Handwriting-OOV word-recognition using web resources

Cristina Oprean, Chafic Mokbel, Laurence Likforman-Sulem, Adrian Popescu

Reconnaissance de mots manuscrits hors-vocabulaire en utilisant des ressources web

Cristina Oprean [1], **Chafic Mokbel** [2], **Laurence Likforman-Sulem** [1], **Adrian Popescu** [3]

1. *Institut Mines-Telecom/Telecom ParisTech and CNRS LTCI, Paris (France)*
   likforman@telecom-paristech.fr
2. *University of Balamand, Faculty of Engineering, P.O. Box 100 Tripoli (Liban)*
   chafic.mokbel@balamand.edu.lb
3. *CEA, LIST, LVIC, 91190 Gif-sur-Yvette (France)*
   adrian.popescu@cea.fr

RÉSUMÉ. *Les systèmes de reconnaissance de l'écriture manuscrite s'appuient sur des dictionnaires prédéfinis obtenus à partir de corpus d'entraînement. La taille de ces dictionnaires résulte d'un compromis entre le taux de reconnaissance des mots du vocabulaire (DV) et leur couverture. Si la taille est petite, beaucoup de mots hors vocabulaire (HV) seront non reconnus. Pour améliorer la reconnaissance des mots HV, sans augmenter le dictionnaire statique, nous introduisons une étape supplémentaire qui exploite des ressources web. Après une classification des mots en DV-HV, Wikipédia est utilisée pour créer des dictionnaires dynamiques pour chaque mot HV. Un décodage final est effectué sur le dictionnaire dynamique afin de déterminer le mot le plus probable pour la séquence HV. Nous validons notre approche par des expériences menées avec un système de reconnaissance BLSTM sur la base RIMES. Les résultats montrent que des améliorations sont obtenues par rapport à la reconnaissance avec dictionnaire statique.*

ABSTRACT. *Handwriting recognition systems rely on predefined dictionaries. Small and static dictionaries are often exploited to obtain high in-vocabulary (IV) accuracy at the expense of coverage. Thus the recognition of out-of-vocabulary (OOV) words is not handled efficiently. To improve OOV recognition while keeping IV dictionaries small, we introduce a multi-step approach that exploits web resources. After an IV-OOV classification, Wikipedia is used to create OOV sequence-adapted dynamic dictionaries. A second decoding is done the dynamic dictionary to determine the most probable word for the OOV sequence. We validate our approach with experiments conducted on the RIMES dataset using a BLSTM recognizer. Results show that improvements are obtained compared to handwriting recognition with static dictionary.*

MOTS-CLÉS : *reconnaissance du texte manuscrit, dictionnaires dynamiques, Wikipédia, BLSTM.*

KEYWORDS: *handwriting recognition, adapted dynamic dictionaries, Wikipedia, BLSTM.*

DOI:10.3166/DN.17.3.77-96

1. Introduction

Les systèmes de reconnaissance vocale et d'écriture manuscrite dépendent fortement des ressources linguistiques disponibles, telles que des dictionnaires statiques ou des modèles de langage, comme les n-grammes. La performance de ces systèmes est tributaire d'un choix approprié de la taille du dictionnaire. Les dictionnaires de petite taille génèrent des résultats médiocres à cause de la couverture d'un faible nombre de mots. De même, pour les dictionnaires très riches, un grand nombre de confusions conduit également à une baisse du taux de reconnaissance. De plus, la complexité calculatoire de la reconnaissance augmente avec la taille du dictionnaire et rend difficile l'utilisation de grands dictionnaires dans les applications du monde réel.

Les systèmes de reconnaissance sont classiquement implémentés à l'aide de modèles de Markov cachés (HMM) qui analysent des séquences d'observations. Ces dernières années, des réseaux de neurones récurrents (RNN) ont été utilisés avec un grand succès pour la reconnaissance de la parole et aussi pour la reconnaissance de l'écriture manuscrite (Liwicki *et al.*, 2007). Contrairement aux HMMs, qui sont des modèles génératifs, les réseaux de neurones sont discriminatifs. Au vu de leurs performances supérieures (Graves, 2012) nous choisissons de travailler avec des RNN et, plus précisément, des réseaux de neurones bidirectionnels avec des cellules à mémoire à court terme persistantes (*BLSTM - Bidirectional Long Short Term Memory*). Ceux-ci prennent en compte l'information passée et future en parcourant une image de gauche à droite et de droite à gauche.

Une hypothèse de monde clos consiste à considérer que tous les mots à reconnaître sont dans le dictionnaire. Les mots sont appelés DV (mots dans le vocabulaire). Le système de reconnaissance associe chaque séquence d'observations avec des modèles de mots. La limite la plus importante de ce type d'approche est que seulement les mots du dictionnaire pourront être reconnus correctement. Afin de passer à une hypothèse de monde ouvert, une partie des séquences peut être classée comme des mots hors vocabulaire (HV). Dans ces cas, les modèles de mots sont remplacés par des modèles à base des caractères en boucle, aussi appelés modèles de remplissage. Les performances de ces derniers sont généralement réduites et des méthodes complémentaires sont nécessaires afin d'améliorer la reconnaissance des mots HV.

Les types principaux de mots HV dans un dictionnaire sont : des entités nommées (des prénoms, des noms, des noms de lieux géographiques, les numéros de téléphone, des dates, des noms d'entreprises, l'âge des personnes, des numéros de compte bancaire) qui n'ont pas été rencontrées dans les ressources d'entraînement, des mots associés à des nouvelles thématiques qui apparaissent au fil du temps, des mots d'autres langues incorporés dans des textes, des formes grammaticales des verbes ou des noms qui n'étaient pas présents dans le corpus d'apprentissage (par exemple « signaux », mais pas « signalais »), des codes.

Nous présentons une approche de reconnaissance de l'écriture manuscrite qui utilise des corpus ouverts du web pour améliorer la reconnaissance des mots HV, en adaptant le dictionnaire, qualifié de dynamique, pour une meilleure couverture lexi-

cale du texte à reconnaître. La figure 1 donne un aperçu de l'approche. Chaque image de mot est reconnue avec les modèles à base de mots et à base de caractères et est classée en DV (dans le vocabulaire) ou HV (hors vocabulaire) en comparant leurs log-probabilités à un seuil de confiance. À ce stade, les mots classés comme DV sont affectés nécessairement à l'un des mots du dictionnaire statique. Pour les mots classés comme HV les ressources web (Wikipédia) sont utilisés afin de leur associer des dictionnaires dynamiques, en utilisant les sorties du BLSTM. Enfin, un second décodage est exécuté pour récupérer l'élément du dictionnaire dynamique qui est le plus similaire au mot à reconnaître. Les deux types de décodage s'appuient sur des systèmes de reconnaissance BLSTM, avec et sans dictionnaire, respectivement.

Il y a plusieurs façons de traiter les mots HV dans la littérature. Certains auteurs préfèrent augmenter le vocabulaire utilisé pour tenter une couverture exhaustive du domaine. La méthode souffre d'une complexité de calcul élevée et plusieurs mots similaires peuvent être introduits (Koerich *et al*., 2003), créant ainsi des confusions lors de la reconnaissance. Une autre façon de traiter les HV est d'introduire des systèmes avec un vocabulaire ouvert qui tendent à être plus rapides et plus souples. Ces systèmes sont construits en utilisant des modèles de remplissage ou de n-grammes de caractères (Brakensiek *et al*., 2000 ; Bazzi *et al*., 1999). Ils facilitent la reconnaissance de tous les types de mots et semblent des bons candidats pour résoudre le problème des mots HV. Cependant, lorsqu'aucun dictionnaire n'est utilisé, les performances de reconnaissance diminuent de façon drastique, en raison de la confusion augmentée entre les caractères. Dans (Brakensiek *et al*., 2000) une diminution d'environ 26 % du taux de reconnaissance des mots est signalée sans utilisation de dictionnaire. Plus récemment, l'élargissement du vocabulaire a été réalisé en décomposant le lexique par analyse morphologique (Hamdani *et al*., 2013). Le nouveau vocabulaire est une combinaison des mots et des sous-mots obtenus après le processus de décomposition. Bien que théoriquement intéressante, cette méthode complexe ne produit qu'une légère amélioration du taux de reconnaissance.

