# Utilisation du Web pour la reconnaissance de mots manuscrits hors-vocabulaire

**Cristina Oprean*** — **Laurence Likforman-Sulem*** — **Adrian Popescu**** — **Chafic Mokbel*****

* *Institut Mines-Telecom/Telecom ParisTech and CNRS LTCI, 46 rue Barrault, 75013 Paris (France), oprean@telecom-paristech.fr*
** *CEA, LIST, LVIC, 91190 Gif-sur-Yvette (France), adrian.popescu@cea.fr*
*** *University of Balamand, Faculty of Engineering, P.O. Box 100 Tripoli (Liban), chafic.mokbel@balamand.edu.lb*

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 la couverture du dictionnaire. Si la taille est limitée, beaucoup de mots hors vocabulaire (HV) restent non reconnus. Pour améliorer la reconnaissance des mots HV, sans augmenter le dictionnaire, nous introduisons une étape supplémentaire qui exploite des ressources Web. Après une première classification des séquences des caractères 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 avec des expériences menées sur la base RIMES. Les résultats montrent que des améliorations sont obtenues par rapport à la reconnaissance avec un dictionnaire statique.*

ABSTRACT. *Handwriting recognition systems rely on predefined classifiers. Small and static dictionaries are usually exploited to obtain high in-vocabulary (IV) accuracy at the expense of coverage. Thus the recognition of out-of-vocabulary (OOV) words cannot be handled efficiently. To improve OOV recognition while keeping IV dictionaries small, we introduce a multi-step approach that exploits Web resources. After an initial IV-OOV classification, external resources are used to create OOV sequence-adapted dynamic dictionaries. A final CTC-based decoding is performed over the dynamic dictionary to determine the most probable word for the OOV sequence. We validate our approach with experiments conducted on the RIMES dataset. Results show that improvements are obtained compared to standard handwriting recognition.*

MOTS-CLÉS : *reconnaissance d'écriture manuscrite, dictionnaires dynamiques, Wikipédia*

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

CIFED 2014, pp. 217–232, Nancy, 18-21 mars 2014

## 1. Introduction

Les systèmes de reconnaissance vocale et de l'écriture manuscrite dépendent fortement des ressources linguistiques disponibles, telles que des dictionnaires (listes de mots) ou des modèles de langage (n-grammes par exemple). Quand les dictionnaires sont utilisés, la performance de ces systèmes est tributaire à un choix approprié de la taille du dictionnaire. Les dictionnaires de petite taille génèrent des résultats médiocres à cause de la modélisation d'un faible nombre de mots. Inversement, pour les dictionnaires très riches, un grand nombre de confusions conduira également à un faible taux de reconnaissance. En outre, la complexité calculatoire de la reconnaissance augmente avec la taille du dictionnaire et l'utilisation de grands dictionnaires est rendue difficile dans les applications du monde réel.

Les systèmes de reconnaissance sont classiquement implémentés à l'aide de modèles de Markov cachés (HMM) qui correspondent à des séquences observées des mots du dictionnaire. Dernièrement, 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 [LIW 07]. Contrairement aux HMMs, qui sont des modèles génératifs, les réseaux de neurones sont discriminatifs. Au vu de leurs performances supérieures [GRA 12] nous choisissons de travailler avec des RNN et, plus précisément, des réseaux de neurones bidirectionnels avec des cellules de mémoire (*BLSTM - Bidirectional Long Short Term Memory*). Ceux-ci prennent en compte l'information contextuelle en parcourant une image de gauche à droite et de droite à gauche.

Une hypothèse de monde clos peut être faite afin d'apparier toutes les séquences testées avec des mots du dictionnaire et on parle de reconnaissance avec des modèles de mots. La limite la plus importante de ce type d'approche est que seulement les mots du dictionnaire seront reconnus correctement. Alternativement, une partie des séquences testées 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, 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.

