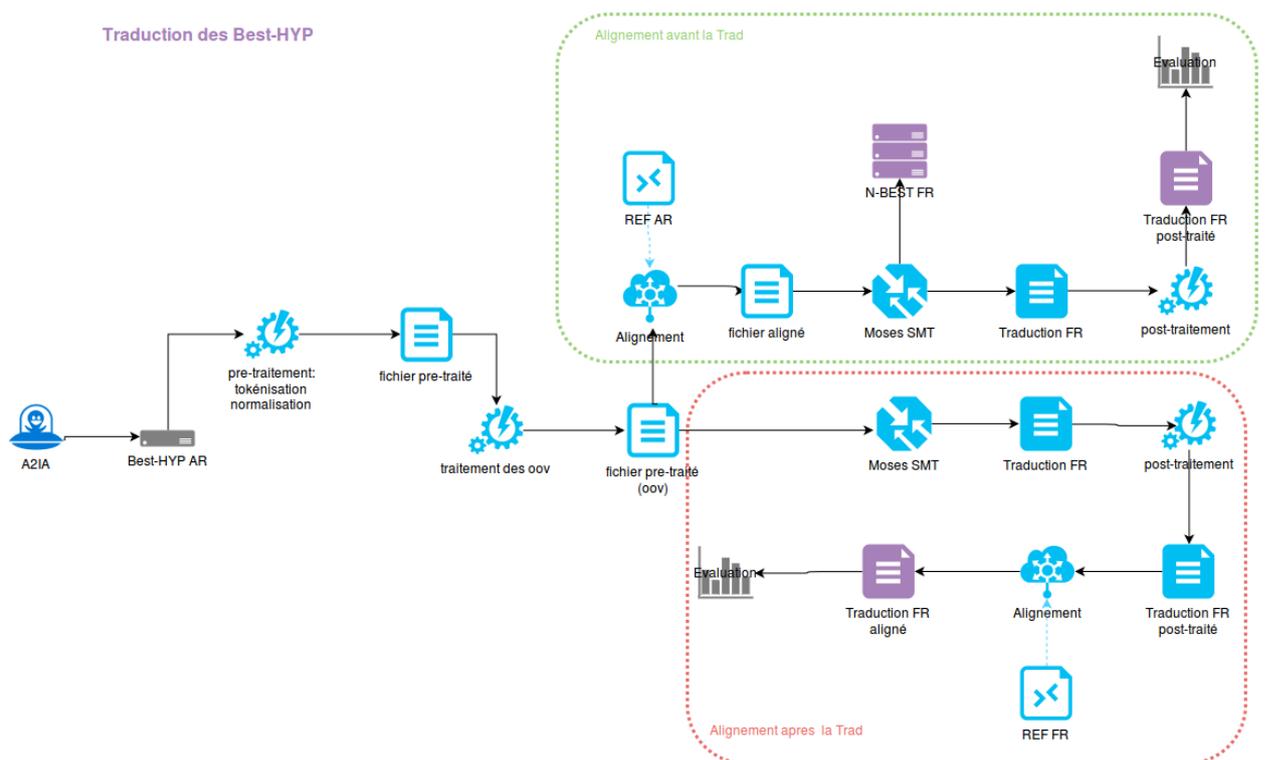


Figure 16 :Processus de traduction des sorties OCR (BEST-HYP)

Fichiers d'entées :

Nous recevons les sorties OCR sous le format XML qui comporte plusieurs documents dans le même fichier, donc avant tout on applique 2 étapes préalables :

- extraire tous les documents séparément.
- Convertir les documents XML en TXT.

Après nous appliquons notre prétraitement sur les fichiers à traduire (les prétraitements sont les mêmes que ceux appliqués sur les corpus d'entraînement arabe).

Alignement :

Pour pouvoir évaluer notre sortie de traduction il faut avoir le même nombre de lignes pour les sorties de traduction et pour les fichiers des références et dans notre cas les sorties OCR n'ont pas forcément le même nombre de lignes c'est pour ça nous avons ajouté une phase d'alignement après traduction en français.

Pour cette tâche nous avons utilisé l'outil MwerAlign (Matusov et al. 2005) qui permet de faire une nouvelle segmentation des phrases qui ne correspondent pas aux segments de référence, cet outil est aussi utilisé pour l'évaluation des traductions de la langue parlée.

Résultats :

Le tableau ci-dessous représente les résultats de traduction des sorties OCR en utilisant notre nouveau système (l'alignement est fait après la traduction).

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	20,39	14,67	16,67
Test	15,42	13,83	15,46

Tableau 16 Évaluation de nouveau système (alignement après la traduction) en score BLEU

Le tableau ci-dessous représente les mêmes résultats que le tableau ci-dessus, sauf que l'alignement est fait avant la traduction coté arabe, en supposant les références arabe connues.

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	22.64	15.34	17.94
Test	19.66	14.56	16.27

Tableau 17 Évaluation de nouveau système (alignement avant la traduction) en score BLEU

Le tableau ci-dessous représente les pourcentages de mots hors vocabulaire (OOV).

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	2.12 %	2.26 %	1.48 %
Test	2 %	2.5 %	1.71 %

Tableau 18 Evaluation de nouveau système en score OOV

Nb : Dans le reste de notre rapport nous allons présenter que les résultats de traduction en utilisant l'alignement après la traduction pour pouvoir les comparer entre eux, parce que lors de l'utilisation des graphes de mots nous n'aurons plus la possibilité d'aligner les documents avant la traduction.

3.3. Traitement des mots hors vocabulaire

Puisqu'on travaille sur des documents OCRisés en arabe, on est confronté au problème des mots hors vocabulaire qui peuvent être dus à des erreurs d'OCR ou à une couverture non suffisante de notre corpus d'entraînement.

Pour traiter les mots hors vocabulaire (OOV), nous proposons la méthode décrite dans la figure 17.

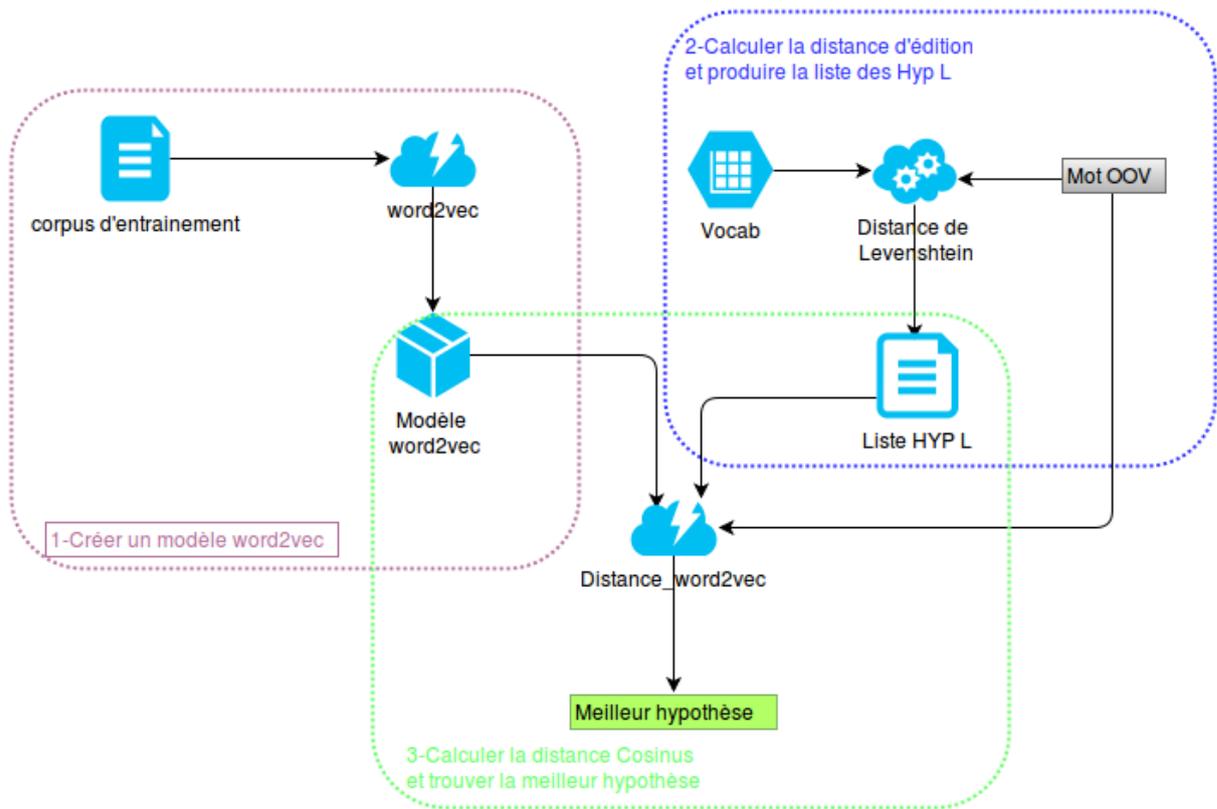


Figure 17 :Traitement des mots hors-vocabulaire

Cette méthode consiste à chercher un mot proche du mot hors-vocabulaire, présent dans le corpus d'apprentissage, selon sa forme de surface (en utilisant une distance d'édition) et selon son contexte (en utilisant une représentation distribuée du contexte des mots).

Plus précisément, celle-ci se décompose en trois étapes :

1- Créer un modèle vectoriel de mots (par exemple avec l'outil Word2Vec¹³) en utilisant un corpus d'entraînement qui se compose des corpus arabes source et des sorties OCR (*dev* et *test*). Ce corpus sera nommé *train+dev+test* dans le futur.

2- Pour chaque mot hors vocabulaire des corpus *dev* et *test*, trouver la liste des **L** mots les plus proches (dans le corpus *train*) selon une distance d'édition (type Levenshtein). La distance d'édition retourne un entier noté **n** ($n \leq 2$). Un exemple est donné ci-dessous pour le mot arabe OOV (en orange).

¹³ code.google.com/archive/p/word2vec/



Figure 18 Traitement des OOV étape 2

3- Trouver le meilleur mot hypothèse parmi la liste \underline{L} en calculant la *distance cosinus* entre le mot OOV et chaque mot de la liste \underline{L} grâce à l'outil **word2ec**. Un exemple est donnée ci-dessous où on voit le mot proche retrouvé dans le corpus d'apprentissage (en vert), pour le mot arabe OOV (en orange).



Figure 19 Traitement des OOV étape 3

Les nouveaux résultats après traitement des oov sur les corpus C2, C3 et C4 sont donnés dans les *tableaux ci-dessous*.