La question des mots HV a été plus intensivement étudiée dans la reconnaissance de la parole, où les systèmes doivent gérer ces mots à la volée. Des travaux récents exploitent les ressources externes, pour récupérer les mots HV (Parada *et al*., 2010 ; Oger *et al*., 2009). Le contexte local des HV est utilisé pour récupérer des documents à partir du web qui sont ensuite utilisés pour augmenter le lexique. Notre système est orienté mot et nous n'exploitons pas le contexte local. En outre, nous proposons une nouvelle méthode d'adaptation des dictionnaires pour ne conserver que la partie d'un corpus externe qui est la plus similaire avec les données d'apprentissage et diminuer ainsi la complexité calculatoire.

Le web comprend une grande richesse de sources ouvertes qu'on peut utiliser pour produire des ressources linguistiques. Ce type de ressource est utilisé, par exemple, pour la correction orthographique (Whitelaw *et al*., 2009). Les principales différences entre notre travail et (Whitelaw *et al*., 2009) viennent de la difficulté plus élevée de la reconnaissance des mots HV en écriture manuscrite par rapport à la correction des

fautes d'orthographe et de la façon innovante de créer des dictionnaires dynamiques avec Wikipédia.

Comparé à d'autres corpus web, le choix de Wikipédia comporte deux avantages importants. D'abord, l'encyclopédie couvre un grand nombre de domaines et, en conséquence, peut être utilisée efficacement afin de traiter des corpus d'écriture manuscrite relevant de thématiques diverses. Ensuite, la ressource est librement disponible et mise à jour constamment.

Nous présentons notre approche dans la suite de cet article. La section 2 présente le système de reconnaissance à base de réseaux de neurones récurrents. La section 3 introduit la classification des mots en DV et HV. La création de dictionnaires dynamiques en utilisant Wikipédia est décrite dans la section 4. Finalement, la section 5 est consacrée aux expériences pour la reconnaissance des mots sur la base de données Rimes, base accessible au public.

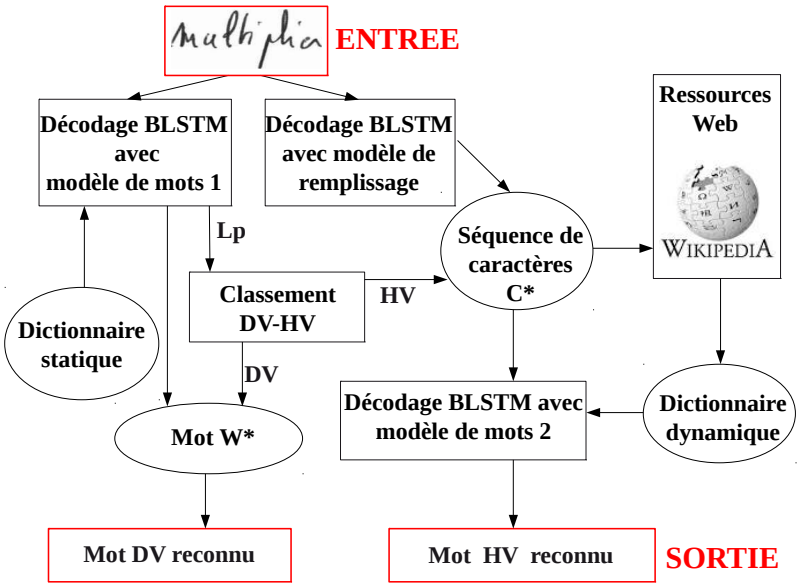

Figure 1. Aperçu de l'approche proposée

2. Description du système de reconnaissance

Dans cette section, nous présentons les principales composantes de notre système de référence. Nous décrivons d'abord le prétraitement, ensuite le module d'extraction de caractéristiques et finalement le système de reconnaissance à base de réseaux de neurones récurrents.

2.1. Prétraitement des images et extraction des caractéristiques

La première étape d'un système de reconnaissance est le prétraitement de l'image d'entrée. Notre module de prétraitement est composé d'algorithmes dont l'objectif principal est de réduire la variabilité due au style d'écriture. Il se décompose en deux étapes principales :

1. Correction de la pente d'écriture qui perturbe l'extraction de la ligne de base dans l'image du mot. Comme certaines des caractéristiques sont basées sur les densités de pixels dans les zones délimitées par les lignes de base inférieures et supérieures, une correction de l'angle d'inclinaison est appliquée. L'algorithme est basé sur l'entropie de la projection du profil (Vinciarelli, Luettin, 2001).

2. Comme le système utilise une approche par fenêtres glissantes, l'écriture oblique peut provoquer un chevauchement des caractéristiques appartenant à des caractères différents dans la même fenêtre verticale. Par conséquent, un angle d'inclinaison est déterminé globalement de l'image, en maximisant une mesure liée à la densité de pixels dans toutes les colonnes de l'image proposée dans (Vinciarelli, Luettin, 2001). Cette correction est ensuite effectuée par une transformation cisaillée.

Lors de la deuxième étape, les caractéristiques extraites sont basées sur les travaux de (Bianne-Bernard *et al.*, 2011) et (Al-Hajj-Mohamad *et al.*, 2005), qui ont prouvé leur efficacité en reconnaissance de l'écriture latine et arabe. Les lignes de base supérieure et inférieure de l'écriture sont automatiquement déterminées pour les images de mots prétraités. Ensuite, une fenêtre glissante est superposée sur l'image pour extraire une séquence de vecteurs de caractéristiques. Une fenêtre glissante de hauteur égale à celle de l'image d'entrée et de largeur $w = 9$ pixels est utilisée dans ce travail. Dans chaque fenêtre glissante, divisée en 20 cellules, 37 caractéristiques sont extraites. Deux fenêtres glissantes consécutives ont un décalage de $\delta = 3$ pixels (figure 2). Les caractéristiques extraites comprennent :

– 2 caractéristiques représentant les transitions de premier-plan/arrière-plan ;

– 12 caractéristiques pour la configuration des concavités ;

– 3 caractéristiques pour la position du centre de gravité - la première caractéristique donne la position par rapport à des lignes de base, la seconde est la distance en nombre de pixels par rapport à la ligne de base inférieure, et la dernière représente la différence entre les centres de gravité des deux fenêtres voisines ;

– 9 caractéristiques correspondant à la densité de pixels dans chaque colonne ;

– 3 caractéristiques correspondant à la densité de pixels dans la fenêtre glissante, au-dessus et en dessous des lignes de base ;

– 8 caractéristiques directionnelles correspondant à l'histogramme des gradients pour les 8 orientations de 0 à $7 * \pi/4$, avec un pas de $\pi/4$.

Les séquences de vecteurs de caractéristiques extraites sont ensuite envoyées au système de reconnaissance basé sur les réseaux des neurones récurrents.