Nous présentons une approche de reconnaissance de l'écriture manuscrite qui utilise des corpus ouverts du Web pour améliorer la reconnaissance des mots HV. La figure 1 donne un aperçu de l'approche. Chaque séquence est décodée avec les modèles à base de mots et de caractères et est classée en DV ou HV en se basant sur la log-vraisemblance de ces deux types de mots. À ce stade, les séquences classées comme DV sont affectés à l'un des mots du dictionnaire statique. Les séquences classées comme HV sont comparées aux ressources Web (Wikipédia) afin de leur associer des dictionnaires dynamiques. Des méthodes d'amélioration de la reconnaissance de mots HV sont proposées, en utilisant les sorties BLSTM. Enfin, un second décodage est exécuté pour récupérer l'élément du dictionnaire dynamique qui est le plus similaire à la séquence testée.

Il y a plusieurs façons de traiter les mots HV dans la littérature du domaine. 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 [KOE 03], 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 caractères arbitraires (modèles de remplissage) ou des modèles de n-grammes des caractères [BRA 00, BAZ 99]. 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 des diminuent de façon drastique, en raison de la confusion augmentée entre les caractères. Dans [BRA 00] une diminution d'environ 26% du taux de reconnaissance de mots est signalé lorsqu'il n'y a pas de dictionnaire.

Plus récemment, l'élargissement du vocabulaire a été réalisée en décomposant le lexique en se basant sur une analyse morphologique [HAM 13]. 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 reconnaissance 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 [PAR 10, OGE 09]. Le contexte local est utilisé pour récupérer des documents à partir du Web qui sont utilisés pour augmenter le lexique .

Le Web comprend une grande richesse de sources ouvertes qu'on peut utiliser pour produire des ressources linguistiques, y compris des dictionnaires pour la reconnaissance des mots HV dans les domaines de la parole ou de la correction orthographique [WHI 09]. Les principales différences entre notre travail et [WHI 09] viennent de la difficulté plus élevée de la reconnaissance des mots HV en écriture manuscrite par rapport à la détection des fautes d'orthographe et de la façon innovante de créer des dictionnaires dynamiques avec Wikipédia. Dans [OGE 08] et [PAR 10] le contexte local est exploité pour récupérer des documents sur le Web qui sont 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.

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 d'un grand nombre de domaines. Ensuite, la ressource est librement disponible et en constante croissance.

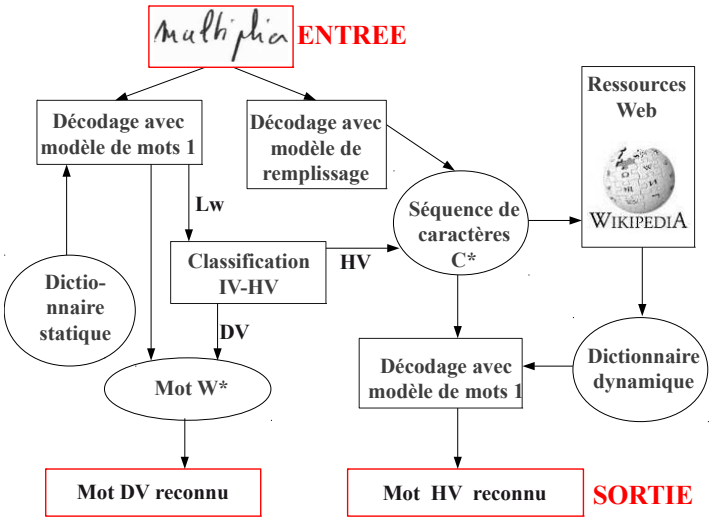



Figure 1 – Aperçu de l'approche proposée.

Le papier est organisé de la manière suivante : La section 2 présente le système de reconnaissance BLSTM. La section 3 introduit la classification des DV et HV mots. La création de dictionnaires dynamiques en utilisant Wikipédia est décrite dans la section 4. Enfin, la section 5 est consacrée aux expériences pour la reconnaissance des mots sur la base des données Rimes, base accessible au public.

