Traitement des Mots Hors Vocabulaire pour la Traduction Automatique de Document OCRisés en Arabe

Kamel Bouzidi[1], Zied Elloumi[1, 3], Laurent Besacier[1], Benjamin Lecouteux[1], Mohamed-Faouzi Benzeghiba[2]

(1) LIG, Univ. Grenoble Alpes, Campus Saint-Martin d'Hères, Grenoble, France
(2) A2IA, 39 Rue de la Bienfaisance, 75008 Paris, France
(3) LNE, F-78190 Trappes, France

<prénom.nom>@univ-grenoble-alpes.fr[1]
<faouzi.benzeghiba>@a2ia.com[2]
<prénom.nom>@lne.fr[3]

RÉSUMÉ

Cet article présente un système original de traduction de documents numérisés en arabe. Deux modules sont cascadés : un système de reconnaissance optique de caractères (OCR) en arabe et un système de traduction automatique (TA) arabe-français. Le couplage OCR-TA a été peu abordé dans la littérature et l'originalité de cette étude consiste à proposer un couplage étroit entre OCR et TA ainsi qu'un traitement spécifique des mots hors vocabulaire (MHV) engendrés par les erreurs d'OCRisation. Le couplage OCR-TA par treillis et notre traitement des MHV par remplacement selon une mesure composite qui prend en compte forme de surface et contexte du mot, permettent une amélioration significative des performances de traduction. Les expérimentations sont réalisés sur un corpus de journaux numérisés en arabe et permettent d'obtenir des améliorations en score BLEU de 3,73 et 5,5 sur les corpus de développement et de test respectivement.

ABSTRACT

This article presents a new system that automatically translates images of arabic documents. Two modules are involved: an optical character recognition (OCR) module in Arabic and an Arabic-French machine translation module (MT). The OCR-MT coupling has not been much studied in the literature previously and the originality of this work consists in proposing a close coupling between OCR and MT as well as a specific processing of out-of-vocabulary (OOV) words due to OCR errors. The OCR-MT coupling based on an hypothesis lattice, as well as our OOV processing by replacement (according to a composite measure that takes into account surface form and context of the word) allow a significant improvement in translation performance. Our experiments are carried out on a challenging corpus of arabic newspapers digitized and we obtain BLEU improvements of 3,73 and 5,5 on our development and test corpora respectively.

MOTS-CLÉS: Traitement Automatique des Langues Naturelles, Traduction Automatique Probabiliste, Mots Hors Vocabulaire, Graphes de Mots, Plongements de Mots, Reconnaissance Optique de Caractères.

KEYWORDS: Natural Language Processing, Statistical Machine Translation, Out-Of-Vocabulary Words, Word Lattice, Word Embeddings, Optical Character Recognition.

1 Introduction

Ce travail est effectué dans le cadre du projet TRIDAN sur la recherche d'informations dans les documents numérisés multilingues. L'application visée s'adresse à des utilisateurs souhaitant réaliser une requête de recherche d'informations dans leur langue sur des documents (initialement sous forme papier puis numérisés) dans une langue qu'ils ne connaissent pas (recherche d'information cross-lingue à partir de documents papiers numérisés). Ce projet vise à mettre au point des techniques permettant un couplage innovant entre la lecture automatique de documents numérisés (LAD), la traduction automatique et l'extraction d'informations. Cependant, dans cet article, nous considérons seulement le problème du couplage entre les deux premières étapes (traduction automatique de documents OCRisés[1]). Chacune de ces techniques prise séparément a fait récemment l'objet d'avancées technologiques importantes, mesurées par des évaluations internationales et par leur utilisation de plus en plus fréquente 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 des différentes étapes.

Les techniques de traduction automatique et d'extraction d'information sont généralement développées pour s'appliquer à 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 optique de caractères manuscrits ou typographiés, nécessite la prise en compte des erreurs et des incertitudes dans la suite de la chaîne de traitement. Ces erreurs d'OCRisation peuvent conduire à l'ajout de mots erronés (qui sont donc considérés comme *hors vocabulaire*) dans les données à traduire, augmentant ainsi la difficulté de la tâche.

Les mots hors vocabulaire considérés ici sont donc soit dus à des erreurs d'OCRisation (erreurs de transcription), soit simplement inconnus de notre système de traduction (dus à une couverture insuffisante du corpus d'entraînement de notre système de TA). En revanche, comme nous allons utiliser des plongements de mots dans cet article, il convient de préciser que l'approche proposée ici ne sera pas capable de prendre en compte un autre type de mots inconnus : ceux qui ne possèdent pas de représentation distribuée.