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	20,59	14,62	16,71
Test	16,07	13,82	15,48

Tableau 19 Évaluation du nouveau système (traitement des OOV) en score BLEU (alignement apres la traduction)

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	1.15	1.16	0.75
Test	1.15	1.48	0.58

Tableau 20 Evaluation de nouveau système (traitement des OOV) en score OOV

Nous observons que le taux de mots hors vocabulaire diminue effectivement, mais l'impact sur le score BLEU final reste faible à ce stade.

3.4. Nouveau modèle de langue (Interpolation linéaire)

Une nouvelle version du modèle de langue français est créée. Celui-ci est construit à partir de 6 modèles de langue obtenus à partir de 6 corpus différents.

Tous les modèles sont interpolés linéairement à l'aide de la boîte à outils SRILM (Stolcke, 2002) pour obtenir le modèle de langue final.

En effet, cette approche d'interpolation de plusieurs modèles de langue peut donner des résultats meilleurs que l'entraînement d'un seul modèle à partir d'un seul corpus.

Le *tableau 21* montre les perplexités des différents modèles estimées sur la partie *dev* des corpus C2+C3+C4 concaténés.

Corpus	Perplexité REF1	Perplexité REF2
EP7 (Europarl)	474	446
NU (Nations Unies)	449	414
News (Journaux)	468	396
Open sub (Sous-Titres)	786	730
Trame (issu de PEA-TRAD)	402	373
Wit3 (Talks)	627	594
Modèle interpolé	239	228

Tableau 21 Perplexité des modèles des langues sur le corpus dev (C2+C3+C4)

Les nouveaux résultats de traduction avec ce nouveau modèle de langue cible, sur les corpus C2, C3 et C4, sont donnés dans le *Tableau 22* (sans traitement des OOV) et dans le *Tableau 23* (avec traitement des OOV).

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	21,17	14,63	16,4
Test	16,41	16,16	16,18

Tableau 22 Évaluation du nouveau système (nouveau modèle de langue - pas de traitement des OOV)

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	21,34	14,31	16,28
Test	16,59	16,05	16,13

Tableau 23 Évaluation du nouveau système (nouveau modèle de langue - avec traitement des OOV)

4. Traduction de graphes

Le système de reconnaissance optique de caractères donne comme sorties une liste de N meilleures hypothèses (N-best) qui contient plus d'informations que la simple meilleure hypothèse. En vue d'utiliser un maximum d'informations, nous allons exploiter cette sortie en la transformant en un graphe de mots.

4.1. Treillis des mots

Un Treillis des mots (ou un graphe de mots) est un graphe acyclique orienté avec un point de départ unique et des arcs marqués avec des mots et leurs poids (Dyer et al. 2008) Nous avons 3 types de treillis des mots comme il représente la figure suivante :

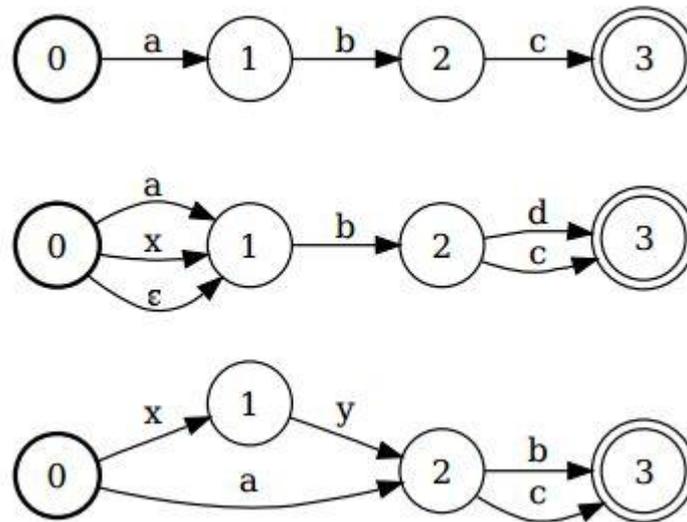


Figure 20 : 3 exemples des treillis: (a) phrase; (b) réseaux de confusion (c) treillis non- linéaire

Contrairement aux réseaux de confusion qui imposent que chaque chemin doive passer à travers chaque nœud, les treillis de mots (non-linéaire) peuvent représenter un ensemble fini de chaînes par exemple pour les chaînes [ab] et [ac] dans l'exemple de figure 20 (c) on n'est pas obligé de passer par le nœud numéro 1.

Cependant, en général un treillis des mots peut représenter un nombre exponentiel de phrases dans l'espace polynôme.

Voici un exemple de treillis montrant les moyens possibles pour représenter certains mots composés en allemand :

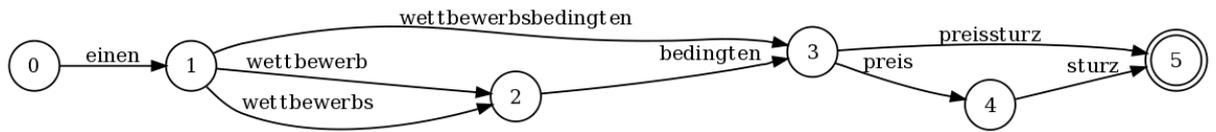


Figure 21 : exemple de treillis allemand

Notre outil de traduction Moses est capable de décoder l'entrée représentée sous forme du graphe de mots.

Quand Moses traduit une entrée codée comme un graphe de mots, il choisit de maximiser la probabilité de traduction le long de tout chemin dans l'entrée c'est-à-dire qu'il fait la traduction de tous les chemins possibles dans le graphe et nous donne la traduction avec le score le plus élevé.

Comment représenter les Graphes comme entrée :

```
(
  (
    ('einen', 1.0, 1),
  ),
  (
    ('wettbewerbsbedingungen', 0.5, 2),
    ('wettbewerbs', 0.25, 1),
    ('wettbewerb', 0.25, 1),
  ),
  (
    ('bedingen', 1.0, 1),
  ),
  (
    ('preissturz', 0.5, 2),
    ('preis', 0.5, 1),
  ),
  (
    ('sturz', 1.0, 1),
  ),
)
```

Figure 22 : Exemple de représentation de Graphe (PLF)

Les graphes de mots sont encodés en ordonnant les nœuds dans un ordre topologique et à l'aide de cet ordre des ID numériques consécutifs seront assignés aux nœuds. Puis, en procédant dans l'ordre par les nœuds, chaque nœud comporte à ses bords sortants les poids qui leur sont associés. Par exemple, le graphe ci-dessus peut être écrit dans le format Moses (PLF : Python lattice format).

Le deuxième chiffre est la probabilité associée à un arc. Le troisième chiffre est la distance entre le nœud de départ et le nœud d'arrive de l'arc, il faut aussi que les nœuds soient numérotés dans l'ordre topologique pour le calcul de la distance.

Configuration de Moses pour traduire des treillis :

Pour indiquer à Moses qu'il va lire des graphes au format PLF, il faut spécifier -inputtype 2 (0 : pour les textes; 1 : pour les réseaux de confusion) sur la ligne de commande ou dans le fichier de configuration moses.ini.

4.2. Chaîne de traduction des graphes

La prise en compte de l'ambiguïté issue du système OCR, nécessite une modification de l'architecture générale du système de traduction.

L'architecture générale de notre module « traduction », qui prend en entrée des sorties OCR du type « N-best », est donnée dans la figure ci-dessous.

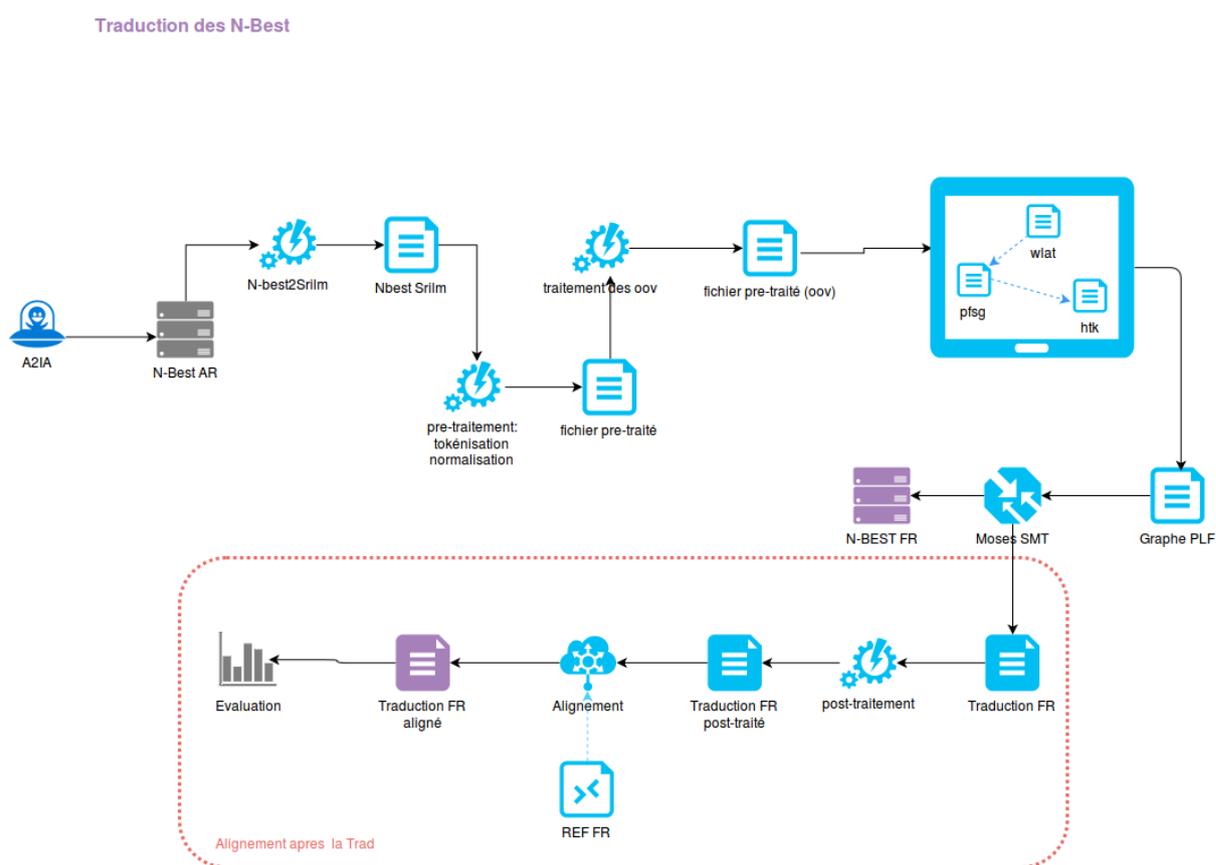


Figure 23 : Processus de traduction des Graphes de mots (N-Best)

Convertir des N-best en graphe :

Pour convertir les N-best en treillis de mot il faut passer par des étapes intermédiaires.