## 2. Description du système de reconnaissance

Afin de tenir compte des évolutions récentes dans le domaine, nous effectuons des expériences avec un système basé sur les réseaux bidirectionnels (BLSTM) [MOR 13] qui utilise un approche de fenêtre glissante. L'entrée du système est une séquence de fenêtres (trames). Chaque fenêtre est divisée en 20 cellules. Un ensemble de 37 caractéristiques est extrait pour chaque trame, augmenté avec leurs dérivées du premier ordre. 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 et 8 caractéristiques directionnelles correspondant à l'histogramme des gradients pour les 8 orientations de  de 0 à 7 $\pi/4$  , avec un pas de $\pi/4$. La fenêtre

glissante est de longueur $w = 9$ pixels. Elle est décalée de $\delta = 3$ pixels à chaque pas de temps.

Le classificateur BLSTM est composé de deux réseaux de neurones récurrents : en avant et en arrière. La valeur d'une unité de sortie au temps $t$ est la combinaison linéaire des sorties des couches cachées avant et arrière au temps $t$. Les unités de neurones cachés sont des blocs de mémoire appelés "long-short term memory" (LSTM). Les couches cachées en avant et en arrière contiennent chacune 100 blocs de mémoire. La couche de sortie est composée 79 neurones, correspondant aux différents caractères de l'alphabet français (majuscules minuscules et signes de ponctuation). L'apprentissage avec BLSTM est basé sur la méthode de descente du gradient. Après chaque époque d'entraînement, le taux d'erreur de reconnaissance est évalué sur un ensemble de validation. Afin d'éviter le surapprentissage, le réseau est arrêté si taux d'erreur ne diminue pas pendant 20 époques.

Le BLSTM calcule les sorties de réseau correspondant à chaque trame qui est à son tour associée à une classe de caractères. Ces sorties sont normalisées, offrant pour chaque classe de caractères, la probabilité a posteriori. Ensuite, un algorithme de passage de jeton arrière-avant, dénommé CTC (Connectionist Temporal Classification) prend ces probabilités a posteriori en entrée et fournit un mot du dictionnaire ou, dans le cas sans dictionnaire, une chaîne des caractères. Nous utilisons l'implémentation du CTC introduite dans les travaux de [GRA 06].

## 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 de log-vraisemblance pour la classification des mots DV-HV.

### 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 [BAZ 00], [BIS 05] ou des scores de confiance [WES 01], [SUN 03]. 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 [QUI 07], [FIS 10], [CUA 02].

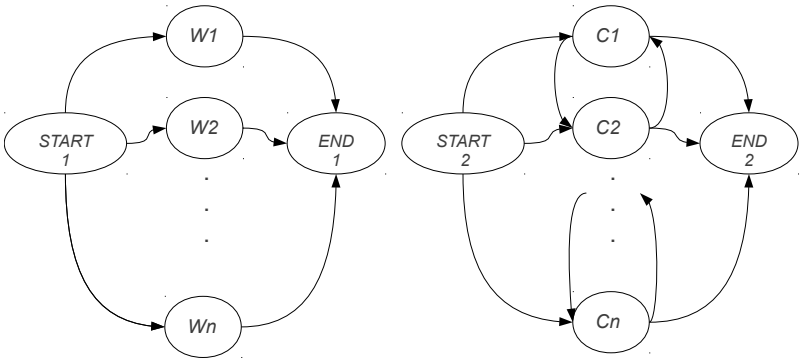



Figure 2 – Réseau des mots (a) réseau des caractères (b).

Dans un modèle de remplissage, les mots HV sont représentés par des réseaux formés de caractères indépendants, délimités par des modèles d'espace (fig. 2-b). Les mots DV sont représentés par des réseaux de mots plus sophistiqués construits sur un dictionnaire statique prédéfini (fig. 2-a). Pour une séquence donnée, la vraisemblance du vecteur des caractéristiques observées est calculée en utilisant le BLSTM.

### 3.2. *Classification DV-HV des images de mots*

Les séquences sont classées comme des DV ou HV en utilisant la log-vraisemblances de la trame calculée à l'aide du réseau des mots. Soit $X = x_1, x_2, ...x_N$ la séquence des trames observées, où est $N$ le nombre de trames. Sa valeur de log-vraisemblance $Lw$ est obtenue en utilisant l'algorithme de passage du jeton CTC sur les sorties du réseau BLSTM.
$Lw = log\ P(X|W^*) = \max_W \log\ P(X|W)$ où $W^*$ est le meilleur mot du dictionnaire pour la trame $X$.