Contributions. Dans cet article, nous proposons un traitement spécifique des mots hors vocabulaire (selon notre définition ci-dessus) en les remplaçant par des mots proches – connus du système de traduction automatique (proches selon un critère composite entre forme de surface et contexte d'utilisation). Par ailleurs, nous prenons en compte l'ambiguité du traitement OCR via un graphe de mots qui est proposé en entrée du système de traduction automatique. La combinaison de ces deux approches (remplacement des mots hors vocabulaire et traduction d'un treillis d'hypothèses) permet des améliorations importantes de la qualité de traduction arabe-français mesurée sur un corpus de journaux en arabe numérisés et OCRisés.

Plan. Après une section 2 consacrée à l'état de l'art, nous présentons les corpus et systèmes construits dans la section 3. La section 4 concerne le cœur de notre approche : le traitement des mots hors vocabulaire. La section 5 est quant à elle consacrée aux résultats expérimentaux tandis que la section 6 présente une rapide conclusion.

[1] OCR signifie « Optical Character Recognition ». Nous nous autorisons dans cet article à utiliser les termes OCRisation et OCRisés qui sont largement utilisés dans la communauté de la lecture automatique de documents numérisés.

2 État de l'art

2.1 Interfaçage OCR et TA (Traduction Automatique)

De nombreuses recherches ont été consacrées au couplage transcription-traduction mais ceci plutôt dans le domaine de la traduction de parole où une transcription bruitée, issue du module de reconnaissance vocale (ou transcription automatique) est fournie en entrée du système de TA. On peut citer les travaux suivants : (Zhou et al., 2007 ; Déchelotte et al., 2007 ; Besacier et al., 2010) sur le couplage entre la transcription de parole et la traduction. Ils consistent notamment à : (a) traduire un treillis de mots ou une chaîne de N meilleures hypothèses plutôt qu'une seule hypothèse de transcription ; (b) remettre en forme la sortie de transcription pour la rendre plus «simple» à traduire (re-ponctuation, re-capitalisation). L'article de synthèse de Segal et al. (2015) est une revue détaillée des problèmes principaux à résoudre en traduction de la parole, appliquée au cas des traductions de conférences (*TED Talks*). Les auteurs quantifient l'impact des erreurs de transcription, de segmentation et de reponctuation sur les performances des systèmes de traduction de parole (écart de plus de dix points BLEU sur la qualité de la traduction automatique par rapport à une traduction de l'écrit) et pointent les voies d'amélioration les plus prometteuses.

En ce qui concerne l'interfaçage OCR-TA, on trouve beaucoup moins de travaux mais on peut tout de même citer (Afli et al., 2015 ; Afli & Way 2016) qui propose de corriger les erreurs d'OCR en utilisant un système de post-édition automatique fondé sur la traduction probabiliste à base de fragments. Le couplage de la reconnaissance de documents et de la traduction automatique est cependant abordé dans le projet MADCAT financé par la DARPA (Macrostie & al., 2010).

Les principales différences entre une hypothèse de transcription de parole et une hypothèse OCR sont les suivantes : même si les systèmes d'OCR utilisent aussi un modèle de langue, en plus du modèle optique, les erreurs d'OCR sont souvent proches du mot de référence « à un ou deux caractères près », tandis que les erreurs de transcription de parole sont issues d'une interaction complexe entre modèle de langue et modèle acoustique. Il nous semble donc possible de corriger certaines erreurs d'OCR en remplaçant les mots mal reconnus par des mots « proches » connus du système de traduction. En ce qui concerne la mise en forme, les problèmes de re-capitalisation et de re-ponctuation ne se posent pas a priori pour l'OCR. Cependant, en nous inspirant des approches de couplage en traduction de parole, nous proposons d'utiliser un graphe de mots issu de l'OCR pour prendre en compte l'ambiguité de la transcription.

2.2 Traitement des mots hors vocabulaires en TA

Dans le domaine de traduction automatique statistique, les mots hors vocabulaire (MHV) sont les mots inconnus du système de traduction, c'est à dire n'ayant pas été rencontrés dans les données d'apprentissage. Les mots hors vocabulaire sont responsables de la dégradation de la performance d'un système de TA car, en plus de ne pas être traduits, ils influent négativement sur la traduction des mots voisins et par la suite sur la traduction de toute la phrase.

Le traitement des mots hors vocabulaire est largement abordé dans les systèmes de reconnaissance automatique de parole (SRAP) (Bousquet-Vernhettes, 2002; Lecouteux et al., 2009 ; Kombrink et al., 2012 ; Bazzi, 2002) et dans les systèmes de traduction automatique (Oprean et al. ,2014 ; Habash , 2008 ; Marton et al., 2009). La plupart de ces travaux consistent à augmenter la

couverture du corpus d'entraînement, tandis que, dans notre travail, nous essayons de traiter plus spécifiquement les mots hors vocabulaire issus du système OCR.

