Ajout de nouveaux noms propres au vocabulaire d'un système de transcription en utilisant un corpus diachronique

Irina Illina * — Dominique Fohr * — Georges Linarès **

* *Équipe parole, LORIA-INRIA, 54602 Villers-les-Nancy, France*

irina.illina@loria.fr ; dominique.fohr@loria.fr

** *LIA, Université d'Avignon, 84911 Avignon, France*

georges.linares@univ-avignon.fr

RÉSUMÉ. Les noms propres sont souvent indispensables pour comprendre l'information contenue dans un document. Notre travail se concentre sur l'augmentation automatique du vocabulaire d'un système de transcription automatique de la parole (RAP) à partir d'un corpus diachronique. Nous faisons l'hypothèse que certains noms propres apparaissent dans des documents relatifs à la même période temporelle et dans des contextes lexicaux similaires. Trois méthodes de sélection de noms propres sont proposées pour augmenter de façon dynamique le vocabulaire en utilisant des informations lexicales et temporelles. Les méthodes sont fondées sur des statistiques de cooccurrences dans des fenêtres de taille fixe, sur l'information mutuelle et sur le modèle vectoriel. Différents paramètres de sélection de noms propres sont également étudiés afin de limiter l'augmentation du vocabulaire. Les résultats de reconnaissance montrent une réduction significative du taux d'erreur de noms propres en utilisant un vocabulaire augmenté.

ABSTRACT. Proper names are usually keys to understand the information contained in a document. Our work focuses on increasing the vocabulary size of a speech transcription system by automatically retrieving proper names from contemporary diachronic text corpus. We assume that some proper names appear in documents relating to the same time period and in similar lexical contexts. We proposed methods that dynamically augment the automatic speech recognition system vocabulary using lexical and temporal features. Three proposed selection methods are based on co-occurrences statistics inside windows of fixed size, on mutual information and on vector space model. Different metrics for proper name selection in order to limit the vocabulary augmentation are studied. Recognition results show a significant reduction of the proper name error rate using augmented vocabulary with retrieved proper names.

MOTS-CLÉS : reconnaissance de la parole, mots hors vocabulaire, noms propres, augmentation du vocabulaire.

KEYWORDS: speech recognition, out-of-vocabulary words, proper names, vocabulary augmentation.

1. Introduction

Nous nous intéressons à la reconnaissance automatique de la parole (RAP) et plus particulièrement à la transcription de documents audio. Même en utilisant un très grand vocabulaire, les systèmes RAP sont confrontés au problème des mots hors vocabulaire (*Out Of Vocabulary*, OOV). Ces mots OOV sont des mots qui se trouvent dans le signal de parole, mais pas dans le vocabulaire du système RAP. Le système RAP ne pourra pas les transcrire correctement et les remplacera par un ou plusieurs mots du vocabulaire, touchant négativement l'intelligibilité de la transcription.

Les noms propres (NP) sont en constante évolution et aucun vocabulaire statique ne pourra contenir tous les noms propres existants : les NP représentent environ 10 % des mots des articles de journaux en anglais ou en français et ils sont vitaux pour caractériser le contenu d'un texte (Friburger *et al*., 2002). Bechet et Yvon dans (Bechet *et al*., 2000) ont montré que 72 % des mots OOV sont des NP dans le cas d'un vocabulaire de 265 k mots.

Dans cet article, notre but est d'ajouter des NP pertinents dans le système de reconnaissance pour diminuer le taux d'erreur en mot. Pour cela trois étapes sont nécessaires :

- sélectionner des NP pertinents ;
- les ajouter dans le vocabulaire et les phonétiser ;
- les ajouter dans le modèle de langage.

Dans cet article, nous nous focalisons sur la première partie, c'est-à-dire la sélection des noms propres pertinents.

1.1. *Sélection des NP pertinents*

Pour rechercher de nouveaux mots, il faut disposer d'un corpus de texte, qui peut être soit un ensemble de textes déjà disponibles, soit des documents collectés sur Internet (Yu *et al.*, 2000). Kemp *et al.* (1998) et Manning *et al.* (2008) utilisent des techniques issues de la recherche d'information pour sélectionner des articles collectés sur Internet.