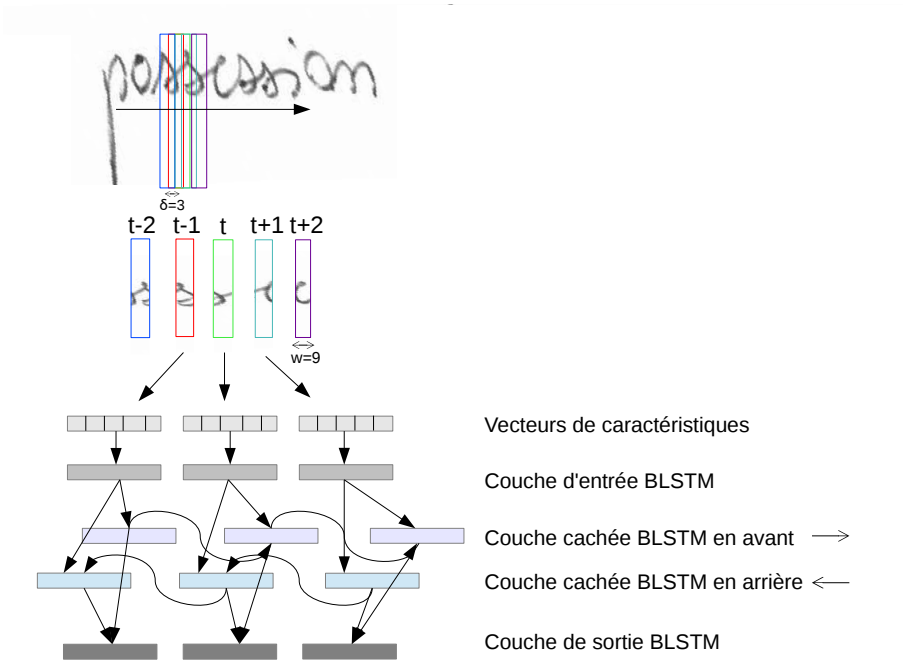

Figure 2. Fenêtres glissantes pour le mot « possession » de la position horizontale t-2 à t+2. Le BLSTM est déroulé pendant trois pas de position horizontale

2.2. Le système de reconnaissance BLSTM

Les réseaux de neurones récurrents sont une classe de réseaux de neurones artificiels dans laquelle les connexions entre les unités cachées assurent un comportement temporel dynamique et le stockage des informations. Les RNN bidirectionnels (Schuster *et al.*, 1997) traitent les données en avant et en arrière en utilisant deux couches récurrentes séparées. Ainsi, les RNN bidirectionnels profitent du contexte passé et futur de la séquence donnée en entrée. La passe avant traite la séquence de gauche à droite, tandis que la passe arrière traite la séquence d'entrée dans la direction opposée. Les deux couches cachées sont reliées aux mêmes couches d'entrée et de sortie.

Le BLSTM est un type d'architecture RNN dans lequel les unités de sommation dans la couche cachée sont remplacées par des blocs de mémoire, introduits afin de résoudre le problème de fuite du gradient « vanishing gradient problem » (Hochreiter *et al.*, 2001) . Chaque LSTM est composé d'une cellule mémoire de type carrousel et de trois portes (entrée, sortie, oubli) qui permettent de laisser passer ou de bloquer l'information en entrée du réseau ou celle provenant des sorties des blocs de la couche cachée. La couche d'entrée est constituée par les caractéristiques extraites dans la fenêtre glissante. La couche de sortie comprend autant des cellules que le nombre des symboles et des lettres utilisés dans le lexique : 79 cellules correspondant à tous les 79

caractères (a–z, A–Z, 0–9, « / », « ’ », « », « – ») nécessaires pour la modélisation de la base de données RIMES.

En nous inspirant des travaux développés dans (Graves *et al*., 2009), nous considérons une architecture BLSTM dont chaque réseau (avant ou arrière) comprend une couche cachée de 100 blocs mémoire. Pour l'entraînement du réseau, la méthode de « Back-Propagation Through Time » (Werbos, 1988 ; Williams, Zipser, 1995) est utilisée. Les poids du réseau sont mis à jour en utilisant la méthode de descente du gradient. Après chaque cycle d'entraînement, le taux d'erreur de reconnaissance est évalué sur un ensemble de validation. Afin d'éviter le surapprentissage, l'apprentissage du réseau est arrêté si le taux d'erreur ne diminue pas pendant 20 cycles.

Le BLSTM calcule les sorties de réseau correspondant aux classes de caractères. Ces sorties étant normalisées, on obtient donc pour chaque trame sa probabilité a posteriori. Ensuite, un algorithme de type « Forward-Backward », dénommé CTC (Connectionist Temporal Classification) (Graves *et al*., 2006) prend ces probabilités a posteriori en entrée et fournit un mot du dictionnaire ou, dans le cas des modèles de remplissage, une chaîne de caractères.

3. Le traitement des mots HV and DV

L'utilisation de modèles de mots pour la reconnaissance des séquences de caractères manuscrits donne de bons résultats pour les mots DV si un bon compromis entre la couverture et le pouvoir discriminant est trouvé. Pour les mots de HV restants (par exemple, mots peu fréquents, entités nommées, codes, etc.), des méthodes alternatives sont nécessaires. Une classification efficace des mots DV et HV conditionne l'obtention d'une performance élevée du système global. Dans cette section, nous présentons brièvement notre approche basée sur la log-probabilité pour la classification DV-HV des mots.

3.1. Modélisation des mots DV-HV

Il existe de nombreuses approches dans la littérature pour la détection des mots HV pour la reconnaissance de la parole et de l'écriture manuscrite. Elles sont basées sur des modèles de remplissage, des modèles hybrides de sous-mots (Bazzi, Glass, 2000), (Bisani, Ney, 2005) ou des scores de confiance (Wessel *et al*., 2001), (Sun *et al*., 2003). Des systèmes de reconnaissance de l'écriture manuscrite combinant des modèles de remplissage et des scores de confiance ont été développés pour détecter les mots HV dans des phrases ou des textes (Quiniou, Anquetil, 2007), (Fischer *et al*., 2010), (Cuayáhuitl, Serridge, 2002).

Dans un modèle de remplissage, les mots HV sont représentés par des réseaux formés de caractères indépendants, délimités par des modèles d'espace (figure 3 b). Les mots DV sont représentés par des réseaux de mots plus sophistiqués construits sur un dictionnaire statique prédéfini (figure 3 a). Pour une séquence donnée, la probabilité du vecteur des caractéristiques observées est calculée en utilisant le BLSTM.

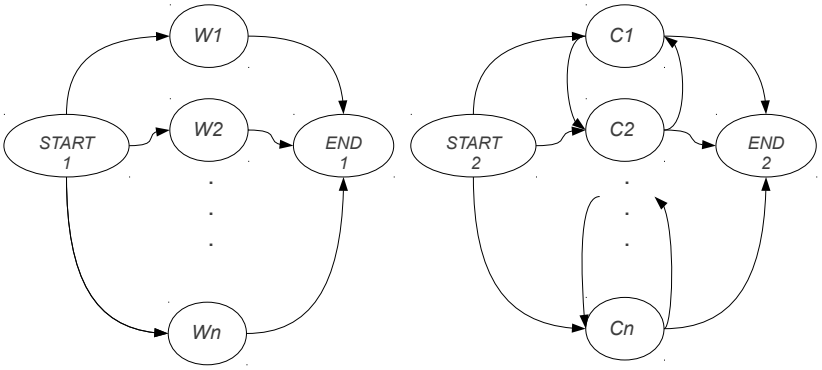

Figure 3. Réseau des mots (a) réseau des caractères (b)

3.2. Classification DV-HV des images de mots

Les séquences sont classées comme DV ou HV en utilisant les log-probabilités calculées à l'aide du réseau des mots. Soit $X = x_1, x_2, ...x_N$ la séquence de trames observée, où N est le nombre de trames. Sa valeur de log-probabilité Lp est obtenue en utilisant l'algorithme CTC sur les sorties du réseau BLSTM.

$$Lp = log\ P(W^*|X) = \max_W \log\ P(W|X) \tag{1}$$

où W^* est le meilleur mot du dictionnaire pour la séquence X. Pour une séquence donnée, le score de reconnaissance pour un mot W du dictionnaire est calculé en utilisant le système de reconnaissance de BLSTM et la couche CTC (section 2.2). Ce score est une log-probabilité Lp qui correspond à une somme de chemins associés au mot W. Le CTC construit une matrice dont les lignes correspondent aux caractères de W entourés de blancs (ou espaces). Le score Lp correspond au score le plus élevé parmi les mots du dictionnaire.