2.3 Représentations distribuées de mots ou plongements de mots

Nous introduisons les représentations distribuées de mots (ou plongements ou encore *word embeddings* en anglais) car elles vont nous servir à déterminer un indicateur de proximité entre mot inconnu issu de l'OCR et mot candidat à son remplacement. Ces représentations des mots dans un espace continu ont pris une place centrale dans la recherche en traitement automatique du langage naturel ces dernières années (Bengio et al., 2003; Turian et al., 2010; Collobert et al., 2011; Huang et al., 2012 ; Servan et al., 2016). Les représentations distribuées de mots ont été beaucoup utilisées pour la TA statistique (Cho et al. , 2014), les systèmes de questions-réponses (Belinkov et al. , 2015), la recherche d'informations (Shen et al. , 2014), par exemple.

L'idée principale est que la représentation d'un mot peut être obtenue en fonction de son contexte (les mots qui l'entourent) (Baroni & Zamparelli, 2010). Les mots sont projetés vers un espace continu de dimension prédéfinie et les mots ayant des contextes similaires sont, de fait, proches dans cet espace continu. Comparées à une représentation discrète, les représentations distribuées permettent d'induire des relations morpho-syntaxiques et sémantiques (Blacoe & Lapata, 2012; Mikolov et al., 2013). Par ailleurs, des similarités entre les mots peuvent être calculées en utilisant les représentations vectorielles via, par exemple, une distance de type *cosinus*. Dans nos expériences, nous proposons d'utiliser une des représentations les plus employées : celle proposée par Mikolov et al. (2013a). Dans cette représentation, deux modèles (CBOW et SKIPGRAM) sont proposés pour apprendre les relations entre les mots selon leur contexte (Mikolov et al., 2013).

3 Traduction de documents OCRisés en arabe

3.1 Corpus pour la traduction automatique

Apprentissage :

Pour l'entraînement de notre système de traduction nous avons utilisé des corpus bilingues parallèles arabe-français similaires à ceux utilisés par Mirkin & al. (2014) (MultiUN, News-Commentary, Opensub, Trame, Wit3) et un corpus monolingue français supplémentaire pour le modèle de langue cible (Europarl). Les statistiques de ces corpus sont données dans le *tableau 1*. *Europarl* est un corpus de transcriptions et traductions de séances parlementaires, *MultiUN* est extrait du site Web des nations-unies après avoir été nettoyé par le laboratoire DFKI, *News Commentary* contient des extraits de diverses publications de presse, *Opensub* correspond à des sous-titrages de films, Trame correspond à environ 90 heures des discours radio et télévises en arabe enregistrés, transcrits et ensuite traduits en français, et enfin *Wit3* est une collection de séminaires transcrits et traduits (*TED talks*).

Corpus	Nombre des phrases	Nombre des mots	
		AR	FR
Europarl	2 007 723	0	52 525 000
MultiUN	9 929 567	222 387 310	285 520 384
News Commentary	90 753	2 180 814	2 372 649
Opensub	4 381 835	27 739 977	32 269 908
Trame	20 539	546 257	758 030
Wit3	87 732	1 946 275	2 436 720

Tableau 1: Corpus d'apprentissage de notre système de TA arabe-français

Développement et Test :

Pour les corpus de développement et de test nous avons utilisé des documents de type *journaux* extraits de la base MAURDOR[2] auxquels nous avons eu accès. Ces documents sont des images numérisées de journaux en arabe. Pour ces images, on dispose de la transcription exacte et d'une transcription automatique (taux d'erreur mots d'environ 15 % du système OCR). La transcription exacte a également été traduite manuellement vers le français par deux traducteurs. On disposera donc de deux références pour évaluer la qualité des traductions. Le tableau 2 montre le nombre de phrases pour les corpus de développement et de test. Il est important de noter que l'OCRisation induit une segmentation en phrases qui peut être différente de la segmentation de référence (marque de ponctuation mal reconnue ou introduite par erreur). Nous discuterons des conséquences de ce problème pour l'évaluation plus loin dans le texte.

Corpus	#phrases (Dev)	#phrases (Tst)
Journaux Ref OCR	250	267
Journaux Hyp OCR	282	295

Tableau 2: Corpus de développement et de test de journaux numérisés en arabe et traduits en français.

3.2 Prétraitements

Suppression des diacritiques. Comme tous les corpus à notre disposition ne contenaient pas de textes voyellés, nous avons supprimé les diacritiques de nos corpus d'apprentissage en arabe.
Suppression du Tatweel. Le Tatweel est un effet typographique utilisé avec les systèmes d'écriture arabe pour allonger les caractères mais il ne change pas le sens de mots. Nous l'avons donc supprimé de tous nos corpus.
Normalisation du caractère Hamza. Le caractère *Hamza (ء)*peut prendre plusieurs formes et a été normalisé de آ ou أ ou إ vers ا.