Pour sélectionner les mots pertinents à partir de ces documents, plusieurs méthodes ont été proposées : des méthodes fondées sur les fréquences ou les cooccurrences et d'autres s'appuyant sur des représentations continues des mots. Allauzen *et al.* (2003) ont proposé une adaptation vectorielle du vocabulaire qui optimise directement la couverture lexicale sur un corpus de développement par combinaison linéaire des fréquences de mots calculés sur des corpus d'entraînement. Ohtsuki *et al.* (2005b) calculent la pertinence d'un mot en se fondant sur la notion de *concept vectors*, qui sont calculés à partir des cooccurrences de mots. Bigot *et al.*

(2013a) utilisent une représentation continue qui tient compte de la position du mot dans la phrase. Dans notre travail, nous avons étudié ces deux représentations de l'espace, l'une continue et l'autre discrète.

Dans le cadre de la sélection des NP, nous nous concentrons sur l'exploitation du contexte lexical et des informations temporelles de documents issus d'un corpus *diachronique*[1] (Allauzen *et al.*, 2005). Notre hypothèse est que l'information temporelle est un élément important pour capturer des dépendances NP-contexte (Kobayashi *et al*., 1998). Notre approche a été inspirée par Bigot *et al.* (2013) et Oger *et al.* (2008). Oger propose de chercher les OOV à partir du Web en utilisant comme requête les n-grams mal reconnus et donc le contexte très local du OOV. Bigot utilise la notion de contexte lexical pour trouver les noms de personnes OOV à partir de la sortie du système RAP et des documents supplémentaires, mais sans localisation temporelle des contextes. Dans notre article, nous utilisons également la notion du contexte pour les noms propres, mais nous nous concentrons sur l'exploitation de la temporalité des documents en utilisant un corpus diachronique. Nous supposons que les NP sont souvent liés à un événement qui se produit dans une période de temps spécifique. Nous émettons l'hypothèse que les NP évoluent dans le temps, et que pour une période donnée, les mêmes NP vont apparaître conjointement dans des documents qui sont produits dans la même période. Cette idée de temporalisation du lexique de reconnaissance a été évaluée notamment par Bertoldi *et al.* (2001), Federico *et al.* (2004) et Martins *et al.* (2010) qui proposent une mise à jour régulière du modèle de langage et du lexique, puis par Parada *et al.* (2010) pour la prédiction des OOV dans les sorties d'un système de reconnaissance. Ces approches sont des approches *a priori*, qui augmentent le lexique avant tout décodage sur la seule base des dates de production des documents. Ce type de technique présente l'inconvénient d'une augmentation très large du lexique, qui ne tient pas compte du contexte d'apparition des mots manquants. Notre proposition est d'utiliser une première passe de décodage pour extraire une information relative au contexte lexical temporalisé des OOV qui devrait permettre de modéliser plus précisément le contexte des mots manquants et d'éviter l'augmentation excessive du vocabulaire.

Pour obtenir un bon compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire, nous proposons différentes méthodes de filtrage de NP fondées sur des cooccurrences dans des fenêtres, sur l'information mutuelle (Church *et al.*, 1989) et sur le modèle vectoriel[2] (Salton *et al.*, 1975) avec comme distance la similarité cosinus. L'extension du vocabulaire est faite de façon dynamique, spécifique pour chaque fichier, afin d'ajouter des noms propres en relation avec le

1. Un corpus de documents datés permettant d'étudier l'évolution des actualités et des noms propres qui en dépendent.

2. *Vector Space Model*

contenu et la date du document à transcrire et ainsi d'éviter un accroissement excessif du vocabulaire.

1.2. *Phonétisation des noms propres sélectionnés*

Dans un grand nombre de langues, dont le français, la difficulté de la conversion graphèmes-phonèmes est due au fait qu'il n'y a pas de correspondance directe entre la graphie d'un mot et la suite de phonèmes correspondant à sa prononciation. Une autre difficulté est que certains mots, surtout les noms propres, peuvent avoir plusieurs prononciations possibles. Les principales méthodes utilisées pour la conversion graphèmes-phonèmes sont : des systèmes à base de règles, des arbres de décision (Pagel *et al.,* 1998), des réseaux de neurones (Jensen *et al.,* 2000), des modèles de Markov (Taylor, 2005), des *Joint-Multigram Models* (JMM) (Bisani *et al.,* 2008) et des champs conditionnels aléatoires (*Conditional Random Fields*, CRF) (Jouvet *et al*., 2012). Dans notre travail, nous avons choisi d'utiliser les CRF car ils donnent des performances similaires voire supérieures à celles obtenues avec les JMM qui sont considérés comme une méthode « état de l'art ».