Les sorties N-best de l'OCR sont sous le format représenté dans la figure ci-dessous, c'est un format spécifique donné par le système d'OCR utilisé par A2IA

```
.
HYP_1-14
(s> (1.69629) (-1.26576>
(6.39922-) (10.9022) أزمة
(2.72041-, 6.39658-, 4.8816-) (5.63672, 2.86317, 1.36126) اخلاقية
(2.55469, 0-) (8.03904, 0.0810335) امة
(9.723-, 9.28856-, 4.64468-) (6.1688, 5.34742, 8.61105) أخلاقية
(2.5431-) (5.40521) تكة
(s> (1.06445) (0/>
.
HYP_1-15
(s> (1.69629) (-1.26576>
(4.16418-, 0.0265942-) (5.27242, 4.78405) زمة
(2.72041-, 6.39658-, 4.8816-) (5.28711, 2.86317, 1.36126) اخلاقية
(2.51492-) (6.15621) مة
(9.723-, 9.28856-, 4.64468-) (6.1688, 5.34742, 8.61105) أخلاقية
(2.5431-) (5.40521) تكة
(s> (1.06445) (0/>
.
HYP_1-16
(s> (1.69629) (-1.26576>
(3.6548-, 2.11551-, 0) (1.49314, 4.18062, 6.75868) ارقفة
(2.72041-, 6.39658-, 4.8816-) (5.25488, 2.86317, 1.36126) اخلاقية
(2.51492-) (6.15621) مة
(9.723-, 9.28856-, 4.64468-) (6.1688, 5.34742, 8.61105) أخلاقية
(2.5431-) (5.40521) تكة
(s> (1.06445) (0/>
.
```

Figure 24 Format des N-Best fournis par l'OCR (A2IA)

Dans notre cas nous avons des 100-best, pour chaque hypothèse nous avons l'entête de l'hypothèse (HYP_X_Y) le X représente le numéro de paragraphe et le Y représente le numéro de l'hypothèse, après nous trouvons la liste des mots composent notre hypothèse et pour chaque mots deux scores (score linguistique et score optique).

Convertir les N-best en N-best_srilm :

Le format supporté par l'outil SRILM est sous la forme suivante :

Ascore Lscore nwords w1 w2 w3 ...

Ascore: Score optique

Lscore: Score de modèle de langue

nwords : le nombre de mots dans la phrase

wi : la liste des mots de la phrase

ازمة اخلاقية مة اخلاقية تكة	5	1.69566017336187	1.72742412811647
ازمة اخلاقية زمة اخلاقية تكة	5	1.68611050843878	1.72493039873331
ازمة اخلاقية زمة اخلاقية تكم	5	1.67142593763329	1.72111865425425
ازمة اخلاقية زمة اخلاقية تكة	5	1.70702671263283	1.74736798777099
ازمة اخلاقية زمة اخلاقية تكة	5	1.68153862725128	1.71593725067829
ازمة اخلاقية امة اخلاقية تكة	5	1.69600811447358	1.74311341563717
ازمة اخلاقية امة اخلاقية تكة	5	1.69153978029408	1.73572584442078
زمة اخلاقية مة اخلاقية تكة	5	1.67120640154771	1.709750642076
ارقة اخلاقية مة اخلاقية تكة	5	1.68559091820935	1.72916835478209
ازمة اخلاقية مة اخلاقية تكة	4	1.74481061959081	1.78746345009847
ازمة اخلاقية امة اخلاقية تكة	5	1.67454094861786	1.71189644420258
زمة اخلاقية زمة اخلاقية تكة	5	1.68322229025401	1.73118143321974
ازمة اخلاقية مة اخلاقية تكة	5	1.69118823937762	1.71976238373099
ازمة اخلاقية امة اخلاقية تكة	5	1.67454094861786	1.7197540251524

Figure 25 Exemple des Nbest-Srilm

Une fois que nous avons nos données sous le format nbest-srilm, nous appliquons nos prétraitements (tokenisation, normalisation... etc.).

Et pour arriver au format PLF lisible par *Moses*, à l'aide de l'outil SRILM (Stolcke, 2002) pour générer un graphe au format HTK¹⁴ et finalement nous utilisons l'outil htk2plf développé par (Sylvain Raybaud 2010) pour convertir les fichiers HTK en graphe de mots (PLF).

¹⁴ www.seas.ucla.edu/spapl/weichu/htkbook/node457_mn.html

Si on compare les résultats du *Tableau 24* avec ceux du *tableau 22* (les résultats de traduction avant l'utilisation des graphes), on voit que le traitement des graphes OCR améliore les performances de traduction.

4.3. Traitement des mots hors vocabulaire

Le fait d'avoir un graphe en entrée nous permet d'ajouter plus d'une hypothèse de mot connu remplaçant le mot hors vocabulaire (et on laisse aussi la phrase initiale contenant le mot OOV dans le graphe).

Nous décidons d'ajouter les 4 meilleures hypothèses. En d'autres termes, si on considère les 100 meilleures hypothèses de l'OCR qui ont 10 phrases possédant des mots OOV, alors le graphe sera construit à partir de $100+(4*10)=140$ hypothèses. Ces dernières améliorations constituent notre meilleur système, on voit aussi que le nombre de mots OOV de notre corpus a beaucoup diminué grâce à la prise en compte des graphes et à notre traitement spécifique.

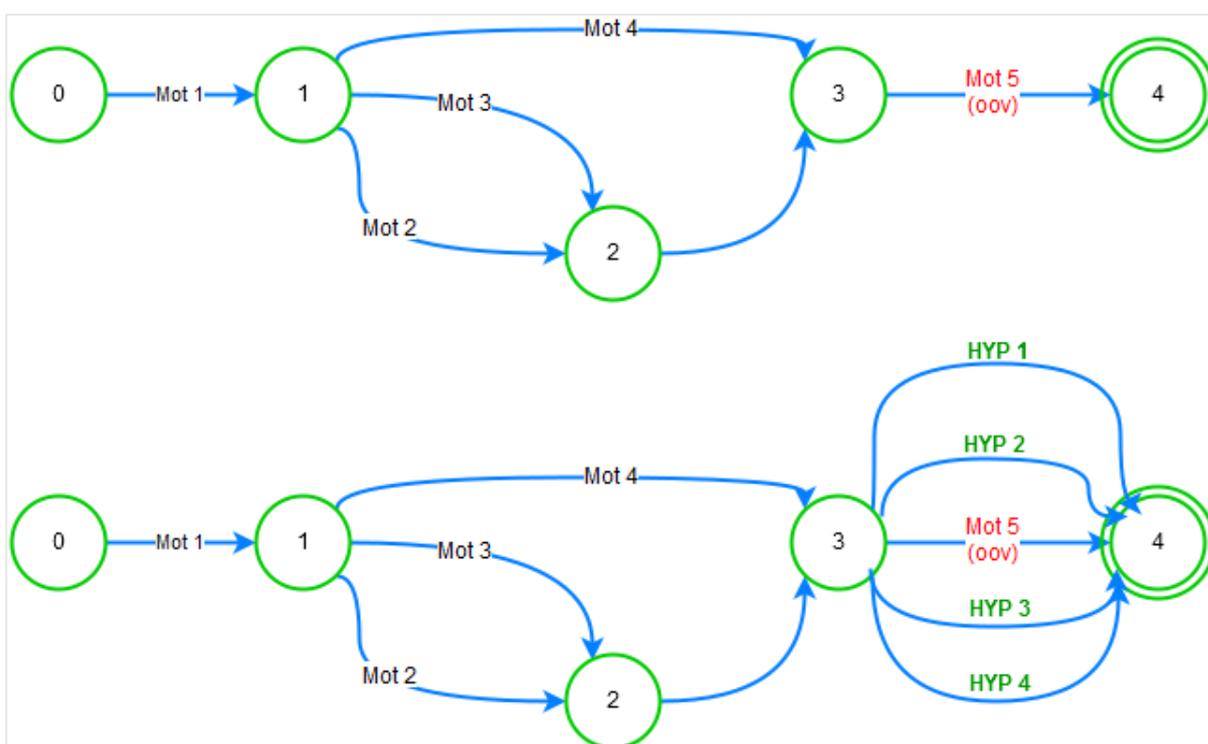


Figure 27 Exemple de Graphe avant et après traitement des OOV

Les nouveaux résultats sur les corpus C2, C3 et C4 sont donnés dans le *Tableau 26* (avec traitement spécifique des OOV) en score bleu et en score OOV dans le *Tableau 27*.

Pour le traitement spécifique des OOV :

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	24,9	14,5	16,88
Test	21,91	17,5	15,98

Tableau 26 Évaluation du nouveau système (traitement de graphes OCR - avec traitement des OOV)

Corpus	Journaux C2	Courrier C3	Courrier C4
Dev	0.12	0.16	0.41
Test	0.09	0.32	0.37

Tableau 27 Évaluation du nouveau système (traitement de graphes OCR - avec traitement des OOV) en score OOV

5. Bilan

Dans cette section nous allons présenter la progression des performances de notre système de traduction en fonction de différentes évolutions sur les corpus C2, C3 et C4 respectivement:

- (1) : système initial – Baseline
- (2) : Création d'un nouveau modèle de traduction
- (3) : L'ajout de traitement des OOV
- (4) : Création d'un nouveau modèle de langue (méthode d'interpolation linéaire)
- (5) : L'ajout de traitement des OOV à notre nouveau système
- (6) : L'utilisation des Graphes
- (7) : L'utilisation des Graphe + traitement des OOV