La valeur Lp est utilisée pour classer la séquence décodée comme HV ou DV. Si $\frac{Lp}{N} \geq thre$ la séquence candidate est affectée à la classe DV, en supposant qu'elle appartient au dictionnaire statique. Sinon, elle est considérée comme un mot HV. La valeur de $thre$ est optimisée (voir section 5) en utilisant une base de validation.

4. Création des dictionnaires dynamiques pour la reconnaissance des mots HV

Nous décrivons maintenant la procédure de construction des dictionnaires dynamiques pour les mots HV à partir des séquences de caractères reconnues par un modèle de remplissage. Lorsqu'une séquence est classée comme HV, la sortie obtenue avec le modèle de remplissage est comparée au dictionnaire dynamique afin de trouver le

mot le plus probable. La création des dictionnaires dynamiques repose sur Wikipédia. Les principaux avantages de Wikipédia sont sa disponibilité (en téléchargement libre), son caractère structuré et sa couverture importante (plus d'un million d'articles en français).

Wikipédia est une corpus vaste, qui décrit un grand nombre de concepts et qui est donc approprié pour la création de dictionnaires offrant une bonne couverture de la langue. La version de Wikipédia en français utilisée ici est celle de Septembre 2012. Les 410 482 articles contenant au moins 100 mots distincts ont été retenus. Un de nos objectifs est de tester la création d'un dictionnaire adapté au domaine défini par les documents d'apprentissage inclus dans le corpus RIMES. A cette fin, nous avons utilisé la similarité cosinus (Singhal, 2001) entre la représentation TF–IDF (Term Frequency – Inverse Document Frequency) de la partie entraînement de RIMES, considérée comme un document unique (voir la section 5) et la représentation TF–IDF de chaque article Wikipédia (Salton *et al.*, 1975). Par conséquent, nous nous intéressons à une première sélection d'un sous-ensemble Wikipédia qui est le plus pertinent pour le domaine cible. TF–IDF mesure l'importance d'un mot dans un document par le TF, mais il tient compte aussi de sa distribution au sein de la collection Wikipedia à travers l'IDF. Autrement dit, l'importance d'un terme dans un document est directement proportionnelle à son nombre d'occurrences dans le document et en relation inverse avec le nombre de différents documents de la collection dans lequel il apparaît.

Ainsi nous pouvons ordonner les articles en fonction de leur similarité avec le domaine délimité par RIMES. Nous construisons deux variantes de dictionnaires dynamiques. L'un est dit dictionnaire adapté et est basé sur les premiers 20 000 articles les plus similaires à la représentation du corpus RIMES. L'autre est dit générique et il est construit de la même façon mais en prenant les premiers 200 000 articles. Dans chaque dictionnaire, les mots sont ordonnés en utilisant leur fréquence d'apparition dans l'ensemble des articles exploités. La création des dictionnaires adaptés au domaine ou génériques est illustrée dans la figure 4.

Le dictionnaire adapté favorise des termes relatifs au domaine. Le dictionnaire générique capte des propriétés plus génériques des mots et est plus compréhensif. Nous illustrons la sélection des termes avec le mot *facture*, un mot très caractéristique pour la base de données RIMES. Il apparaît 1 440 fois dans l'ensemble des documents adaptés au domaine et de 1 459 fois dans le dictionnaires générique.

Pour construire le dictionnaire dynamique on détermine les mots du dictionnaire Wikipedia les plus proches de la séquence de lettres décodée par le modèle de remplissage. Dans ce but, on utilise la distance Levenshtein (Levenshtein, 1966), une mesure classique de la différence entre deux séquences de caractères. Elle calcule le nombre de modifications nécessaires pour passer d'une chaîne à l'autre. Pour une séquence candidate donnée, les mots du dictionnaire sont ordonnés en fonction de leur distance Levenshtein. À distance égale, les mots sont triés en fonction de leur fréquence dans les documents. Seulement les k mots les plus fréquents pour lesquels la différence entre la longueur des mots du dictionnaire et celle de la séquence décodée est au plus l sont retenus dans le dictionnaire dynamique. Les valeurs des paramètres k et l sont

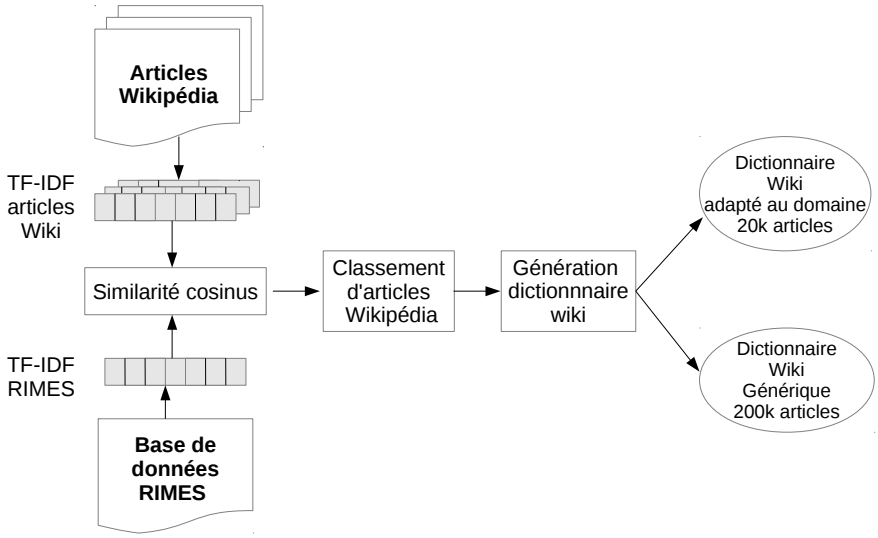

Figure 4. Création des dictionnaires adaptés au domaine et génériques en utilisant Wikipédia

déterminées empiriquement sur une base de validation, en examinant la différence entre les longueurs de mots du dictionnaire et les séquences déterminées avec le modèle de remplissage. L'augmentation de la valeur l permettrait de retenir plus de mots du dictionnaire Wikipédia. Toutefois, elle n'aurait qu'une influence marginale sur les résultats car la distance entre le décodage correct et celle de la séquence prédite automatiquement est usuellement parmi les plus réduites obtenues pour cette paire.