[2] http://www.maurdor-campaign.org

Autres prétraitements. Correction des caractères mal-encodés, tokenisation côté arabe avec l'outil Opennlp[3], tokenisation côté français avec le script `tokenizer.pl`[4] (fourni avec la suite Moses), normalisation de la ponctuation et des caractères spéciaux avec le script `normalize-punctuation.perl`[5] (fourni avec la suite Moses), transformation de toutes les données en minuscule et vers un encodage UTF-8, suppression des phrases plus longues que 100 mots.

3.3 Système OCR utilisé

Le système de transcription de texte arabe est un modèle hybride neuronal/Markovien développé par A2IA[6]. Il utilise un réseau neuronal récurrent (RNN) multidimensionnel qui intègre des cellules de type LSTM (Long Short Term Memory) (Hochreiter & Schmidhuber, 1997) (Graves 2012; Moysset et al., 2014). L'architecture utilise des couches de convolution pour apprendre les bonnes caractéristiques pour la tâche de la reconnaissance de caractères. Le réseau est entraîné avec une fonction de coût de type CTC (*Connectionnist Temporal Classification*) et ne nécessite pas une segmentation explicite des données en termes de caractères.

Le système intègre un modèle de langue hybride mots/PAW (*Part-of-Arabic Word) (BenZeghiba et al., 2015)* qui sont des unités sous-lexicales d'ordre 3. Il est entraîné par interpolation sur plusieurs sources textuelles. L'espace de décodage est défini comme un automate d'états finis pondérés (Mohri 1997) et l'algorithme de Viterbi est utilisé pour trouver la ou les meilleure(s) hypothèse(s). Ce système d'OCRisation obtient un taux d'erreur de mots de 15.5% et 16.5% sur nos corpus *Dev* et *Test* respectivement. En sortie d'OCR, la taille du vocabulaire mesurée sur *Dev* et *Test* est de 3329 et 3391 mots respectivement.

3.4 Système de traduction automatique probabiliste

Le système de traduction utilisé est un système probabiliste à base de fragments (*phrase-based*) construit à partir de la boîte à outils open source *moses* (Koehn et al., 2007). Nous utilisons également IRSTLM (Federico et al., 2008) pour créer un modèle de langue 5-gramme en français qui résulte de l'interpolation entre des modèles appris sur chaque sous-corpus mentionné dans le *tableau 1*.

Après avoir créé le modèle de traduction (paramètres par défauts avec heuristiques d'alignement et de modèle de re-ordering obtenus avec les options `-alignment grow-diag-final-and -reordering msd-bidirectional-fe`), nous utilisons le corpus de développement pour optimiser les poids du modèle log-linéaire avec l'outil MERT (*Minimum Error Rate Training*) (Och, 2003). Ce système de traduction donne comme résultat sur les références OCR un score BLEU (Papineni &al, 2002) de 29,79 sur le corpus *Dev* et de 26,46 sur le corpus *Test*. Ce score BLEU est évalué avec les 2 références disponibles sur les corpus *Dev* et *Test*.

Le système de reconnaissance optique de caractères produit une liste de N meilleures hypothèses que nous exploitons en la transformant en un graphe de mots. En effet, *Moses* est capable de décoder une entrée représentée sous forme des graphes de mots (Dyer et al., 2008).

Pour convertir les N meilleures sorties du système OCR en graphe de mots au format PLF lisible par *Moses*, nous passons par les étapes suivantes : conversion vers le format N-best_SRILM, puis

[3] opennlp.apache.org
[4] www.statmt.org/wmt08/scripts.tgz (scripts/tokenizer.perl)
[5] www.statmt.org/wmt11/normalize-punctuation.perl
[6] www.a2ia.com

utilisation de l'outil SRILM (Stolcke, 2002) pour générer un graphe au format HTK et enfin utilisation de l'outil `htk2plf`[7] pour convertir les fichiers HTK en graphe des mots au format PLF.

La traduction de graphes de mots introduit un nouveau paramètre dans le modèle log-linéaire et une nouvelle optimisation des poids du modèle est alors nécessaire. Ceci est réalisé à nouveau avec l'algorithme de minimisation du taux d'erreur MERT sur le corpus de *développement*. Les poids trouvés à l'issue de cette optimisation sont ensuite appliqués à l'identique pour décoder le corpus de *test*.