La valeur $Lw$ est utilisé pour classer la séquence décodée comme HV ou DV. Si $\frac{Lw}{N} \geq thre$ 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 en utilisant une base de validation (section 5).

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

Nous faisons l'hypothèse que, en raison des mauvaises performances des modèles de remplissage (de caractères), la création de dictionnaires dynamiques qui

contiennent des chaînes similaires à une séquence HV détectée améliore sa reconnaissance. Lorsqu'une séquence est classée comme HV, la sortie obtenue avec le modèle des caractères est comparée au dictionnaire dynamique pour trouver le mot le plus probable qui correspond à la séquence candidate. La création des dictionnaires dynamiques repose sur Wikipédia. Les principaux avantages de Wikipédia sont sa disponibilité (en téléchargement libre) et sa couverture importante (elle contient des centaines de milliers d'articles pour la langue française).

### 4.1. *Dictionnaires dynamiques basés sur Wikipédia*

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 qui 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 entre la représentation TF-IDF de RIMES (voir la section 5) et la représentation TF-IDF de chaque article Wikipédia. 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 qui 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 3.

Le dictionnaire adapté favorise des termes relatifs au domaine. Le dictionnaire générique capture des propriétés plus génériques des mots et est plus compréhensif. Nous illustrons la sélection des termes avec le mot *facture*, un mot est très caractéristique pour la base de données RIMES. Il apparaît $1440$ fois dans l'ensemble des documents adaptés au domaine et de $1459$ fois dans le dictionnaires générique. Comme attendu, il apparaît de manière quasiment exclusive dans les documents les plus similaires à RIMES.

La distance Levenshtein [LEV 66] est 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 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

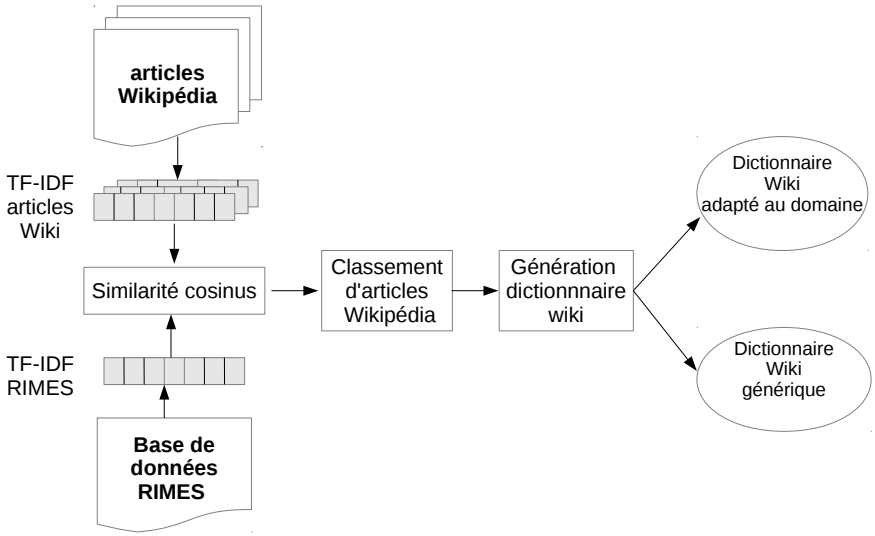
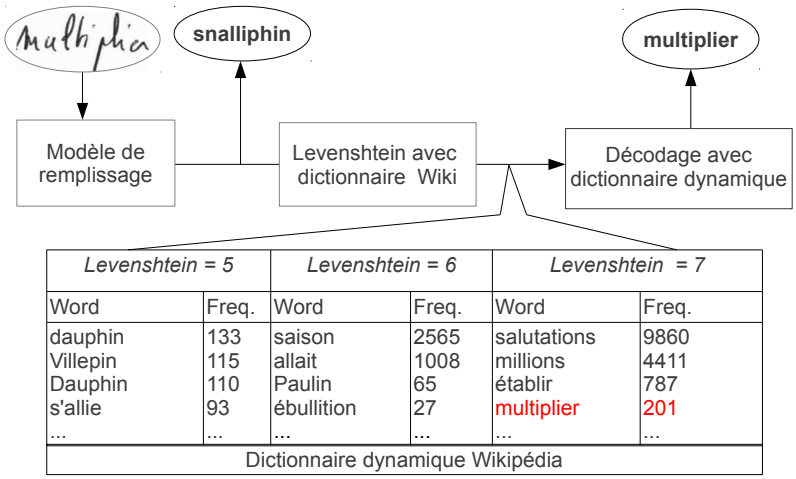