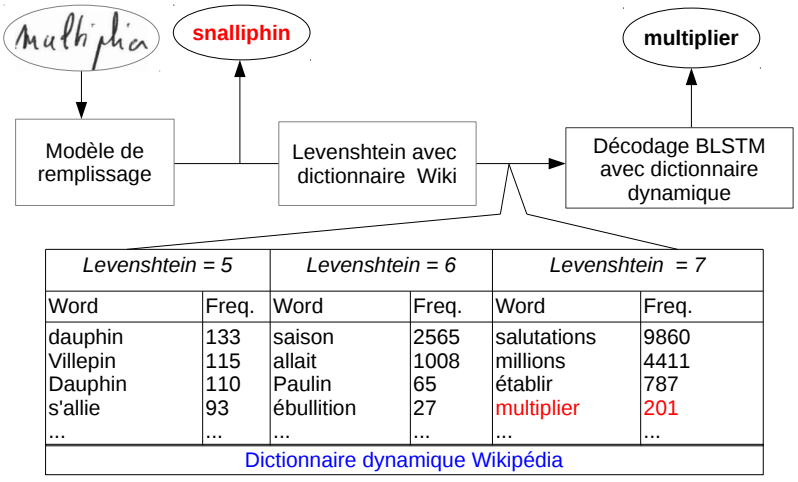

Figure 5. Reconnaissance de mots hors vocabulaire basée sur Wikipédia

Les auteurs de (Damerau, 1964) montrent que 80 % des fautes d'orthographe ont une distance d'édition 1. En calculant sur la base de validation la moyenne des distances Levenshtein entre la séquence de caractères obtenue après le premier décodage et le vrai mot nous obtenons une valeur de 2,80. En outre, le pourcentage global de reconnaissances des caractères erronés est de 35 %. Ces résultats montrent que le problème abordé ici est plus difficile que la correction des fautes d'orthographe et il est nécessaire de conserver un nombre relativement élevé de mots similaires dans le dictionnaire dynamique. Alors, pour les expériences nous avons retenu des tailles du dictionnaire de $k = 100, 200, 500, 1000$ et $l = 5$. La valeur de l, qui donne la différence entre les longueurs de la séquence et du mot du dictionnaire dynamique est choisie afin de prendre en compte les confusions qui arrivent souvent par deux, par exemple : « m » est reconnu comme le groupe « rn » ou « nm », etc.

Les listes créées avec Levenshtein sont utilisées pour mettre en place des dictionnaires adaptés. Par exemple, le mot *multiplier* (figure 5) a d'abord été décodé comme *snalliphin*. En calculant la distance de Levenshtein avec l'ensemble des mots contenus dans les dictionnaires Wikipédia générique ou adapté, on obtient les groupes représentés dans la figure 5. Notez que le vrai mot *multiplier* se trouve à une distance de Levenshtein 7. Même si le mot n'a pas une fréquence élevée dans les documents de Wikipédia et la distance Levenshtein n'est pas minimale, le mot peut encore être récupéré grâce à l'utilisation du second décodage.

5. Expériences

Nous menons les expériences sur la base de données RIMES (Grosicki, El-Abed, 2011a) afin d'évaluer l'efficacité de la méthode de création des dictionnaires dynamiques proposée ici. La métrique utilisée dans toutes les expériences est le taux de reconnaissance, calculé comme le rapport entre le nombre de mots correctement reconnus et la taille de l'ensemble de test.

La base de données RIMES a été créée en demandant à des volontaires d'écrire des lettres relatives à des scénarios tels que la modification de comptes bancaires, des déclarations de dommages ou de paiement etc. Un exemple des documents RIMES peut être visualisé dans la figure 6.

La consigne a été donnée aux volontaires d'écrire librement en utilisant de l'encre noire sur du papier blanc. Les documents obtenus ont été numérisés en niveaux de gris. À partir de ces documents sont extraits et étiquetés des blocs de texte, des lignes de texte et des mots. Des compétitions de reconnaissance de mots et de lignes de texte ont été organisées en utilisant cette base de données en 2009 (Grosicki, Abed, 2009) et 2011 (Grosicki, El-Abed, 2011a). RIMES 2011 est divisée en trois ensembles : apprentissage, validation et test, composés de 51 739, 7 464 et 7 776 images de mots respectivement. Les dictionnaires statiques correspondants contiennent 4 972, 1 588 et 1 612 mots uniques.

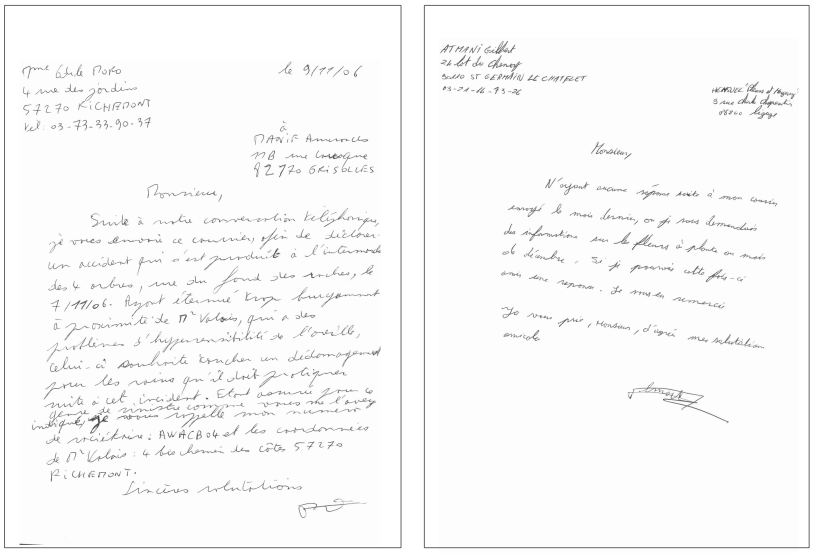

Figure 6. Exemples des documents RIMES

5.1. Détection de mots HV

L'ensemble de test contient ainsi 460 mots HV réels, dont certains sont des séquences spéciales (codes postaux, dates, numéros de téléphone, etc.). Celles-ci ne peuvent pas être reconnues avec les techniques présentées ici et nous nous focalisons sur des mots du vocabulaire obtenu avec Wikipédia. L'utilisation de BLSTM pour reconnaître de nouveaux mots passe par le calcul d'un seuil $thre$ (section 3.2) qui classe les mots de la base de test en HV ou DV. Pour régler les paramètres pour la classification des mots DV-HV, nous utilisons la base de validation qui contient 5,5 % de mots HV, un ratio de mots HV proche de celui de la base de test. Les listes de mots HV et DV de la base de validation sont décodées avec le dictionnaire statique d'apprentissage et les deux distributions de log-probabilité obtenues sont utilisées pour trouver un seuil qui partage les deux classes, tout en maximisant le taux de reconnaissance DV. Les deux distributions sont centrées en deux moyennes distinctes $mean_{HV}$ et $mean_{DV}$, mais il y a un chevauchement important qui ne permet pas une séparation parfaite.

Un seuil $thre$ proche de la moyenne $mean_{DV}$ minimise le nombre de mots HV réels affectés à la classe DV (faux négatifs), mais il classifie incorrectement un grand nombre de DV comme étant des HV. Inversement, pour une valeur de seuil $thre$ proche de $mean_{HV}$, le nombre de confusions HV-DV est réduit tandis qu'une proportion plus importante des HV est classée comme des DV.