4 Traitement des mots hors vocabulaire

Nous rappelons que les mots hors vocabulaire (MHV) peuvent être dus à des erreurs d'OCR ou a une couverture non suffisante de notre corpus d'entraînement. La solution que nous proposons en 4.1 a pour but de limiter l'impact des seules erreurs de l'OCR, tandis que le traitement en 4.2 est plus général.

4.1 Traitement des MHV sur les chaînes de mots

Notre approche consiste à chercher et remplacer les mots hors vocabulaire par des mots proches connus de notre système de traduction. Notre critère de remplacement est le suivant :

- Les mots qui remplaceront le MHV doivent avoir une forme de surface proche de celle du mot inconnu (nous utiliserons pour cela une distance d'édition au niveau des caractères),

- Les mots qui remplaceront le MHV doivent avoir un contexte d'utilisation proche de celui du mot inconnu (nous utiliserons pour cela une distance cosinus entre représentations distribuées des mots).

Plus précisément, notre méthode se décompose en trois étapes :

1- Créer un modèle vectoriel de mots (par exemple avec l'outil *Word2Vec*[8]) à partir de toutes les données disponibles en arabe (apprentissage, développement et test) concaténées. La seule contrainte imposée par cette étape est que le corpus à traduire doit être connu au départ, afin que tous les mots *source* à traduire possèdent une représentation distribuée.

2- Pour chaque mot hors vocabulaire des corpus *dev* et *test (*c'est à dire *pour tout mot non connu du système de TA, donc non présent dans le corpus d'apprentissage décrit dans le tableau 1)*, trouver la liste des *L* mots les plus proches (dans le corpus *apprentissage*) selon une distance d'édition au niveau des caractères (type distance de *Levenshtein*). La distance d'édition retourne un entier noté n et nous gardons comme candidats tous les mots dont la distance d'édition est inférieure ou égale à 2 (mots proches du MHV en terme de forme de surface). Un exemple est donné dans la figure 1 pour un MHV en arabe.

[7]Développé par Sylvain Raybaud du LORIA
[8]https://code.google.com/archive/p/word2vec/

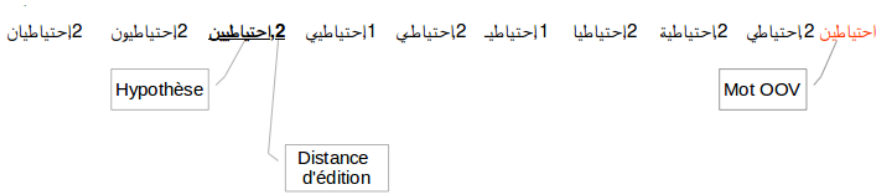

Figure 1 : Traitement des MHV (étape 2)

3- Trouver le meilleur mot qui remplacera le MHV parmi la liste *L* en calculant la *distance cosinus* entre le MHV et chaque mot de la liste *L* grâce à une représentation distribuée. Un exemple est donné dans la figure 2 où l'on voit le mot le plus proche retrouvé en fonction de sa distance cosinus par rapport au mot MHV dans le corpus d'apprentissage (notre distance *cosinus* est comprise entre 0 et 2 puisqu'elle est égale à *1-similarité cosinus*).

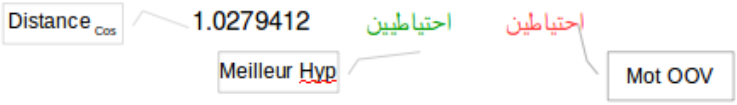

Figure 2 : Traitement des MHV (étape 3)

4.2 Traitement des MHV sur les graphes de mots

Le critère de remplacement fondé sur une mesure de similarité composite (pris en compte de la forme de surface **et** du contexte) reste le même. Cependant, le fait d'avoir un graphe nous permet d'ajouter plus d'un mot candidat au remplacement du MHV. Ainsi, dans l'étape 3, les M mots les plus proches en fonction de la distance cosinus par rapport au MHV sont sélectionnés (M est fixé empiriquement à 4 dans les expériences reportées plus loin).