Figure 3 – Création des dictionnaires adaptés au domaine et génériques en utilisant Wikipédia.




| *Levenshtein = 5* | | *Levenshtein = 6* | | *Levenshtein = 7* | |
|---|---|---|---|---|---|
| Word | Freq. | Word | Freq. | Word | Freq. |
| dauphin | 133 | saison | 2565 | salutations | 9860 |
| Villepin | 115 | allait | 1008 | millions | 4411 |
| Dauphin | 110 | Paulin | 65 | établir | 787 |
| s'allie | 93 | ébullition | 27 | multiplier | 201 |
| ... | ... | ... | ... | ... | ... |

Dictionnaire dynamique Wikipédia

Figure 4 – Reconnaissance de mots hors vocabulaire basée sur Wikipédia.

la valeur $l$ permettrait de retenir plus de mots du dictionnaire Wikipédia mais 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 petite.

Les auteurs de [DAM 64] 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 de 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 mis la taille du dictionnaire à $k = 100, 200, 500, 1000$ 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* 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 fig. 4. Notez que le mot vrai *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 [GRO 11a] afin d'évaluer l'efficacité de la méthode de création des dictionnaires dynamiques proposée dans la section précédente. 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. 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 [GRO 09] et 2011 [GRO 11a]. 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.

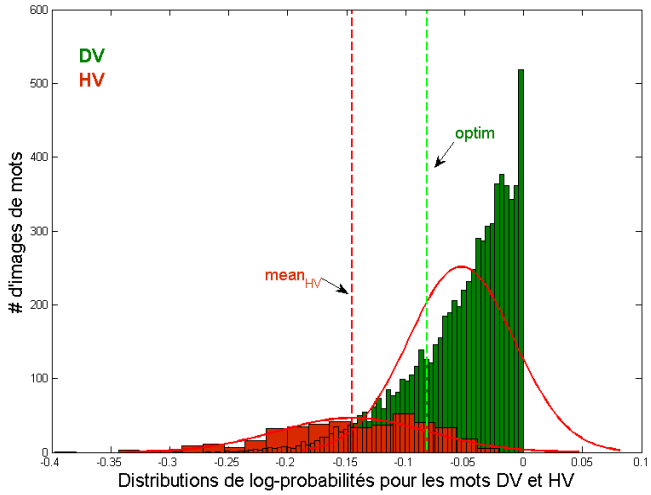



Figure 5 – Distributions des log-vraisemblances pour les mots HV et DV sur la base de validation RIMES 2011

### 5.1. *Détection de mots HV*

La classification DV-HV est fondée sur les distributions de log-vraisemblance des mots HV et DV, détaillée dans la section 3.2. L'ensemble de validation est utilisé pour régler les paramètres de classification des mots HV-DV et la base de test pour évaluer les performances du système. Environ 6% des mots de notre base de test sont des HV, n'étant pas présents dans l'ensemble d'apprentissage. 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. 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-vraisemblance 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.

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. Toutefois, la quantité 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 à grande échelle a été menée sur les 7 776 mots de la base de test RIMES 2011.

Lorsque $thre$=$mean_{DV}$, $3612$ images de mots ont été classées comme des HV, comparativement à $568$ lorsque $thre$=$mean_{HV}$. Les mots détectés comme des DV ont été déjà décodés avec le vocabulaire d'apprentissage.