1.3. *Ajout des noms propres dans le modèle de langage*

Concernant l'ajout des mots dans le modèle de langage, plusieurs solutions sont envisageables. La solution la plus simple serait d'affecter à chacun des nouveaux noms propres ajoutés, la probabilité du mot inconnu (« *unk* »). Les inconvénients majeurs de cette méthode sont que la même probabilité est affectée à tous les nouveaux NP et que cette probabilité est fortement surestimée (il y a de très nombreuses occurrences de « *unk* » dans le corpus d'apprentissage). Le risque est d'obtenir de nombreuses insertions de ces nouveaux NP lors de la reconnaissance. La deuxième solution consiste à adapter le modèle de langage en utilisant un corpus de textes contenant ces nouveaux NP (Bellegarda, 2004). La sélection du corpus d'adaptation peut être effectuée à l'aide de méthodes issues de la recherche d'information (Chen *et al*., 2003). Une autre solution, plus coûteuse en temps de calcul consiste à réestimer entièrement le modèle de langage pour ce nouveau vocabulaire augmenté. La meilleure façon d'intégrer les nouveaux noms propres dans le modèle de langage n'entre pas dans le champ d'application de cet article.

Cet article est organisé de la façon suivante : la section 2 présente la méthodologie proposée pour la sélection des nouveaux NP à partir du corpus diachronique. La section 3 décrit les expériences réalisées dont les résultats sont discutés dans la section 4.

2. Méthodologie

Notre idée consiste à extraire des noms propres OOV automatiquement à partir du corpus diachronique, en utilisant le contexte lexical et temporel. Nos méthodes d'extraction des noms OOV sont fondées sur l'idée que les noms propres manquants se retrouvent probablement dans des documents contemporains, c'est-à-dire correspondant à la même période de temps que le document que nous voulons transcrire. Nous émettons l'hypothèse que les NP évoluent dans le temps et que pour une date donnée, les mêmes noms propres apparaîtront dans les documents qui appartiennent à la même période. Par exemple, dans un document de mars 2011 contenant les NP « *Japon* » et « *Fukushima* », il y a de fortes chances que les NP « *TEPCO* », « *Daiichi* » et « *Naoto Kan* » apparaissent ensemble.

À partir du corpus diachronique, nous proposons d'extraire un sous-corpus synchronique composé des documents qui sont contemporains d'un document à transcrire et de construire un vocabulaire augmenté pour ce document audio à transcrire. Dans la suite de cet article, nous appellerons « *documents sélectionnés* » ces documents du sous-corpus synchronique. En résumé, nous avons un document audio à transcrire qui contient des mots OOV et nous disposons d'un corpus diachronique de textes, utilisé pour rechercher de nouveaux NP. Un vocabulaire augmenté est construit pour chaque document à transcrire car les noms propres ajoutés sont spécifiques à la date et au contenu lexical du document à transcrire. Cette extension dynamique du vocabulaire est utile pour traiter des documents couvrant différentes périodes temporelles et des thématiques variées.

Nous supposons que, pour une date donnée, un nom propre du document à transcrire apparaîtra avec d'autres NP des *documents sélectionnés* correspondant à la même période de temps. Parmi ces NP, nous faisons l'hypothèse qu'un certain nombre de NP seront présents dans le document à transcrire, c'est-à-dire seront des NP OOV. L'idée est d'exploiter la relation entre les NP et leur contexte temporel et lexical pour un enrichissement du vocabulaire.