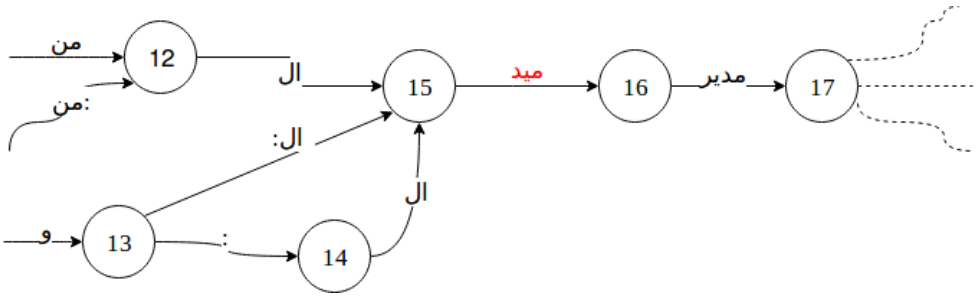

Figure 3 : Exemple de graphe avant traitement des MHV

Les figures 3 et 4 montrent un exemple ou l'arc correspondant à un mot inconnu (en rouge – figure 3) est augmenté par 4 autres arcs passant par les 4 meilleurs candidats au remplacement du mot inconnu selon la procédure décrite dans la partie 4.1 (voir figure 4). Il est important de noter que dans ce cas, le MHV est conservé dans un chemin du graphe et nous laissons le décodeur de TA choisir quel chemin sera le plus prometteur au final.

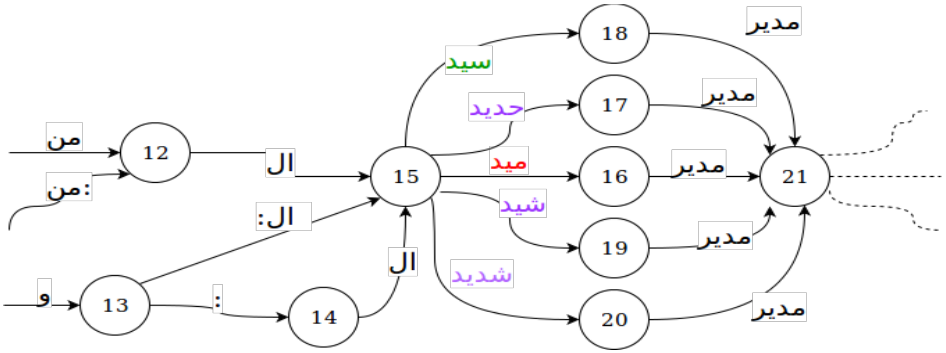

Figure 4 : Exemple de graphe après traitement des MHV

5 Expérimentation et résultats

Dans cette section, nous présentons nos résultats de traduction automatique arabe-français de documents OCRisés avec et sans traitement des MHV. Il est important de souligner que pour pouvoir évaluer nos sorties de traduction il faut avoir le même nombre de lignes pour les sorties de traduction (hypothèses) et pour les traductions manuelles (références). Cependant, dans notre cas, les sorties du système OCR sont utilisées pour obtenir les hypothèses tandis que nos références sont obtenues à partir des transcriptions manuelles des documents. Comme le montre le tableau 2, on n'a donc pas forcément le même nombre de lignes entre hypothèse et référence. Une phase d'alignement entre les sorties de traduction automatique et les références en français (traductions manuelles) est donc nécessaire. Le même genre de problème peut se poser en traduction de la parole (phrases bien formés en *référence* et tours de parole issus d'une segmentation automatique en *hypothèse*). Ainsi, nous profitons de la disponibilité de l'outil *MwerAlign* (Matusov et al. 2005) qui est utilisé pour l'évaluation des traductions de la parole lorsque il y a une dfférence de segmentation entre hypothèses traduites et référence. Cette nouvelle phase de post-traitement avant d'évaluer le score BLEU introduit une dégradation légère des résultats (car le re-alignement par *MwerAlign* n'est pas toujours parfait) mais elle est nécessaire à la bonne évaluation des performances de la chaîne complète de traduction.

Le *tableau 3* ci-dessous montre l'évaluation de différentes sorties de notre système de traduction de documents OCRisés en arabe avec ou sans traitement des mots hors vocabulaire et avec ou sans traduction des graphes d'hypothèses.

Système	Dev (BLEU)	Test (BLEU)	Dev (%MHV)	Test (%MHV)
Baseline	21,17	16,41	1,96 %	2,00 %
+ traitement MHV	21,34	16,59	1,04 %	1,04 %
+Graphe	24,63	21,75	1,43 %	1,62 %
+traitement MHV +Graphe	**24,90**	**21,91**	**0,20 %**	**0,20 %**

Tableau 3: Performances (score BLEU) de traduction de documents OCRisés en arabe avec ou sans traitement des mots hors vocabulaire et avec ou sans traduction de graphes. La traduction du texte arabe exacte (pas d'erreur OCR) donne un BLEU de 29,79 sur le dev et 26,46 sur le test.