Les mots classés comme des HV (leur score de log-probabilité normalisé $\frac{Lw}{N}$ < $thre$ ) sont traités avec la méthode décrite dans la section 4.1. Dans [OPR 13], nous avions montré que $thre$=$mean_{HV}$ maximise le taux de reconnaissance total. 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 proposé dans la section 3.2.

Les tailles des vocabulaires Wikipédia sont de 76 566 et 137 200 en ne considérant que les articles adaptés au domaine (cas "adapté") et tous les 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 d'un premier décodage du système de reconnaissance utilisant le 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 issue du 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 et les résultats préliminaires obtenus ont montré que les valeurs citées ci-dessus assurent un bon compromis entre le taux de reconnaissance et la complexité de calcul.

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.

| **Type du dictionnaire** | **Adapté** | | | | **Générique** | | | |
|---|---|---|---|---|---|---|---|---|
| Taille dict. dynamique | 100 | 200 | 500 | 1000 | 100 | 200 | 500 | 1000 |
| *séparation 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% |

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

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 quand la détection DV-HV est parfaite, le taux de reconnaissance avec un dictionnaire adapté au domaine est de 42.82 %, contre 42.60% pour le générique. Lorsque le seuil de séparation est $thre = mean_{HV}$ le meilleur taux de reconnaissance de 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.

| **Détection DV-HV** | **Taux de reconnaissance** [%] |
|---|---|
| DV pur | 75,45 |
| DV-HV réel | 75,87 |
| DV-HV théorique | **77.98** |

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

Dans le tableau 2, nous présentons une comparaison des résultats obtenus avec un système sans traitement des mots HV (DV pur), avec notre approche (DV-HV réel), ainsi qu'un résultat théorique (DV-HV théorique) dont le rôle est de montrer l'amélioration qui pourrait être atteinte dans le cas d'une séparation optimale DV-HV. Étant donné le chevauchement entre les distributions des mots DV-HV, la séparation obtenue est imparfaite mais une légère amélioration est toutefois obtenue par comparaison à une approche sans traitement des mots HV (75,87% vs. 75,45%). La différence entre DV-HV réel et DV-HV théorique 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 DV pur est égal à $[73.78\%, 77.11\%]$. Pour le système DV-HV réel, il est égal à $[74.19\%, 77.54\%]$ et pour le système DV-HV théorique, à $[76.25\%, 79.7\%]$ avec un niveau de risque $\alpha = 5\%$. Le résultat obtenu pour le système DV-HV théorique est significatif dans le cas de 6% de mots HV, car la valeur obtenue (77.98%) est hors de l'intervalle de confiance obtenu pour le système DV pur. Dans le cas où seulement 6% des mots sont HV l'amélioration d'environ 0.5% entre le système DV réel et le système DV pur n'est pas significative.

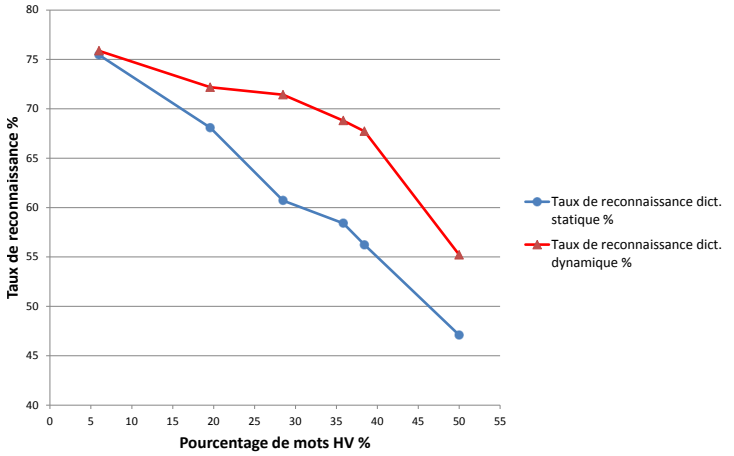



Figure 6 – Taux de reconnaissance en fonction du pourcentage de mots HV dans la base de test

Toutefois, elle peut devenir significative dans le cas où le nombre de mots HV est plus élevé. 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 DV pur et DV-HV réel (Figure 6) augmente avec le pourcentage de mots HV. Pour 50% de mots HV, le taux de reconnaissance passe de 47.08% (DV pur) à 55.20% (DV-HV réel). Ce résultat montre que l'intérêt de notre approche est d'autant plus grand que le nombre de mots hors vocabulaire augmente.