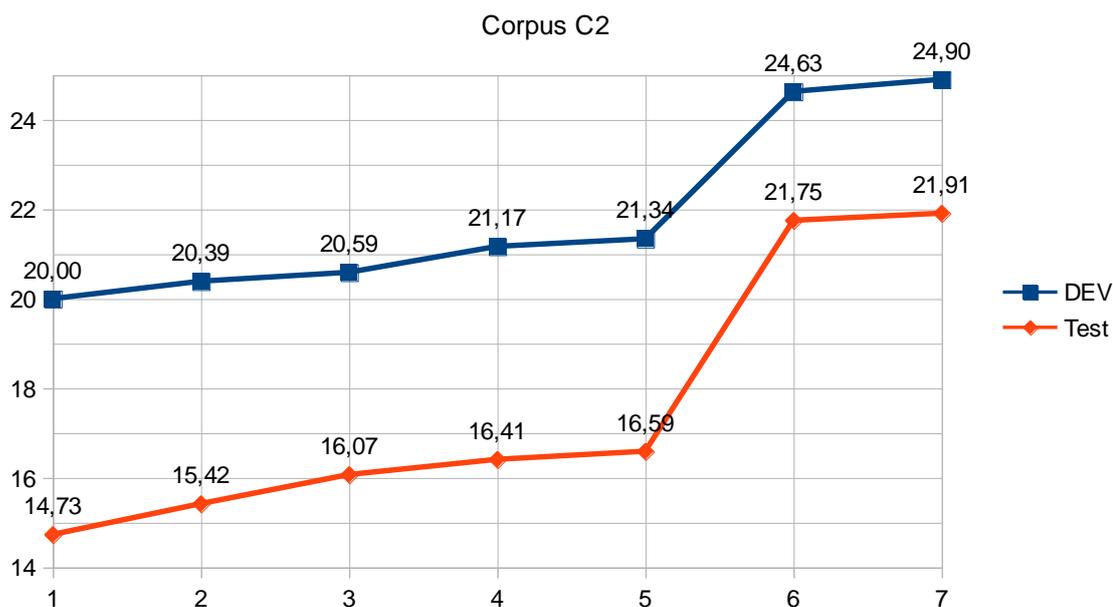


Figure 28 : progression des performances en score Bleu sur le corpus C2

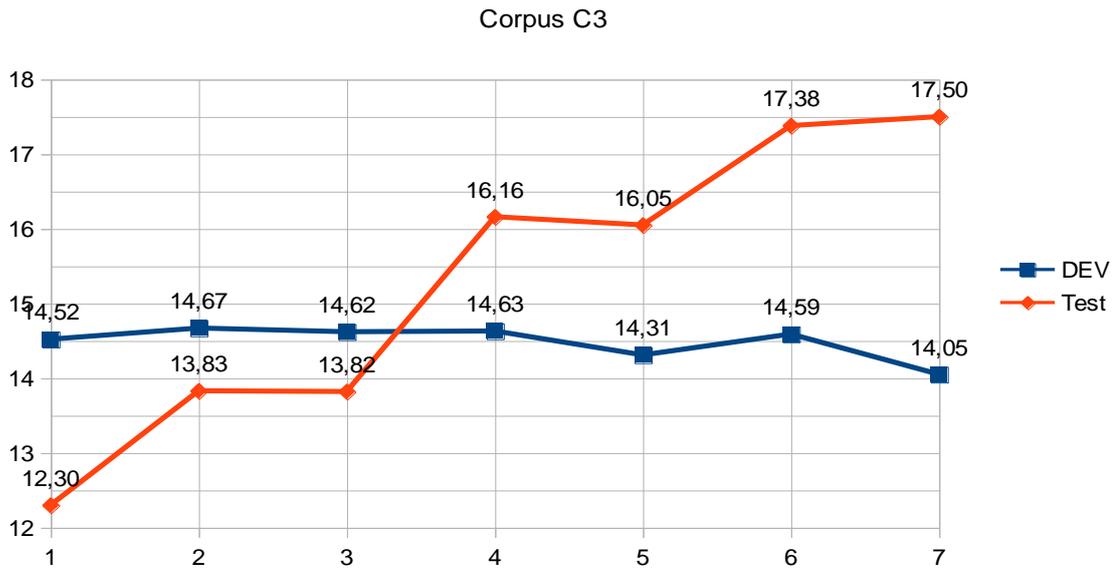


Figure 29 : progression des performances en score Bleu sur le corpus C3

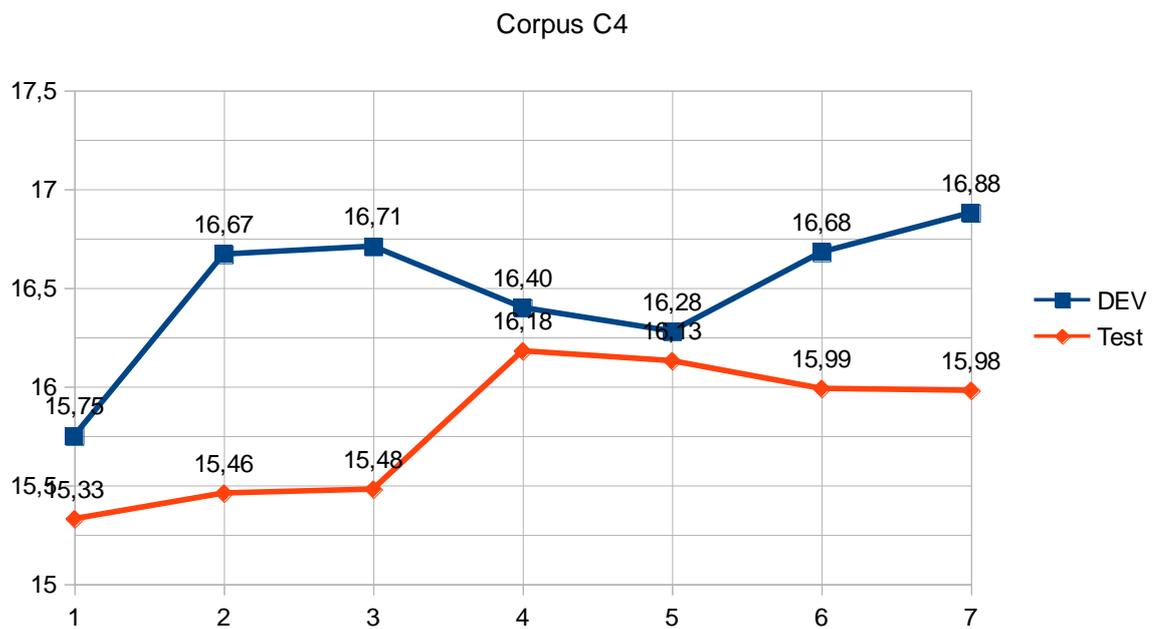


Figure 30 : progression des performances en score Bleu sur le corpus C4

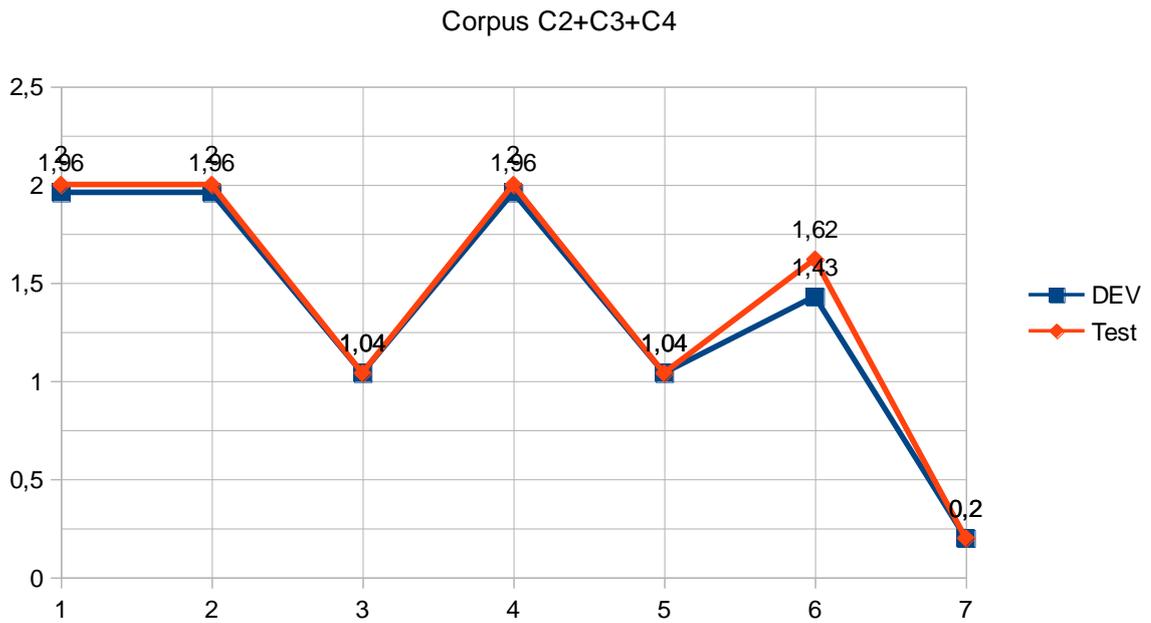


Figure 31 progression des performances en score OOV sur le corpus C2+C3+C4

Après avoir comparé les différents résultats nous pouvons interpréter en premier lieu l'influence du nouveau modèle de langue créé par la méthode de l'interpolation linéaire entre plusieurs modèles, et en second lieu l'effet de l'utilisation des graphes qui améliore bien les performances de traduction et nous permet d'utiliser des données bruitées à l'entrée de notre système.

Par ailleurs, le traitement particulier des mots hors vocabulaire en se basant sur la distance l'édition et la distance cosinus nous aide beaucoup à diminuer le nombre de mots OOV dans notre sortie finale de traduction ainsi qu'à améliorer notre score de traduction.

Conclusion et perspectives

Conclusion

Dans notre travail de recherche, nous nous sommes intéressés à la traduction automatique des sorties OCR et l'utilisation des entrées bruitées (N-best) ainsi que le traitement des mots hors vocabulaire.

Au début de notre rapport nous avons commencé par une description générale du domaine de reconnaissance automatique de caractères et la traduction automatique statistique.

De plus, nous avons présenté les données utilisées pour l'apprentissage de notre système ainsi que les différents prétraitements appliqués sur ces corpus parallèles telles que la normalisation de caractères, la suppression de diacritique et la tokenisation etc. Ensuite nous avons fait une première traduction de sortie OCR (les meilleures hypothèses) après la création de notre système de traduction.

Par ailleurs, nous avons ajouté notre méthode de traitement des mots hors vocabulaire qui se base sur la recherche de mots les plus proches de notre mot OOV en utilisant la distance d'édition et la distance cosinus à notre chaîne de traduction. Après ce traitement, nous n'avons pas eu une grande amélioration en termes de score BLEU (gain entre 0 et 0,40 pt). Par contre, le nombre de mots hors vocabulaire dans la sortie de traduction a diminué de 1% en termes de score OOV.

En outre, nous avons modifié notre système de traduction en ajoutant une nouvelle version de modèle de langue en utilisant la méthode d'interpolation linéaire de plusieurs modèles. En effet, nous avons eu des améliorations intéressantes en terme de score BLEU (gain jusqu'à 2,5 pt).