Les résultats montrent que notre approche de traitement de mots hors vocabulaire améliore les résultats de traduction mais que c'est surtout la traduction de graphes d'hypothèses qui permet des gains de performance importants. A notre connaissance, peu de travaux antérieurs ont démontré l'efficacité du couplage étroit entre module OCR et module de traduction. Ce résultat est donc la principale contribution de l'article. Le traitement des MHV obtient également des gains encourageants mais il semble que le système remplace parfois à tort les mots inconnus. Par ailleurs, on mesure que la perte due aux erreurs d'OCR (dont le taux d'erreurs mots était de 15 % environ) donne une dégradation de 5 points de BLEU (de 29,79 à 24,90 sur le *Dev* et de 26,46 à 21,91 sur le *Test*). Si on compare aux travaux de Segal & al. (2015) sur la traduction de parole (écart de plus de dix points BLEU sur la qualité de la traduction automatique par rapport à une traduction de l'écrit), il semble que les erreurs de l'OCR soient un peu plus faciles à corriger que celles issues de la traduction de la parole. Cette dernière affirmation nécessite cependant d'être confirmée par de plus larges investigations.

Les exemples ci-dessous présentent des sorties de traduction avant et après le traitement des MHV.

-Les colons à la protection de l'armée que de nombreuses dans les territoires occupés à trois reprises, notamment en envoyant des **احتياطين.** - …/... si elle **ايحرك** les débats publics et de gérer les affaires du Conseil le contenu soit présent ou **غاتبا**, derrière le Conseil en 2011 **مجتمعاسياسيا** et un autre civil et médiatiquement …/...
-Les colons à la protection de l'armée que de nombreuses dans les territoires occupés à trois reprises, notamment en envoyant des **réservistes**. - …/... 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 **مجتمعاسياسيا** et un autre civil et médiatiquement …/...

Au final, après combinaison graphe+traitement MHV, les performances de notre système de traduction arabe-français progressent au niveau du score BLEU de 5,5 pts sur le *Test (cette amélioration est significative).* Les exemples ci-dessous comparent les sorties de traduction de chaînes de mots et la traduction de graphe des mots.

Ref	Certaines entreprises de presse représentent elles une menace pour la **paix** mondiale ?
Hyp 1-best	Certaines institutions médiatiques une menace pour la **communauté** mondiale ?
Hyp Graphe	Certaines institutions médiatiques une menace pour la **paix** mondiale ?

Ref	C'est aussi une langue rituelle dans un nombre d'églises chrétiennes dans le monde arabe - pierre **abramovitch**
Hyp 1-best	Aussi divisée principal auprès de plusieurs églises chrétiennes dans le monde arabe

	- pierre **ابراموفيتشي**
Hyp Graphe	Aussi divisée principal auprès de plusieurs églises chrétiennes dans le monde arabe - pierre **abramovich**

Ref	Afin d'instaurer une paix sociale la gauche a accepté par exemple de faire des concessions aux **syndicats** et d'accorder des augmentations de salaire …/...
Hyp 1-best	Pour l'achat de la paix sociale a approuvé la gauche par exemple la renonciation **للنفقابات** augmentation des salaires …/...
Hyp Graphe	Pour l'achat de la paix sociale a approuvé la gauche par exemple la renonciation des **syndicats** sur des augmentations salariales …/...

6 Conclusion

Dans cet article, nous nous sommes intéressés à la traduction automatique des sorties d'un système OCR et avons proposé une amélioration par l'utilisation de graphes d'hypothèses OCR et le traitement des mots hors vocabulaire (MHV). Notre technique de traitement des MHV est fondée sur l'observation que les erreurs d'OCRisation sont souvent des substitutions de mots similaires en terme de forme de surface. Nous proposons donc de remplacer les mots inconnus du système de traduction, par des mots similaires selon un score composite qui prend en compte la similarité des formes de surface et le contexte d'utilisation des mots (ce dernier étant mesuré par une distance cosinus entre plongements de mots). Nous avons montré aussi l'efficacité d'un couplage étroit entre OCR et traduction par utilisation de graphes de mots dans notre chaîne de traduction. Ce couplage permet d'obtenir des améliorations importantes du score BLEU mesuré avec deux références tandis que notre méthode originale de traitement des MHV permet des gains additionels (mais plus faibles). D'autres pistes d'améliorations pourraient être envisagées. Dans ce travail nous avons utilisé l'outil *Moses* pour créer un système de traduction probabiliste à base de fragments mais dans un futur proche, il est envisageable d'utiliser un système de traduction neuronal. Ceci semble d'autant plus pertinent que le système d'OCRisation est lui-même fondé sur un réseau de neurones récurrent (RNN). L'utilisation d'un système de traduction neuronal ouvre donc la possibilité à d'autre types de couplage innovants entre OCR et TA. Notre traitement des MHV est également perfectible puisqu'il remplace actuellement tous les mots inconnus selon notre méthode, alors que certains ne sont pas dûs à des erreurs d'OCR mais simplement à une couverture insuffisante de notre corpus d'apprentissage de TA.