D'autres résultats sont disponibles pour cette base, notamment ceux de la compétition ICDAR 2011 [GRO 11b] 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 5740 mots issus de la base d'apprentissage et celle de test. Comme nous avons présenté dans la Section 5, notre dictionnaire contient environ 5000 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, basés sur des ressources externes à grande échelle.

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 classés 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 les ressources disponibles librement. Elle pourrait être appliquée à d'autres tâches qui impliquent 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. Tout d'abord, 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 [SHA 95, MOR 03], seront rajoutés au système dans l'avenir afin d'améliorer les performances.

Deuxièmement, tout en étant simple et efficace, le choix des mots similaires (calcul de la distance Levenshtein, nombre d'occurrences des mots dans les documents, la différence de longueur du mot trouvé avec la séquence des caractères HV) peut être amélioré en ajoutant d'autres méthodes. Vu que les sorties de BLSTM sans dictionnaire sont assez propres, une méthode de regroupement pourra être appliquée sur les mots sélectionnés initialement, 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 [KAU 97, SEN 96] 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 réduits 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.

Quatrièmement, l'amélioration de la méthode de classement des mots HV-DV s'avère importante, vu l'augmentation potentielle des performances (DV-HV théorique 77.98% vs. DV-HV réel 75.87%). Comme les deux distributions de log-probabilité sont superposées, d'autres critères de séparation doivent êtres considérés.

Enfin, le temps d'exécution pour la création d'un dictionnaire dynamique correspondant à une séquence de caractères candidate prend en moyenne 10.9s avec le dictionnaire adapté au domaine et 19.6s avec le dictionnaire générique. Cependant, ici l'accent n'était pas mis sur l'optimisation du calcul de la distance Levenshtein. Nous prévoyons d'optimiser le calcul de la distance Levenshtein pour une mise en oeuvre pratique du système.

## 7. Bibliographie

[BAZ 99] BAZZI I., SCHWARTZ R. M., MAKHOUL J., An Omnifont Open-Vocabulary OCR System for English and Arabic , *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, n° 6, 1999, p. 495-504.

[BAZ 00] BAZZI I., GLASS J. R., Modeling out-of-vocabulary words for robust speech recognition , *INTERSPEECH*, 2000, p. 401-404.

[BIS 05] BISANI M., NEY H., Open vocabulary speech recognition with flat hybrid models , *INTERSPEECH*, 2005, p. 725-728.

[BRA 00] BRAKENSIEK A., WILLETT D., RIGOLL G., Unlimited Vocabulary Script Recognition Using Character N-Grams , *In Proc. 22. DAGM-Symposium Tagungsband*, Springer-Verlag, 2000.

[CUA 02] CUAYÁHUITL H., SERRIDGE B., Out-of-Vocabulary Word Modeling and Rejection for Spanish Keyword Spotting Systems , *MICAI'02*, 2002.

[DAM 64] DAMERAU F., A technique for computer detection and correction of spelling errors , *Commun. ACM*, vol. 7, 1964, p. 171-176.

[FIS 10] FISCHER A., KELLER A., FRINKEN V., BUNKE H., HMM-based Word Spotting in Handwritten Documents Using Subword Models , *ICPR*, 2010, p. 3416-3419.

[GRA 06] GRAVES A., FERNÁNDEZ S., GOMEZ F. J., SCHMIDHUBER J., Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks , *ICML*, 2006, p. 369-376.

[GRA 12] GRAVES A., *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385 de *Studies in Computational Intelligence*, Springer, 2012.