Enfin, nous nous sommes concentrés sur l'intégration de l'utilisation des graphes de mots dans notre chaîne de traduction. Nous avons commencé par la conversion des sorties OCR (n-best) vers des graphes de mots et l'adaptation du système de traduction pour traduire ce type d'entrée. Puis, nous avons ajouté un traitement spécifique des mots hors vocabulaire à notre chaîne de traduction. Et après l'évaluation de nos sorties de traduction, nous avons obtenu des améliorations remarquables (jusqu'à 5,3 pt de gain). En terme de score BLEU et de score OOV, le nombre de mots hors vocabulaire a diminué de 1,8 %.

Perspectives

Dans notre travail nous avons utilisé l’outil Moses pour créer un système de traduction de type Phrase-Based. Dans un futur proche, il est envisageable d’utiliser un système de traduction neuronal qui a récemment trouvé une place centrale dans le paysage de traitement automatique des langues (TAL).

Par ailleurs, nous pourrions améliorer notre méthode de traitement de mots hors vocabulaire par l’utilisation de l’outil MultiVec développé récemment par (Bérard et al. 2016) qui permet de calculer la distance entre des mots ou des séquences des mots dans deux langue différentes a la place de word2vec qui utilise juste une seule langue.

Annexes

Annexe 1 : Extrait de sortie OCR (Corpus C2)

الشرعي فوق أراضيهم الاحتلال، وايضا الى رموز الاستيطان غير الفلسطينيين يصوبون الى الجنود، رمز فالشبان . حياتهم باتت في خطر ويدركون أن يعرفون ذلك والمستوطنون يسكا مليون دولار 500 بكلفة تبلغ – المتتالية بشتى شبكة واسعة من الطرق سبيل حمايتهم قامت الحكومات الاسرائيلية في المدن والقرى الفلسطينية اليهود وحدهم بهدف الالتفاف على منذ لحم ومن خلال تفوق – اقلقت عشرات المرات والتي تربط القدس بمجمع اتزيون جنوب بيت – امانا . أكثر هذه الطرق لكن الفلسطينيين بدء الانتفاضة بعد تعرضها لنيران بالطبع . تعرضها لنيران الفلسطينيين بعد ارسال الاحتياطيين في الاراضي المحتلة ثلاث مرات بما في ذلك المستوطنون بحماية الجيش الذي ضاعف عديده طالب

كانت هي أبحرك النقاشات العامة الزعيم المجاهد الأكبر الذي لولاه لما ستار مجلس كان يقف رجل تونس الواحد الأوحد وراء مجتمعا سياسيا وأخر مدنيا 2011 حاضرا أو غائبا، بينما يقف وراء مجلس شؤون للمجلس ويحدد محتوياته سواء كان ويدير الذي في مناقشة الشأن العام والهم السياسي بامتياز، ورغبة جامعة من كافة التونسيين وثالثا إعلاميا رنخبة وطنية كانت مقيم حرموا منه لأكثر من نصف قرن المؤسسة الحاكمة وقبل بإعلان وحرية دون سواهم وشرعن الانقلاب على منح للمجلس التأسيسي السلطة البورقيلية في تك الدستور الذي من المفترض له وصفوا قبل أن يكملوا مهمتهم في وضع الذي تلاه بورقيلية على مسامع النواب فهو الجمهورية لمجموعة من الأحزاب الحكم 2011 بينما أعطى مجلس النظام السياسي ملكيا كان أم جمهوريا، أن يقر طبيعة

Les jeunes palestiniens - soldats, symbole de l'occupation et à la colonisation illégale au-dessus de leurs terres.

Et les colons savent. Ils savent que leur vie est en danger.

Pour protéger les gouvernements israéliens successifs divers un vaste réseau de routes – pour un coût de 500 millions de dollars à des juifs seuls dans le but de contourner les villes et les villages palestiniens.

Mais la méthode la plus sûre. - Et reliant Jérusalem complexe Etzion au sud de Bethléem en – est fermé des dizaines de fois depuis le début de l'Intifada, après avoir fait l'objet de tirs palestiniens. Après avoir fait l'objet de tirs palestiniens.

Les colons à la protection de l'armée que de nombreuses dans les territoires occupés à trois reprises, notamment en envoyant des احتياطين.

Derrière le Conseil était debout hommes de la Tunisie un seul dirigeant moudjahid plus que Lula si elle يحرك les débats publics et de gérer les affaires du Conseil le contenu soit présent ou غاتبا, derrière le Conseil en 2011 مجتمعا سياسيا et un autre civil et médiatiquement رنخبة national résident excellence, et le désir ardent de tous les Tunisiens à la discussion publique et politique qui ont été privés de plus d'un demi-siècle

Dans ta octroi de l'Assemblée constituante l'autorité بورقبيبة et de la liberté de communication et entreprennent de coup de l'institution dirigeante avant la déclaration de la République lue par Bourguiba à des députés avez apprise il était avant leur mission dans l'élaboration de la Constitution, qui reconnaît la nature du système politique, Royal ou républicain, alors que le Conseil en 2011 la disposition du groupe de partis

Annexe 3 : Extrait de la traduction de sortie OCR (Corpus C2) après traitement des mots hors vocabulaire

Les jeunes palestiniens - soldats, symbole de l'occupation et à la colonisation illégale au-dessus de leurs terres.

Et les colons savent. Ils savent que leur vie est en danger.

Pour protéger les gouvernements israéliens successifs divers un vaste réseau de routes – pour un coût de 500 millions de dollars à des juifs seuls dans le but de contourner les villes et les villages palestiniens.

Mais la méthode la plus sûre. - Et reliant Jérusalem complexe Etzion au sud de Bethléem en – est fermé des dizaines de fois depuis le début de l'Intifada, après avoir fait l'objet de tirs palestiniens. Après avoir fait l'objet de tirs palestiniens.

Les colons à la protection de l'armée que de nombreuses dans les territoires occupés à trois reprises, notamment en envoyant des réservistes.

Derrière le Conseil était debout hommes de la Tunisie un seul dirigeant moudjahid plus que Lula si elle déplace les débats publics et de gérer les affaires du Conseil le contenu soit présent ou de gatumba, alors que derrière le Conseil en 2011 مجتمعا سياسيا et un autre civil et médiatiquement رنخبة national résident excellence, et le désir ardent de tous les Tunisiens à la discussion publique et politique qui ont été privés de plus d'un demi-siècle

Dans ta octroi de l'Assemblée constituante l'autorité Bourguiba et de la liberté de communication et entreprennent de coup de l'institution dirigeante avant la déclaration de la République lue par Bourguiba à des députés avez apprise il était avant leur mission dans l'élaboration de la Constitution, qui reconnaît la nature du système politique, Royal ou républicain, alors que le Conseil en 2011 la disposition du groupe de partis

Annexe 4 : exemple de fichier N-best de sortie OCR (C2_Ar_TBKBIS_14)

HYP_14-1
<s> (8.78395) (-13.5884)
2.54031-) (0.522418) اللغة
19.2736-,13.1935,0-,4.11673-,6.1807-) (0.513673,17.7271-,5.29284,3.86122,12.0193) (المزدوجة
3.13013-) (1.24017) للحكم
2.98302-) (0.71185) في
5.96046-,13.8831-,7.24654-,3.84989) (5.97952-,2.5263,6.2176,10.5086-) (07-رومانيا
</s> (1.78516) (0)

HYP_14-10
<s> (8.78395) (-13.5884)
2.54031-) (0.522418) اللغة
19.2736-,13.1935-,4.11673-,6.1807-) (5.29284,3.86122,17.9236,16.774) (المزدوجة
3.13013-) (1.24017) للحكم
2.98302-) (0.71185) في
5.96046-,13.8831-,7.24654-,3.84989) (5.97952-,2.5263,6.2176,10.5086-) (07-رومانيا
</s> (1.78516) (0)

HYP_14-11
<s> (8.78395) (-13.5884)
2.54031-) (0.522418) اللغة
25.7015-,4.11673-,6.1807-) (5.29284,3.86122,24.1937) (المزدوجة
3.13013-) (1.24017) للحكم
2.98302-) (0.71185) في
5.96046-,13.8831-,7.24654-,3.84989) (5.97952-,2.5263,6.2176,10.5086-) (07-رومانيا
</s> (1.78516) (0)

HYP_14-12
<s> (8.78395) (-13.5884)
2.54031-) (0.522418) اللغة
19.2736-,13.1935,0-,4.11673-,6.1807-) (0.513673,17.7271-,5.29284,3.86122,12.0193) (المزدوجة
3.13013-) (1.24017) للحكم
2.98302-) (0.71185) في
3.97364-,7.24654-,3.84989) (5.97951-,2.5263,6.2176,4.17773,12.2908-) (07-رومانيا
</s> (1.78516) (0)

HYP_14-13
<s> (8.78395) (-13.5884)
2.54031-) (0.522418) اللغة
19.2736-,13.1935,0-,4.11673-,6.1807-) (0.513673,17.7271-,5.29284,3.86122,12.0193) (المزدوجة
3.13013-) (1.24017) للحكم
2.98302-) (0.71185) في
17.2798,0-) (13.944,0.242166) (رومانيا
</s> (1.78516) (0)

HYP_14-100
<s> (8.78395) (-13.5884)
2.54031-) (0.522418) اللغة
-,19.2736-,13.1935,0-,4.11673-,6.1807-) (0.513673,17.7271,1.24017-,5.29284,3.86122,12.0193) (المزدوجة
3.13013) للحكم
2.98302-) (0.71185) في
17.2798,0-) (13.944,0.242166) (رومانيا
</s> (1.78516) (0)