Références

Afli, H., & Way, A. (2016). Integrating Optical Character Recognition and Machine Translation of Historical Documents. *LT4DH 2016*, 109.

Afli, H., Barrault, L., & Schwenk, H. (2015). OCR Error Correction Using Statistical Machine Translation. In *16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt*.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1183-1193). Association for Computational Linguistics.

Bazzi, I. (2002). *Modelling out-of-vocabulary words for robust speech recognition* (Doctoral dissertation, Massachusetts Institute of Technology).

Belinkov, Y., Mohtarami, M., Cyphers, S., & Glass, J. (2015). VectorSLU: A continuous word vector approach to answer selection in community question answering systems. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 282-287).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3* (Feb), 1137-1155.

BenZeghiba, M. F., Louradour, J., & Kermorvant, C. (2015, August). Hybrid word/Part-of-Arabic-Word Language Models for arabic text document recognition. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on* (pp. 671-675). IEEE.

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)*.

Besacier, L., Afli, H., Do, T. N. D., Blanchon, H., & Potet, M. (2010). LIG statistical machine translation systems for IWSLT 2010. In *International Workshop on Spoken Language Translation IWSLT 2010).*

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 546-556). Association for Computational Linguistics.

Bousquet-Vernhettes, C. (2002). Traitement des mots mal reconnus en compréhension de la parole. *Journées d'Études sur la Parole (JEP'2002)*, 309-312.

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*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*(Aug), 2493-2537.

Déchelotte, D., Schwenk, H., Adda, G., & Gauvain, J. L. (2007). Improved machine translation of speech-to-text outputs. In *Interspeech 2007* (Vol. 7, pp. 2441-2444).

Dyer, C., Muresan, S., & Resnik, P. (2008). *Generalizing word lattice translation* (No. LAMP-TR-149). *Internal Report Maryland University College Park Institute for Advanced Computer Studies.* 2008.

Federico, M., Bertoldi, N., & Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008* (pp. 1618-1621).

Graves, A. (2012). Offline arabic handwriting recognition with multidimensional recurrent neural networks. In *Guide to OCR for Arabic scripts* (pp. 297-313). Springer London.

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (pp. 57-60). Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 873-882). Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Dyer, C. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 177-180). Association for Computational Linguistics.

Kombrink, S., Hannemann, M., & Burget, L. (2012). Out-of-vocabulary word detection and beyond. In *Detection and Identification of Rare Audiovisual Cues* (pp. 57-65). Springer Berlin Heidelberg.

Lecouteux, B., Linarès, G., & Favre, B. (2009). Détection de mots hors-vocabulaire par combinaison de mesures de confiance de haut et bas niveaux. *MajecSTIC'09*.

Macrostie, E., Rawls, S., and Subramanian, K. (2010). The BBN Document Analysis Service : A Platform for Multilingual Document Translation. In *Document Analysis Systems*, pages 447-453, 2010.

Marton, Y., Callison-Burch, C., & Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 381-390). Association for Computational Linguistics.

Matusov, E., Leusch, G., Bender, O., & Ney, H. (2005). Evaluating machine translation output with automatic sentence segmentation. In *IWSLT 2005* (pp. 138-144).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Mirkin, S., & Besacier, L. (2014). Data selection for compact adapted SMT models. In *Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational linguistics*, *23*(2), 269-311.

Moysset, B., Bluche, T., Knibbe, M., Benzeghiba, M. F., Messina, R., Louradour, J., & Kermorvant, C. (2014). The A2iA multi-lingual text recognition system at the second Maurdor evaluation. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on* (pp. 297-302). IEEE.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 160-167). Association for Computational Linguistics.

Oprean, C., Likforman-Sulem, L., Popescu, A., & Mokbel, C. (2014). Utilisation du Web pour la reconnaissance de mots manuscrits hors vocabulaire. In *CORIA-CIFED* (pp. 217-232).

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318.

Segal N., Bonneau-Maynard H., Yvon F. (2015). « Traduire la parole : le cas des TED Talks », *Traitement automatique des langues (TAL)*, 2015.

Servan, C., Elloumi, Z., Blanchon, H., & Besacier, L. (2016). Word2Vec vs DBnary ou comment (ré) concilier représentations distribuées et réseaux lexico-sémantiques? Le cas de l'évaluation en traduction automatique. In *TALN 2016*.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 101-110). ACM.

Stolcke, A. (2002, September). SRILM - an extensible language modeling toolkit. In *Interspeech 2002* (Vol. 2002, p. 2002).

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394). Association for Computational Linguistics.

Zhou, B., Besacier, L., & Gao, Y. (2007). On efficient coupling of ASR and SMT for speech translation. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* (Vol. 4, pp. IV-101). IEEE.