[GRO 09] GROSICKI E., ABED H. E., ICDAR 2009 Handwriting Recognition Competition , *ICDAR*, 2009.

[GRO 11a] GROSICKI E., EL-ABED H., ICDAR 2011 : French Handwriting Recognition Competition , *ICDAR*, 2011.

[GRO 11b] GROSICKI E., EL-ABED H., ICDAR 2011-French Handwriting Recognition Competition , *Proc. of ICDAR'11*, IEEE, 2011, p. 1459–1463.

[HAM 13] HAMDANI M., EL-DESOKY MOUSA A., NEY H., Open Vocabulary Arabic Handwriting Recognition Using Morphological Decomposition , *International Conference on Document Analysis and Recognition*, Washington DC, 2013, p. 280-284.

[KAU 97] KAUFMANN G., BUNKE H., HADORN M., Lexicon Reduction in an HMM-Framework Based on Quantized Feature Vectors , *Proceedings of the 4th International Conference on Document Analysis and Recognition*, ICDAR '97, 1997, p. 1097–1101.

[KOE 03] KOERICH A. L., SABOURIN R., SUEN C. Y., Large Vocabulary Off-Line Handwriting Recognition : A Survey , *Pattern Analysis and Applications*, vol. 6, 2003, p. 97–121.

[LEV 66] LEVENSHTEIN V., Binary Codes Capable of Correcting Deletions, Insertions and Reversals , *Soviet Physics Doklady*, vol. 10, 1966, page 707.

[LIW 07] LIWICKI M., GRAVES A., BUNKE H., SCHMIDHUBER J., A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks , *In Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007*, 2007.

[MOR 03] MORITA M. E., Automatic Recognition of Handwritten Dates on Brazilian Bank Cheques , PhD thesis, 2003.

[MOR 13] MORILLOT O., LIKFORMAN-SULEM L., GROSICKI E., New baseline correction algorithm for text-line recognition with bidirectional recurrent neural networks , *Journal of Electronic Imaging*, vol. 22, n° 2, 2013, p. 023028-023028.

[OGE 08] OGER S., LINARES G., BÉCHET F., NOCERA P., On-demand new word learning using world wide web , *ICASSP*, 2008, p. 4305-4308.

[OGE 09] OGER S., POPESCU V., LINARÉS G., Using the world wide web for learning new words in continuous speech recognition tasks : Two case studies , *in SPECOM*, 2009.

[OPR 13] OPREAN C., LIKFORMAN-SULEM L., POPESCU A., MOKBEL C., Using the Web to Create Dynamic Dictionaries in Handwritten Out-of-Vocabulary Word Recognition , *ICDAR*, 2013, p. 989-993.

[PAR 10] PARADA C., SETHY A., DREDZE M., JELINEK F., A spoken term detection framework for recovering out-of-vocabulary words using the web , *INTERSPEECH*, 2010.

[QUI 07] QUINIOU S., ANQUETIL É., Use of a Confusion Network to Detect and Correct Errors in an On-Line Handwritten Sentence Recognition System , *ICDAR*, 2007, p. 382-386.

[SEN 96] SENI G., SRIHARI R. K., NASRABADI N., Large Vocabulary Recognition of On-Line Handwritten Cursive Words , *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, n° 7, 1996, p. 757–762.

[SHA 95] SHASTRI L., FONTAINE T., Recognizing Handwritten Digit Strings Using Modular Spatio-temporal Connectionist Networks. , *Connection Science*, vol. 7, 1995, p. 211-246.

[SUN 03] SUN H., ZHANG G., ZHENG F., XU M., Using word confidence measure for OOV words detection in a spontaneous spoken dialog system , *INTERSPEECH'03*, 2003.

[WES 01] WESSEL F., SCHLÜTER R., MACHEREY K., NEY H., Confidence measures for large vocabulary continuous speech recognition , *IEEE TSAP*, vol. 9, 2001, p. 288-298.

[WHI 09] WHITELAW C., HUTCHINSON B., CHUNG G., ELLIS G., Using the Web for Language Independent Spellchecking and Autocorrection , *EMNLP*, 2009, p. 890-899.