1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.7367667139611 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.74073674018054 1.83694276608076 6 ال لغة ال مزدوجة للحكم في رومانيا
1.64387273215271 1.7919160910379 7 ال لغة ال مزدوجة للحكم في رومانيا
1.74069061591709 1.83694245627022 7 ال لغة ال مزدوجة للحكم في رومانيا
1.74073674018054 1.83694276608076 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.68766631744796 1.79115738557396 8 ال لغة ال مزدوج ه للحكم في رومانيا
1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.72454435266365 1.8112994001162 8 ال لغة ال مزدوجة للحكم في رومانيا
1.73520863375186 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.75820816649396 1.85764730971541 6 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 6 ال لغة ال مزدوجة للحكم في رومانيا
1.67819629902188 1.79191581054082 7 ال لغة ال مزدوجة للحكم في رومانيا
1.73934023944993 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.75840152359792 1.83694251697632 7 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.7367667139611 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.64387273215271 1.7919160910379 6 ال لغة ال مزدوجة للحكم في رومانيا
1.76212762051997 1.85764730971541 6 ال لغة ال مزدوجة للحكم في رومانيا
1.64387273215271 1.7919160910379 7 ال لغة ال مزدوجة للحكم في رومانيا
1.71515068865365 1.80604152459656 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 6 ال لغة ال مزدوجة للحكم في رومانيا
1.68290436438604 1.79191581054082 7 ال لغة ال مزدوجة للحكم في رومانيا
1.73520863375186 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.70864630081103 1.83694245375816 8 ال لغة ال مزدوجة للحكم في رومانيا
1.74069061591709 1.83694245627022 7 ال لغة ال مزدوجة للحكم في رومانيا
1.74073674018054 1.83694276608076 7 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.74073674018054 1.83694276608076 7 ال لغة ال مزدوجة للحكم في رومانيا
1.74088359352482 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.64896572684956 1.7919160910379 7 ال لغة ال مزدوجة للحكم في رومانيا
1.69549161436514 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 6 ال لغة ال مزدوجة للحكم في رومانيا
1.67819629902188 1.79191581054082 7 ال لغة ال مزدوجة للحكم في رومانيا
1.73520863375186 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.67737677039889 1.79115738557396 8 ال لغة ال مزدوج ه للحكم في رومانيا
1.74477063991028 1.83694245627022 7 ال لغة ال مزدوجة للحكم في رومانيا
1.75820816649396 1.85764730971541 6 ال لغة ال مزدوجة للحكم في رومانيا
1.6909187559956 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.7367667139611 1.83694245375816 7 ال لغة ال مزدوجة للحكم في رومانيا
1.64387273215271 1.7919160910379 7 ال لغة ال مزدوجة للحكم في رومانيا
1.67819629902188 1.79191581054082 7 ال لغة ال مزدوجة للحكم في رومانيا

Annexe 6 : exemple fichier N-best WLAT (fichier C2_Ar_TBKBIS_14)

```
version 2
name nbest_C2_Ar_TBKBIS_14.txt.nbest.pre
initial 0
final 1
node 0 NULL 0 1 2 0.989822 16 0.0101776
node 1 NULL 1 1
node 2 0.989822 3 0.989822 2 ال
node 3 0.989822 4 0.989822 3 لغة
node 4 0.00975834 19 0.129003 18 0.0102287 15 0.0819857 13 0.0108822 9 0.685426 5 1 4 ال
0.00848555 30 0.018102 29 0.0365842 28 0.009544 21
node 5 0.576051 11 0.109375 6 0.685426 5 مزدوجة
node 6 0.0110177 8 0.117935 7 0.128952 6 للحكمفي
node 7 0.029071 20 0.838469 1 0.86754 7 رومانيا
node 8 0.0834124 1 0.0834124 7 رومانيا
node 9 0.0108822 7 0.0108822 5 مزدوجةللحكمفي
node 10 24 0.0189012 23 0.00993383 17 0.0207524 12 0.0723948 8 0.711521 7 0.860165 6 في
0.00793196 31 0.0093367 26 0.00939358
node 11 0.00937694 25 0.00954233 22 0.741158 10 0.760078 8 للحكم
node 12 0.0207524 1 0.0207524 7 رومانيا
node 13 0.0819857 10 0.0819857 5 مزدوجةللحكم
node 14 0.0102287 11 0.0102287 5 وجة
node 15 0.0102287 14 0.0102287 9 مزد
node 16 0.0101776 4 0.0101776 2 اللغة
node 17 « 10 0.00993383 7 0.00993383
node 18 0.109426 11 0.0195772 6 0.129003 5 مزدوجة
node 19 0.00975834 11 0.00975834 5 مزدوجة
node 20 . 11 0.029071 1 0.029071
node 21 0.009544 11 0.009544 5 مزدوجة
node 22 . 12 0.00954233 10 0.00954233
node 23 0.0189012 1 0.0189012 7 ارومانيا
node 24 0.00939358 1 0.00939358 7 درومانيا
node 25 ، 12 0.00937694 10 0.00937694
node 26 0.0093367 7 0.0093367 10 و
node 27 0.0365842 11 0.0365842 5 ه
node 28 0.0365842 27 0.0365842 9 مزدوج
node 29 0.018102 10 0.018102 5 مزدوجةللحكم
node 30 0.00848555 11 0.00848555 5 مزدوجة
node 31 DIESE 10 0.00793196 7 0.00793196
```

Annexe 7 : exemple fichier N-best PFSG (fichier C2_Ar_TBKBIS_14)

```
name nbest_C2_Ar_TBKBIS_14.txt.nbest.pre
nodes 32 NULL NULL ال لغة ال مزدوجة للحكمي رومانيا رومانيا مزدوجة للحكمي في للحكم رومانيا
مزدوجة للحكم وجة مزد اللغة « مزدوجة مزدوجة . مزدوجة . ارومانيا درومانيا ، و ه مزدوج مزدوجة للحكم مزدجة
DIESE
initial 0
final 1
transitions 54
0 2 -102
0 16 -45878
2 3 0
3 4 0
4 5 -3777
4 9 -45209
4 13 -25013
4 15 -45828
4 18 -20480
4 19 -46299
4 21 -46521
4 28 -33083
4 29 -40119
4 30 -47696
5 6 -18353
5 11 -1739
6 7 -893
6 8 -24601
7 1 -341
7 20 -33961
8 1 0
9 7 0
10 7 -1897
10 8 -24751
10 12 -37246
10 17 -44614
10 23 -38181
10 24 -45173
10 26 -45234
10 31 -46865
11 10 -252
11 22 -43779
11 25 -43954
12 1 0
13 10 0
14 11 0
15 14 0
16 4 0
17 7 0
18 6 -18856
18 11 -1646
.
.
.
31 7 0
```

Annexe 8 : exemple fichier HTK (fichier C2_Ar_TBKBIS_14)

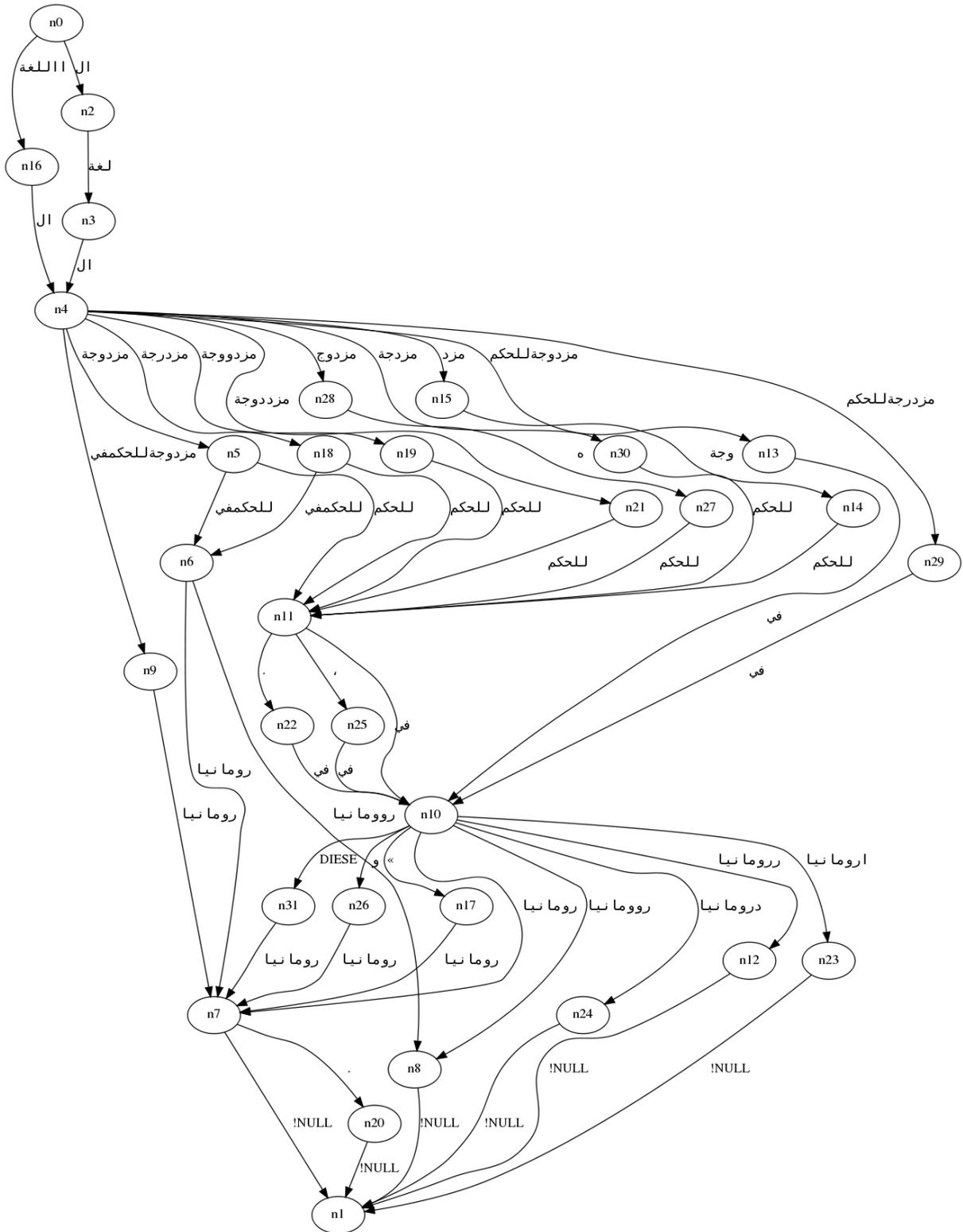
```
# Header (generated by SRILM)
VERSION=1.1
UTTERANCE=nbest_C2_Ar_TBKBIS_14.txt.nbest.pre
base=10
dir=f
start=0
end=1
NODES=32 LINKS=54
# Nodes
I=0
I=1
I=2
I=3
I=4
I=5
I=6
I=7
I=8
I=9
I=10
I=11
I=12
I=13
I=14
I=15
I=16
I=17
I=18
I=19
I=20
I=21
I=22
I=23
I=24
I=25
I=26
I=27
I=28
I=29
I=30
I=31
# Links
J=0 S=0 E=2 W=ال a=-0.00442958
J=1 S=0 E=16 W=اللغة a=-1.99236
J=2 S=2 E=3 W=لغة a=0
J=3 S=3 E=4 W=ال a=0
J=4 S=4 E=5 W=مزوجة a=-0.164025
J=5 S=4 E=9 W=مزوجة للحكمفي a=-1.9633
```