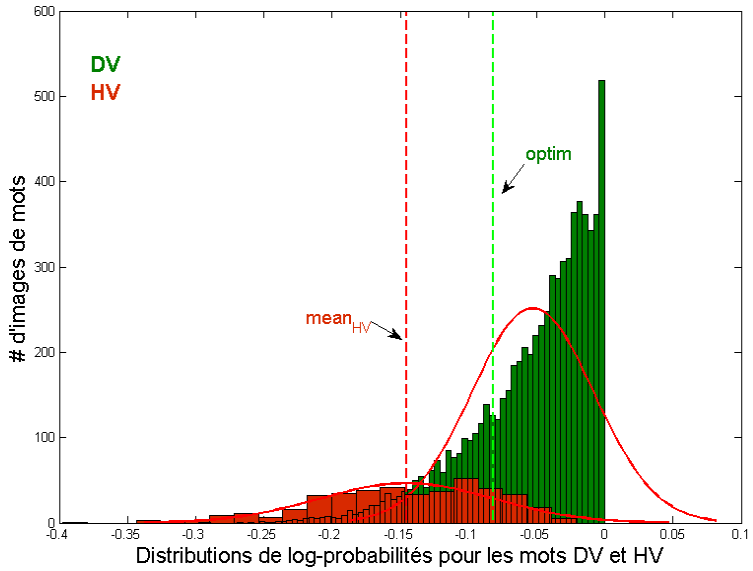

Figure 7. Distributions des log-vraisemblances pour les mots HV et DV sur la base de validation RIMES 2011

Dans une première expérience réalisée sur la base de validation (7 049 DV et 415 HV), pour $thre = mean_{HV}$, 47 % de vrais mots HV sont détectés, tandis que si le seuil est $thre = mean_{DV}$, 96 % des mots HV sont récupérés. Dans le dernier cas, la proportion des mots DV reconnus comme HV augmente de 0,05 % à 42 %, ce qui a un effet négatif sur le taux de reconnaissance global.

5.2. Reconnaissance de mots HV

Une première expérience a été menée sur les 7 776 mots de la base de test RIMES 2011. Notez que les mots détectés comme DV sont directement décodés avec le vocabulaire d'apprentissage. Les mots classés comme HV (leur score de log-probabilité normalisé $\frac{Lp}{N} < thre$) sont traités avec la méthode décrite dans la section 4. Dans (Oprean *et al.*, 2013), nous avions montré que $thre$=$mean_{HV}$ maximise le taux de reconnaissance total. Lorsque $thre$=$mean_{DV}$, 3 612 images de mots ont été classées comme HV, comparativement à 568 lorsque $thre$=$mean_{HV}$. Pour les expériences de cet article nous comparons le taux de reconnaissance d'une *séparation parfaite de mots HV-DV* utilisant le taux de reconnaissance avec la méthode de séparation DV-HV proposée dans la section 3.2.

Les tailles des vocabulaires Wikipédia sont de 76 566 et 137 200, en ne considérant que les 20 000 articles les plus adaptés au domaine (cas « adapté ») et 200 000 articles (cas « générique »), respectivement. Ces valeurs sont obtenues en ne retenant que les

mots ayant au minimum 12 occurrences pour le dictionnaire adapté et au minimum 40 pour le dictionnaire générique.

Pour chaque mot HV, un dictionnaire dynamique est obtenu en comparant la sortie du modèle de remplissage avec le dictionnaire Wikipédia adapté ou générique. Les mots obtenus sont triés en fonction de leur similarité avec la séquence de caractères associée au mot HV à reconnaître. Un deuxième filtre ne retient que les mots dont la différence de longueur entre la séquence de caractères issue du premier décodage et celle du mot du dictionnaire dynamique n'excède pas 5. D'autres valeurs pour le nombre d'articles retenus et pour le nombre d'occurrences de mots dans les documents ont été testées. Les résultats préliminaires obtenus ont montré que les valeurs citées ci-dessus assurent un bon compromis entre le taux de reconnaissance et la complexité calculatoire.

Nous considérons le cas d'un seuil $thre$=$mean_{HV}$ et le cas d'une séparation parfaite de mots HV-DV pour la classification et les dictionnaires Wikipédia adapté et générique pour construire les dictionnaires dynamiques avec des tailles variables : k=$100, 200, 500, 1000$. Les résultats dans le tableau 1 sont obtenus sur un ensemble de test de taille 7 776, distinct de la base de validation.

Tableau 1. Le taux de reconnaissance des mots HV sur l'ensemble de test de la base RIMES 2011

Type du dictionnaire	**Adapté**				**Générique**			
Taille dict. dynamique	100	200	500	1000	100	200	500	1000
séparation DV/HV parfaite	41,95 %	42,39 %	42,60 %	**42,82%**	42,17 %	**42,82 %**	42,39 %	42,60 %
$thre = mean_{HV}$	**33,97 %**	33,62 %	33,45 %	33,45 %	**31,16** %	30,63 %	30,10 %	30,10 %

Les résultats montrent que la qualité des résultats est équivalente pour les dictionnaires Wikipédia adapté et générique. Toutefois, au vu de sa taille significativement plus réduite, l'usage du dictionnaire adapté est préférable pour réduire la complexité des calculs.

Pour une taille du dictionnaire dynamique de 1 000 mots et une détection DV-HV parfaite, le taux de reconnaissance avec un dictionnaire adapté au domaine serait de 42,82 %, contre 42,60 % pour le générique. Lorsque le seuil de séparation est $thre = mean_{HV}$ le meilleur taux de reconnaissance (33,97 %) est obtenu pour un dictionnaire dynamique de taille 100. La distance Levenshtein moyenne (sur la base de validation) entre la séquence de caractères obtenue après le décodage BLSTM avec un modèle de remplissage et le vrai mot est de 2,8. Cette valeur explique, au moins en partie, le fait que la taille du dictionnaire dynamique a une faible influence sur les scores de classification. Le résultat obtenu montre aussi qu'il est possible d'obtenir des bons résultats tout en réduisant la complexité de la deuxième étape de classification avec un dictionnaire adapté.

Dans le tableau 2, nous présentons une comparaison des résultats obtenus avec un système sans traitement des mots HV (*Dict. statique*), avec notre approche (*Dict. dynamique*), ainsi qu'un résultat théorique (*Dict. dynamique sép. parfaite*) dont le rôle est de montrer l'amélioration qui pourrait être atteinte dans le cas d'une séparation

Tableau 2. Taux de reconnaissance mots sur la base de données de test Rimes 2011

Détection DV-HV	**Taux de reconnaissance** [%]
Dict. statique	75,45 $[73, 78, 77, 11]$
Dict. dynamique	75,87 $[74, 19, 77, 54]$
Dict. dynamique sép. parfaite	**77,98** $[76, 25, 79, 7]$

parfaite DV-HV. Dans la suite, nous fournissons les intervalles de confiance de Wald calculés comme

$$\hat{p} \pm k * N^{-\frac{1}{2}}(\hat{p}(1-\hat{p}))^{\frac{1}{2}} \quad (2)$$

où $\hat{p}$ est la proportion de mots bien reconnus, N le nombre de données de test et k est la $100(1-\frac{\alpha}{2})$ ème centile de la distribution normale, avec un risque $\alpha = 5$ %.