J=6 S=4 E=13 W=مزدوجة للحكم a=-1.08625
 J=7 S=4 E=15 W=مزد a=-1.99019
 J=8 S=4 E=18 W=مزدوجة a=-0.889391
 J=9 S=4 E=19 W=مزدوجة a=-2.01064
 J=10 S=4 E=21 W=مزدوجة a=-2.02028
 J=11 S=4 E=28 W=مزدوج a=-1.4367
 J=12 S=4 E=29 W=مزدوجة للحكم a=-1.74226
 J=13 S=4 E=30 W=مزدوجة a=-2.07131
 J=14 S=5 E=6 W=للحكمفي a=-0.797021
 J=15 S=5 E=11 W=للحكم a=-0.07552
 J=16 S=6 E=7 W=رومانيا a=-0.0387806
 J=17 S=6 E=8 W=رومانيا a=-1.06835
 J=18 S=7 E=1 W=!NULL a=-0.0148087
 J=19 S=7 E=20 W=. a=-1.47483
 J=20 S=8 E=1 W=!NULL a=0
 J=21 S=9 E=7 W=رومانيا a=0
 J=22 S=10 E=7 W=رومانيا a=-0.0823815
 J=23 S=10 E=8 W=رومانيا a=-1.07487
 J=24 S=10 E=12 W=رومانيا a=-1.61749
 J=25 S=10 E=17 W=« a=-1.93746
 J=26 S=10 E=23 W=ارومانيا a=-1.6581
 J=27 S=10 E=24 W=درومانيا a=-1.96174
 J=28 S=10 E=26 W=و a=-1.96439
 J=29 S=10 E=31 W=DIESE a=-2.03522
 J=30 S=11 E=10 W=في a=-0.0109437
 J=31 S=11 E=22 W=. a=-1.9012
 J=32 S=11 E=25 W= ،a=-1.9088
 J=33 S=12 E=1 W=!NULL a=0
 J=34 S=13 E=10 W=في a=0
 J=35 S=14 E=11 W=للحكم a=0
 J=36 S=15 E=14 W=وجة a=0
 J=37 S=16 E=4 W=ال a=0
 J=38 S=17 E=7 W=رومانيا a=0
 J=39 S=18 E=6 W=للحكمفي a=-0.818865
 J=40 S=18 E=11 W=للحكم a=-0.0714813
 J=41 S=19 E=11 W=للحكم a=0
 J=42 S=20 E=1 W=!NULL a=0
 J=43 S=21 E=11 W=للحكم a=0
 J=44 S=22 E=10 W=في a=0
 J=45 S=23 E=1 W=!NULL a=0
 J=46 S=24 E=1 W=!NULL a=0
 J=47 S=25 E=10 W=في a=0
 J=48 S=26 E=7 W=رومانيا a=0
 J=49 S=27 E=11 W=للحكم a=0
 J=50 S=28 E=27 W=ه a=0
 J=51 S=29 E=10 W=في a=0
 J=52 S=30 E=11 W=للحكم a=0
 J=53 S=31 E=7 W=رومانيا a=0

Annexe 9 : exemple fichier PLF (fichier C2_Ar_TBKBIS_14)

```
((('9.898524,'eال-01,2),('1.017747,'eاللغة-02,1),),(('1.000000,'eال+00,3),),(('1.000000,'eلغة+00,1),),(('1.000000,'eال+00,1),),(('6.85,'مزدوجة-4488e-01,12),('1.088178,'مزدوجةللحكفي-02,11),('8.198794,'مزدوجةللحكم-02,10),('1.022845,'مزدد-02,8),('1.290057,'مزدرجة-01,7),('9.757982,'مزدووجة-03,6),('9.543771,'مزددووجة-03,5),('3.658474,'مزدوج-02,3),('1.810256,'مزدرجةللحكم-02,2),('8.485745,'مزدوجة-03,1),),(('1.000000,'eللحكم+00,12),),(('1.000000,'eفي+00,14),),(('1.000000,'eه+00,1),),(('1.0,'للحكم-00000e+00,9),),(('1.000000,'eللحكم+00,8),),(('1.000000,'eللحكم+00,7),),(('8.482399,'eللحكم-01,6),('1.517522,'eللحكفي-01,16),),(('1.000000,'eووجة+00,1),),(('1.000000,'eللحكم+00,4),),(('1.000000,'eفي+00,6),),((','رومانيا-1.000000e+00,14),),(('1.595802,'eللحكفي-01,11),('8.403883,'eللحكم-01,1),),((','1.233673e-02,2),('9.751160,'eفي-01,3),('1.255452e-02,1),),(('1.000000,'eفي+00,2),),(('1.000000,'eفي+00,1),),(('8.272152,'هرومانيا-01,9),('8.416470,'هروومانيا-02,8),('2.412737,'هروومانيا-02,6),('«',1.154888e-02,5),('2.197354,'هارومانيا-02,4),('1.092094,'هروومانيا-02,3),('1.085450,'هرو-02,2),('DIESE',9.221042e-03,1),),(('1.000000,'هروومانيا+00,8),),(('1.000000,'هروومانيا+00,7),),(('!NULL',1.000000e+00,8),),(('!NULL',1.000000e+00,7),),(('1.000000,'هروومانيا+00,4),),(('!NULL',1.000000e+00,5),),((','رومانيا-9.145752e-01,2),('8.543779,'هروومانيا-02,1),),(('!NULL',1.000000e+00,3),),(('!NULL',9.664765e-01,2),('1.3.350966e-02,1),),(('!NULL',1.000000e+00,1),),)
```

Annexe 10 : représentation graphique d'un fichier PLF (C2_Ar_TBKBIS_14)



Annexe 11 : la traduction de fichier C2_Ar_TBKBIS

1 Elle ne constitue pas une victoire du Parti social démocrate roumain, le commandant de la victoire de M. Yoon du Pisco, à l'issue des élections législatives et présidentielles de novembre et décembre de l'année, quelle surprise et a déjà ce siècle lidén Shor la démocratie. Il a indiqué cette victoire lidén l'échec de la politique du gouvernement de coalition de centre-droit sur le Parti du Congrès démocrate roumain. La surprise était le résultat pancarte qui Demande-lui du parti, ennemi notoire pour les étrangers et pour la déclaration de l'éducation civique et de l'état civil. Il a soulevé des solutions de ce parti Deuxièmement depuis la première session des élections présidentielles cas de grande préoccupation. Après trois mois, ne bénéficient pas du gouvernement minoritaire dirigé par M. Adrian ناستاز que la majorité au Parlement. Mais une marge de manœuvre de ces personnes, dont la plupart des critiques à leur statut " les anciens communistes » , en dépit du nom

2 " socialiste démocratique » qui تينوه depuis l'année, est que la direction de la gravité de la situation sociale. Les relations entre la Roumanie et du Fonds monétaire international et la Banque mondiale n'a jamais été aussi bien, si les gouvernements du Parti du Congrès démocrate des trois dernières années de souplesse pendant les négociations, actuellement de Bucarest un accord « chargement » s'étend sur mois, après la suspension de l'accord précédent, organisée par le gouvernement ايزاريكو, sous prétexte de « commencer à accélérer les privatisations. " et que la majorité des nouvelles de tirer parti de décollage économique du cycle ancestrales (croissance économique du cycle ancestrales sociale afin de limiter le déficit budgétaire frontières pour cent et taux d'inflation frontières pour cent, contre plus de pour cent). l'autre de cette pression de l'extérieur, consiste dans les négociations d'adhésion à l'Union

3 Européenne qui respecte ناستاز gouvernement à poursuivre, en dépit des capsules de Bruxelles. En janvier, six seulement des dossiers dimanche, les conditions d'adhésion européenne qui est de la Roumanie satisfait à l'instar des autres pays neuf candidats. De ce fait, le « Plan national d'engagement » pris conditions la Roumanie de sa mise en œuvre pendant la période entre les deux, et de renforcer leurs conditions financières, ce plan sous la pression de la politique économique, dans le même temps de préserver les équilibres importants dans les domaines de la pression externe avec les nouveaux dirigeants à critiquer, économiques, financières et monétaires les prescriptions de médicaments Oman note que ce nationale et le gouvernement est préoccupé par la gestion du processus de réforme .

4 Le double langage de la gouvernance en Roumanie

NB : le fichier **C2_Ar_TBKBIS** se compose de 4 fichiers (paragraphes), la 4eme ligne représente la traduction de fichier **C2_Ar_TBKBIS_14**.

Bibliographie

- Abandah, G., & Anssari, N. (2009). Novel moment features extraction for recognizing handwritten arabic letters. *Journal of Computer Science*, 5 (3), 226.
- Al-Rashaideh, H. (2006). Preprocessing phase for arabic word handwritten recognition. , 6(1).
- Belaïd, A. (2001). Reconnaissance automatique de l'écriture et du document. *Pour la science, disponible sur le lien web: <http://webloria.loria.fr/~abelaid/Publications.html>*.
- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *journal of machine learning research*, 3 (Feb), 1137–1155.
- Bérard, A., Servan, C., Pietquin O. & Besacier, L. (2016) « *MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP* ». In The 10th edition of the Language Resources and Evaluation Conference (LREC 2016). http://www.lifl.fr/~pietquin/pdf/LREC_2016_ABMSOPLB.pdf.
- Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., ... & Ureš, L. (1994, March). The Candide system for machine translation. In Proceedings of the workshop on Human Language Technology (pp. 157-162). Association for Computational Linguistics.
- Bottou, L. (2014). From machine learning to machine reasoning. *Machine learning*, 94(2), 133–149.
- Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., . . . Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16 (2), 79–85.
- Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2), 263–311.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Dyer, C., Muresan, S., & Resnik, P. (2008). *Generalizing word lattice translation* (Tech. Rep.). DTIC Document.
- Federico, M., Bertoldi, N., & Cettolo, M. (2008). Irstlm: an open source toolkit for handling large scale language models. In *Interspeech* (pp. 1618–1621).