Étant donné le chevauchement entre les distributions des mots DV-HV, la séparation obtenue est imparfaite. Toutefois, une légère amélioration est obtenue par rapport à une approche sans traitement des mots HV (75,87 % vs. 75,45 %). La différence entre *Dict. dynamique* et *Dict. dynamique sép. parfaite* montre que la marge de progression possible à obtenir avec ce type de classification est encore importante et nos travaux futurs vont se focaliser sur ce point. L'intervalle de confiance pour le système *Dict. statique* est égal à $[73, 78\ \%, 77, 11\ \%]$. Pour le système *Dict. dynamique*, il est égal à $[74, 19\ \%, 77, 54\ \%]$ et pour le système *Dict. dynamique sép. parfaite*, à $[76, 25\ \%, 79, 7\ \%]$ avec un niveau de risque $\alpha = 5$ %. Le résultat obtenu pour le système *Dict. dynamique sép. parfaite* serait significatif dans le cas de 6 % des mots HV, car la valeur obtenue (77,98 %) est hors de l'intervalle de confiance obtenu pour le système *Dict. statique*. Dans le cas où seulement 6 % des mots sont HV l'amélioration d'environ 0,5 % entre *Dict. dynamique* et *Dict. statique* n'est pas significative. Toutefois, elle pourrait devenir significative dans le cas où le nombre de mots HV serait plus élevé, ce qui serait probable pour des bases de test réelles. Nous avons effectué des expériences avec un dictionnaire statique réduit à chaque fois de 10 %, jusqu'à obtenir 50 % de mots HV. La différence entre le taux de reconnaissance des systèmes *Dict. statique* et *Dict. dynamique* (figure 8) augmente avec le pourcentage de mots HV. Pour 50 % de mots HV, le taux de reconnaissance passe de 47,08 % (*Dict. statique*) à 55,20 % (*Dict. dynamique*). Ce résultat montre que l'intérêt de notre approche est d'autant plus grand que le nombre de mots hors vocabulaire augmente.

La figure 9 montre la couverture des dictionnaires en fonction de la proportion de mots HV dans l'ensemble de test. Une classification DV-HV entraîne l'utilisation d'un dictionnaire adéquat : statique pour les mots DV et dynamique pour les mots HV. Quand seulement 6 % de mots HV sont présents dans l'ensemble de test, les couvertures des dictionnaires statique et dynamiques sont très similaires en raison de mauvaises classifications DV-HV. Une classification parfaite maximise la potentialité de cette approche. Si le nombre des HV augmente, les deux couvertures des dictionnaires diminuent. Lorsque les dictionnaires dynamiques sont utilisés, la couverture est plus grande que lorsque le dictionnaire statique est utilisé, parce que des mots similaires sont extraits de la ressource externe. La comparaison effectuée avec divers

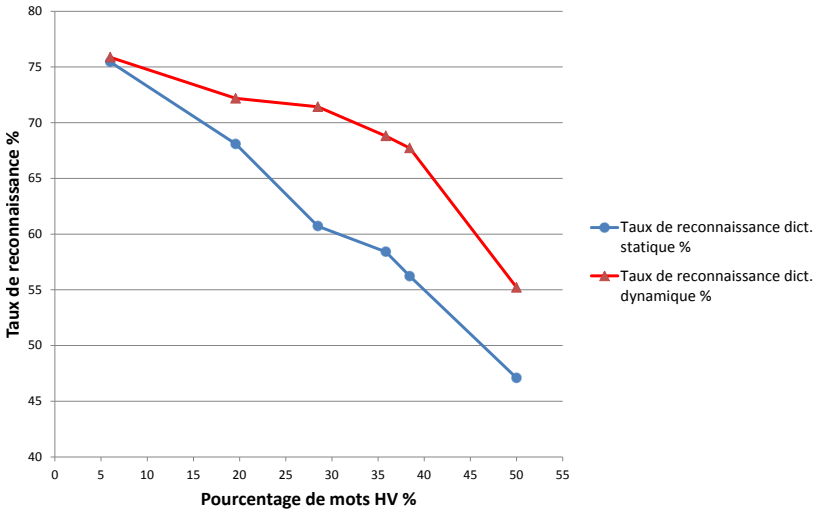

Figure 8. Taux de reconnaissance en fonction du pourcentage de mots HV dans la base de test

pourcentages de HV est utile car elle simule les conditions que l'on rencontre pour les grands ensembles de données avec différentes tailles. Ce résultat indique que les ressources externes pourraient être encore plus utiles dans les grandes bases de test, qui contiennent un plus grand nombre de mots uniques que RIMES. L'amélioration relative apportée par les dictionnaires dynamiques par rapport aux dictionnaires statiques est plus élevée lorsque la proportion des HV augmente, ou de manière équivalente, quand la couverture du dictionnaire statique diminue.

D'autres résultats sont disponibles pour cette base, notamment ceux de la compétition ICDAR 2011 (Grosicki, El-Abed, 2011b) de reconnaissance de mots. Cependant les résultats ne sont pas facilement comparables, car le dictionnaire de la compétition inclut les mots de l'ensemble de test, et donc pas de mots HV. Le dictionnaire de cette compétition contient environ 5 740 mots issus de la base d'apprentissage et celle de test. Comme nous avons présenté dans la section 5, notre dictionnaire contient environ 5 000 mots issus de la base d'apprentissage seule, avec un pourcentage de 6 % de mots HV, ce qui correspond bien à des situations réelles.

6. Conclusion et travaux futurs

Le traitement des mots HV est un défi important en reconnaissance de l'écriture manuscrite. Nous introduisons des méthodes de détection et de correction de ces mots, basées sur des ressources externes à grande échelle.

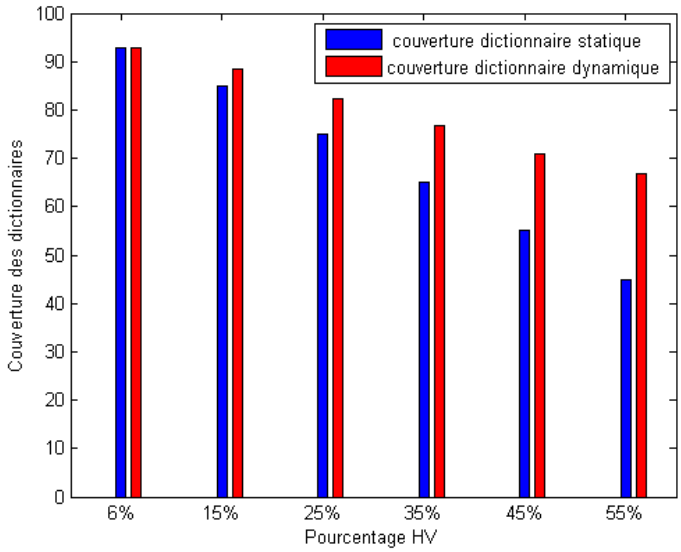

Figure 9. Couvertures des dictionnaires statiques et dynamiques pour différents pourcentages de mots HV

La méthode de détection utilise les valeurs de log-vraisemblance des mots décodés avec un dictionnaire statique et calcule un seuil qui classe de nouveaux mots inconnus dans l'une des catégories HV ou DV.

Une première contribution pour la reconnaissance des mots HV est l'utilisation novatrice de Wikipédia dans une tâche de reconnaissance d'écriture par réseau récurrent bi-directionnel BLSTM. Des dictionnaires dynamiques ont été créés à partir de Wikipédia pour les séquences initialement classées comme des mots HV. Ces dictionnaires sont de deux types : adaptés au domaine délimité par les documents de la base d'apprentissage ou génériques. L'introduction d'une méthode d'adaptation au domaine constitue une seconde innovation de ce travail. Nous n'avons plus besoin d'utiliser Wikipédia en entier, car en gardant seulement les articles les plus similaires à la base d'entraînement, une couverture similaire à celle d'un dictionnaire générique est obtenue. L'avantage du dictionnaire adapté vient de sa taille plus réduite qui permet d'optimiser les calculs sans perte de performance.

Une troisième innovation concerne l'introduction d'une seconde étape de décodage avec BLSTM, afin d'améliorer le processus de reconnaissance par rapport à l'utilisation du dictionnaire statique.

Enfin, la méthode proposée est facilement reproductible puisque nous exploitons des ressources librement disponibles. Elle pourrait être appliquée à d'autres tâches qui impliquent des décodages de séquences non fiables (en reconnaissance de la parole,

OCR, etc.). Tout aussi important, en raison de la dimension multilingue de Wikipédia, la méthode peut être facilement adaptée à un grand nombre de langues.

Les résultats obtenus sont prometteurs et nous poursuivrons les travaux dans plusieurs directions importantes. Premièrement, la reconnaissance des cas particuliers de mots HV (des codes, des dates, des numéros de téléphone, etc.) n'est pas traitée ici car les ressources web ne sont pas adaptées à cette tâche. Des classifieurs dédiés, tels que ceux décrits dans (Shastri, Fontaine, 1995 ; Morita, 2003), seront rajoutés au système dans l'avenir afin d'améliorer les performances.

Deuxièmement, le choix des mots similaires (calcul de la distance Levenshtein, nombre d'occurrences des mots dans les documents, la différence entre la longueur du mot trouvé et celle de la séquence des caractères HV) peut être amélioré en ajoutant d'autres méthodes. Vu que les sorties de BLSTM sans dictionnaire sont relativement propres, une méthode de regroupement pourra être appliquée sur la sélection initiale afin de réduire l'espace de recherche pour le décodage en considérant des mots qui sont plus susceptibles d'être confondus avec la séquence décodée. De plus, d'autres méthodes décrites dans la littérature telles que la longueur et la forme des mots (Kaufmann *et al*., 1997 ; Seni *et al*., 1996) pourraient s'avérer utiles pour un meilleur filtrage des mots et seront testées.

Troisièmement, le nombre de mots HV dans les ensembles de données du monde réel est beaucoup plus élevé que dans RIMES 2011. Dans un cas réel, les performances des modèles à base de mots purs seraient plus réduites et le gain de performance obtenu avec notre méthode est susceptible d'augmenter. Par conséquent, nous allons vérifier l'hypothèse que les améliorations apportées ici sont encore plus élevés pour les bases de données du monde réel.

Finalement, l'amélioration de la méthode de classement des mots HV-DV qui se base sur la simple comparaison de la log-probabilité à un seuil s'avère importante, vu l'augmentation potentielle des performances (*Dict. dynamique sép. parfaite* 77,98 % vs. *Dict. dynamique* 75,87 %). Comme les deux distributions de log-probabilité sont superposées, d'autres critères de séparation doivent êtres considérés.

Bibliographie

Al-Hajj-Mohamad R., Likforman-Sulem L., Mokbel C. (2005). Arabic handwriting recognition using baseline dependent features and hidden markov modeling. In *Proc. of icdar'05*, p. 893-897.

Bazzi I., Glass J. R. (2000). Modeling out-of-vocabulary words for robust speech recognition. In *Interspeech*, p. 401-404.

Bazzi I., Schwartz R. M., Makhoul J. (1999). An omnifont open-vocabulary ocr system for english and arabic. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, n° 6, p. 495-504.

Bianne-Bernard A.-L., Menasri F., El-Hajj R., Mokbel C., Kermorvant C., Likforman-Sulem L. (2011). Dynamic and contextual information in HMM modeling for handwritten word recognition. *IEEE PAMI*, vol. 99, n° 10, p. 2066-2080.

Bisani M., Ney H. (2005). Open vocabulary speech recognition with flat hybrid models. In *Interspeech*, p. 725-728.

Brakensiek A., Willett D., Rigoll G. (2000). Unlimited vocabulary script recognition using character n-grams. In *In proc. 22. dagm-symposium tagungsband.* Springer-Verlag.

Cuayáhuitl H., Serridge B. (2002). Out-of-vocabulary word modeling and rejection for spanish keyword spotting systems. In *Micai'02.*

Damerau F. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, vol. 7, p. 171-176.

Fischer A., Keller A., Frinken V., Bunke H. (2010). HMM-based word spotting in handwritten documents using subword models. In *Icpr*, p. 3416-3419.

Graves A. (2012). *Supervised sequence labelling with recurrent neural networks* (vol. 385). Springer.

Graves A., Fernández S., Gomez F. J., Schmidhuber J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Icml*, p. 369-376.

Graves A., Liwicki M., Fernández S., Bertolami R., Bunke H., Schmidhuber J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, n° 5, p. 855–868.

Grosicki E., Abed H. E. (2009). Icdar 2009 handwriting recognition competition. In *Icdar.*

Grosicki E., El-Abed H. (2011a). ICDAR 2011: French handwriting recognition competition. In *Icdar.*

Grosicki E., El-Abed H. (2011b). Icdar 2011-french handwriting recognition competition. In *Proc. of icdar'11*, p. 1459–1463.

Hamdani M., El-Desoky Mousa A., Ney H. (2013). Open vocabulary arabic handwriting recognition using morphological decomposition. In *International conference on document analysis and recognition*, p. 280–284. Washington DC.

Hochreiter S., Bengio Y., Frasconi P., Schmidhuber J. (2001). *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.*

Kaufmann G., Bunke H., Hadorn M. (1997). Lexicon reduction in an hmm-framework based on quantized feature vectors. In *Proceedings of the 4th international conference on document analysis and recognition*, p. 1097–1101.

Koerich A. L., Sabourin R., Suen C. Y. (2003). Large vocabulary off-line handwriting recognition: A survey. *Pattern Analysis and Applications*, vol. 6, p. 97–121.

Levenshtein V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, vol. 10, p. 707.

Liwicki M., Graves A., Bunke H., Schmidhuber J. (2007). A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *In proceedings of the 9th international conference on document analysis and recognition, icdar 2007.*

Morita M. E. (2003). *Automatic recognition of handwritten dates on brazilian bank cheques*. Thèse de doctorat non publiée.

Oger S., Popescu V., Linarés G. (2009). Using the world wide web for learning new words in continuous speech recognition tasks: Two case studies. In *in specom.*

Oprean C., Likforman-Sulem L., Popescu A., Mokbel C. (2013). Using the web to create dynamic dictionaries in handwritten out-of-vocabulary word recognition. In *Icdar*, p. 989-993.

Parada C., Sethy A., Dredze M., Jelinek F. (2010). A spoken term detection framework for recovering out-of-vocabulary words using the web. In *Interspeech.*

Quiniou S., Anquetil É. (2007). Use of a confusion network to detect and correct errors in an on-line handwritten sentence recognition system. In *Icdar*, p. 382-386.

Salton G., Wong A., Yang C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, vol. 18, n° 11, p. 613-620.

Schuster M., Paliwal K. K., General A. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*.

Seni G., Srihari R. K., Nasrabadi N. (1996). Large vocabulary recognition of on-line handwritten cursive words. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, n° 7, p. 757–762.

Shastri L., Fontaine T. (1995). Recognizing handwritten digit strings using modular spatio-temporal connectionist networks. *Connection Science*, vol. 7, p. 211-246.

Singhal A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, vol. 24, n° 4, p. 35-43.

Sun H., Zhang G., Zheng F., Xu M. (2003). Using word confidence measure for oov words detection in a spontaneous spoken dialog system. In *Interspeech'03.*

Vinciarelli A., Luettin J. (2001). A new normalization technique for cursive handwritten words. *Pattern recognition letters*, vol. 22, n° 9, p. 1043-1050.

Werbos P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, vol. 1, n° 4, p. 339–356.

Wessel F., Schlüter R., Macherey K., Ney H. (2001). Confidence measures for large vocabulary continuous speech recognition. *IEEE TSAP*, vol. 9, p. 288-298.

Whitelaw C., Hutchinson B., Chung G., Ellis G. (2009). Using the web for language independent spellchecking and autocorrection. In *Emnlp*, p. 890-899.

Williams R. J., Zipser D. (1995). *Gradient-based learning algorithms for recurrent networks and their computational complexity.*