- Gahbiche-Braham, S. (2013). *Amélioration des systèmes de traduction par analyse linguistique et thématique: application à la traduction depuis l'arabe* (Unpublished doctoral dissertation). Université Paris Sud-Paris XI.
- Gao, Q., & Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software engineering, testing, and quality assurance for natural language processing* (pp. 49–57). *Interspeech* (Vol. 2002, p. 2002).
- Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, speech, and signal processing, 1995. icassp-95., 1995 international conference on* (Vol. 1, pp. 181–184).
- Knight, K., & Koehn, P. (2003). What's new in statistical machine translation. In *Hlt-naacl* (pp. 5–5).
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521 (7553), 436–444.
- Matusov, E., Leusch, G., Bender, O., Ney, H., et al. (2005). Evaluating machine translation output with automatic sentence segmentation. In *Iwslt* (pp. 138–144).
- Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19–51.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).
- Robadey, L. (2001). *Une methode de reconnaissance structurelle de documents complexes basee sur des patterns bidimensionnels*. University of Fribourg Switzerland
- Schwenk, H. (2010). *Adaptation dun système de traduction automatique statistique avec des ressources monolingues*. TALN, page.
- Stolcke, A., et al. (2002). Srilmm-an extensible language modeling toolkit. In
- Zahra, M. (2006). *Reconnaissance de l'écriture arabe manuscrite a base des machines a vecteurs de support*. Université Badji Mokhtar de Annaba.
- Zaiz, F. (2010). *Les supports vecteurs machines (svm) pour la reconnaissance des caractères manuscrits arabes*. Université Mohamed Khider Biskra.

Liste des figures

Figure 1 :projet TRIDAN, vue d'ensemble	4
Figure 2 : Processus de la traduction automatique statistique	8
Figure 3 : Exemple d'alignement mot-à-mot entre une phrase en français et sa traduction en anglais	9
Figure 4 : Exemple d'alignement d'une phrase en arabe et sa traduction en français.	9
Figure 5 : Exemple d'alignement d'une phrase (Knight et Koehn, 2004).	10
Figure 6 : Structure d'un neurone (Wikipédia)	15
Figure 7 : Modèles de langue neuronal	16
Figure 8 : exemple d'architecture neuronale pour la TA	17
Figure 9 :Processus de production et de reconnaissance de documents.....	18
Figure 10 : Etapes de la reconnaissance de documents (Robadey, 2001)	19
Figure 11 : Scripteur	21
Figure 12 : Schéma général d'un système de reconnaissance de caractères (Zaiz, 2010) ...	22
Figure 13 : Processus de création de notre système de traduction	30
Figure 14 : script RunLM.sh	31
Figure 15 : script RunPhTable.sh	31
Figure 16 :Processus de traduction des sorties OCR (BEST-HYP)	32
Figure 17 :Traitement des mots hors-vocabulaire	35
Figure 18 Traitement des OOV étape 2.....	36
Figure 19 Traitement des OOV étape 3.....	36
Figure 20 : 3 exemples des treillis: (a) phrase; (b) réseaux de confusion (c) treillis non-linéaire	38
Figure 21 : exemple de treillis allemand	39
Figure 22 : Exemple de représentation de Graphe (PLF)	39
Figure 23 :Processus de traduction des Graphes de mots (N-Best).....	40
Figure 24 Format des N-Best fournir par l'OCR (A2IA).....	41
Figure 25 Exemple des Nbest-Srilm.....	42
Figure 26 représentation graphique d'un fichier PLF	43
Figure 27 Exemple de Graphe avant et après traitement des OOV	44
Figure 28 : progression des performances en score Bleu sur le corpus C2	46
Figure 29 : progression des performances en score Bleu sur le corpus C3	47
Figure 30 : progression des performances en score Bleu sur le corpus C4	47
Figure 31 progression des performances en score OOV sur le corpus C2+C3+C4	48

Liste des équations

Équation 1 : Théorème de bayes.....	8
Équation 2 : Recherche de la traduction optimale	8
Équation 3 : Probabilité des segments	11
Équation 4 :Probabilité d'une séquence de mots.....	11
Équation 5 : Fonction de densité utilisée par le décodeur	12

Liste des tableaux

Tableau 1 Corpus Europarl.....	24
Tableau 2 Copus MultiUN	25
Tableau 3 Corpus News-Commentary	25
Tableau 4 Corpus Opensub	25
Tableau 5 Corpus Trame	25
Tableau 6 Corpus Wit3.....	26
Tableau 7 nombre des lignes Copus dev/test.....	26
Tableau 8 Exemple de diacritique obligatoire (position des points)	27
Tableau 9 Exemple de diacritique obligatoire (nombre de points)	27
Tableau 10 diacritiques simples	27
Tableau 11 Diacritiques doubles	28
Tableau 12 Diacritique chadda	28
Tableau 13 Exemple d'utilisation de Tatweel.....	28
Tableau 14 les différents "hamza"	29
Tableau 15 Evaluation de notre premier systeme en score BLEU	32
Tableau 16 Evaluation de nouveau systeme (alignement apres la traduction) en score BLEU.....	33
Tableau 17 Evaluation de nouveau système (alignement avant la traduction) en score BLEU.....	33
Tableau 18 Evaluation de nouveau système en score OOV	34
Tableau 19 Évaluation du nouveau système (traitement des OOV) en score BLEU (alignement apres la traduction)	36
Tableau 20 Evaluation de nouveau système (traitement des OOV) en score OOV	36
Tableau 21 Perplexité des modèles des langues sur le corpus dev (C2+C3+C4)	37
Tableau 22 Évaluation du nouveau système (nouveau modèle de langue - pas de traitement des OOV).....	37
Tableau 23 Évaluation du nouveau système (nouveau modèle de langue - avec traitement des OOV).....	37
Tableau 24 Évaluation du nouveau système (traitement de graphes OCR - sans traitement des OOV).....	43
Tableau 25 Évaluation du nouveau système (traitement de graphes OCR-sans traitement des OOV) en score OOV	43
Tableau 26 Évaluation du nouveau système (traitement de graphes OCR - avec traitement des OOV).....	45
Tableau 27 Évaluation du nouveau système (traitement de graphes OCR - avec traitement des OOV) en score OOV	45

Table des matières

Sommaire.....	
Introduction générale.....	1
1. Contexte du stage	3
1.1. Contexte et objectifs du projet TRIDAN.....	3
1.2. Description du projet TRIDAN	4
1.3. Présentation du porteur et des partenaires	6
2. Etat de l'art.....	7
2.1. Traduction Automatique Statistique (TAS).....	7
2.1.1. Principe de base	7
2.1.2. Les modèles de traduction	8
2.1.2.1. Les modèles de traduction à base de mots	8
2.1.2.2. Les modèles de traduction à base de segments	9
2.1.3. Le modèle de langue.....	11
2.1.4. Le décodage.....	12
2.1.4.1. Moses	13
2.1.5. Évaluation de la traduction automatique	13
2.1.5.1. Evaluation manuelle.....	13
2.1.5.2. Evaluation automatique.....	14
2.2. La traduction automatique neuronale:	14
2.2.1. Réseaux des neurones:.....	14
2.2.2. Modélisation d'un réseau de neurone.....	15
2.2.3. Modèles de langue neuronaux:.....	16
2.2.4. Modèle neuronal pour la Traduction automatique (Séquence vers séquence) : 16	
2.3. Reconnaissance optique des caractères.....	18
2.3.1. Introduction	18
2.3.2. Production et reconnaissance.....	18
2.3.3. Caractéristique de l'OCR	20
2.3.4. Problèmes liés à l'OCR	20
2.3.5. Organisation générale d'un système de reconnaissance	21
2.3.6. Caractéristiques de l'écriture arabe	22
3. Traduction de chaîne de mots.....	24
3.1. Création d'un système de traduction arabe/français	24
3.1.1. Description des corpus	24

3.1.2.	Prétraitement des corpus.....	26
3.1.3.	Création de notre système.....	30
3.2.	Traduction de la meilleure hypothèse d'OCR	32
3.3.	Traitement des mots hors vocabulaire	34
3.4.	Nouveau modèle de langue (Interpolation linéaire).....	36
4.	Traduction de graphes	38
4.1.	Treillis des mots.....	38
4.2.	Chaîne de traduction des graphes	40
4.3.	Traitement des mots hors vocabulaire	44
5.	Bilan	46
Conclusion et perspectives		49
Conclusion		49
Perspectives		50
Annexes		51
Annexe 1 :	Extrait de sortie OCR (Corpus C2)	51
Annexe 2 :	Extrait de la traduction de sortie OCR (Corpus C2)	52
Annexe 3 :	Extrait de la traduction de sortie OCR (Corpus C2) après traitement des mots hors vocabulaire	53
Annexe 4 :	exemple de fichier N-best de sortie OCR (C2_Ar_TBKBIS_14).....	54
Annexe 5 :	exemple fichier N-best SRILM (fichier C2_Ar_TBKBIS_14).....	55
Annexe 6 :	exemple fichier N-best WLAT (fichier C2_Ar_TBKBIS_14).....	56
Annexe 7 :	exemple fichier N-best PFSG (fichier C2_Ar_TBKBIS_14)	57
Annexe 8 :	exemple fichier HTK (fichier C2_Ar_TBKBIS_14)	58
Annexe 9 :	exemple fichier PLF (fichier C2_Ar_TBKBIS_14).....	60
Annexe 10 :	représentation graphique d'un fichier PLF (C2_Ar_TBKBIS_14).....	61
Annexe 11 :	la traduction de fichier C2_Ar_TBKBIS.....	62
Bibliographie		63
Liste des figures.....		65
Liste des équations		65
Liste des tableaux		66
Table des matières		67

MOTS-CLÉS : Traitement automatique des langues naturelles, Traduction Automatique Probabiliste, Graphes de mots (treillis de mots), plongement de mots, reconnaissance optique de caractères.

RÉSUMÉ

Durant ces dernières années, le domaine du traitement automatique des langues naturelles (TALN) a connu des évolutions rapides, et spécialement la recherche des informations dans les documents numérisés, qui nécessite deux domaines connexes : la reconnaissance optique de caractères et la traduction automatique. Dans ce mémoire de recherche nous nous sommes intéressés à la traduction automatique des documents numérisés, soit manuscrite ou typographie. En premier lieu nous avons créé un système de traduction arabe français. En second lieu nous avons amélioré notre système par la traduction des graphes de mots qu'ils sont construits à partir des sorties OCR bruitées (N-best). Et en dernier lieu nous avons ajouté un traitement spécifique de mots hors vocabulaire en se basant sur l'approche de plongement de mots (word2vec).

KEYWORDS: Natural language processing, Statistical Machine Translation, Word lattice, Word Embedding, Optical character recognition.

ABSTRACT

During the last few years, the field of automatic processing of natural languages has seen rapid developments, especially in research on scanned documents, which involves two interrelated fields: Optical character recognition (OCR) and machine translation (MT). In this research paper, we investigate machine translation of scanned documents, whether handwritten or typed. First of all, we created a Arabic-French machine translation system. Next we improved our system by translating a words lattice constructed from noisy OCR outputs (N-best). Finally, we added a specific preprocessing for out-of-vocabulary (OOV) words using word-embeddings (word2vec).