Dans cet article, différentes stratégies de sélection de NP sont proposées pour construire ce vocabulaire augmenté :

– la *méthode de référence* : sélection des documents du corpus diachronique en utilisant uniquement une période de temps correspondant au document à transcrire ;

– la *méthode fondée sur des cooccurrences dans des fenêtres contextuelles :* même stratégie que la méthode de référence mais les NP sont filtrés en tenant compte de leurs proximités et de leurs cooccurrences ;

– la *méthode fondée sur l'information mutuelle* : même stratégie que la méthode de référence, mais l'information mutuelle est utilisée pour sélectionner les noms propres à partir des *documents sélectionnés* ;

– la *méthode fondée sur le modèle vectoriel avec comme distance la similarité cosinus :* même stratégie que la méthode de référence mais les documents sont représentés par le modèle vectoriel (Shingal *et al.*, 2011).

Dans une étude précédente (Nkairi *et al.*, 2013), nous avons présenté les résultats de la méthode de référence et nous avons proposé la méthode fondée sur l'information mutuelle. Dans (Illina *et al.*, 2014) la méthode fondée sur la similarité cosinus a été introduite et quelques résultats préliminaires présentés. Dans le présent article, nous développons plus largement ces deux études : nous avons étudié le comportement des trois méthodes proposées en fonction de différents paramètres, nous avons introduit un corpus de développement afin de sélectionner aux mieux les paramètres et élargi le vocabulaire standard initial. De plus, nous avons proposé un protocole expérimental **contrastif** afin de valider les hypothèses émises.

2.1. *Méthode de référence*

Cette méthode consiste à extraire de nouveaux noms propres à partir du corpus diachronique en utilisant uniquement la date du document à transcrire. Aucun filtrage n'est appliqué dans ce cas et le vocabulaire augmenté risque donc d'être très grand.

2.1.1. *Extraction des NP des* documents sélectionnés

Les documents contemporains du document à transcrire, c'est-à-dire ceux qui correspondent à la même période de temps, sont pris en compte. Un étiquetage morphosyntaxique de ces documents à l'aide de *TreeTagger* (Schmit, 1994) est effectué. Puis nous extrayons les mots qui ont été étiquetés comme « nom propre ». Parmi ceux-ci, nous constituons la liste des nouveaux NP (*listeNouveauxNP*), c'est-à-dire ceux qui ne sont pas dans le vocabulaire standard. Nous supposons que si un nouveau nom propre est présent dans un *document sélectionné* contemporain du document à transcrire, il est possible qu'il soit présent dans ce dernier.

2.1.2. *Augmentation du vocabulaire*

Ensuite, notre vocabulaire est augmenté avec tous les NP de cette liste (*listeNouveauxNP*). Pour mieux tenir compte des aspects temporels, un vocabulaire augmenté est construit pour chaque document à transcrire. Les prononciations de ces NP sont générées en utilisant un dictionnaire phonétique ou un outil graphème-phonème (Illina *et al.*, 2011).

Nous considérons cette méthode comme une méthode de référence car tous les nouveaux NP apparaissant dans les *documents sélectionnés* sont pris en compte. Le problème de cette approche est que si le corpus diachronique est de grande taille, nous risquons d'augmenter la taille du vocabulaire de façon démesurée.

La faiblesse de cette méthode est qu'aucune information sur le document à transcrire, à part sa date, n'est utilisée ici. Dans les méthodes présentées dans les sections suivantes, le contenu lexical du document à transcrire sera pris en compte pour sélectionner des noms propres plus pertinents.

2.2. *Méthode fondée sur des cooccurrences dans des fenêtres contextuelles*

Pour obtenir un meilleur compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire, nous allons filtrer les NP de la *listeNouveauxNP* en utilisant le contenu du document à transcrire. De plus, nous allons utiliser la notion de cooccurrence dans une fenêtre de N mots. Nous faisons l'hypothèse que si un nouveau nom propre apparaît fréquemment à proximité d'un nom propre présent dans le document à transcrire, c'est un bon candidat pour étendre le vocabulaire.

2.2.1 *Extraction des NP de chaque document à transcrire*

Tout d'abord chaque document est transcrit en utilisant le système de RAP et le vocabulaire standard (cf. section 3.1). Il est fréquent que cette transcription contienne un certain nombre d'erreurs, en particulier les noms propres OOV (qui ne peuvent pas être reconnus par le système). Ensuite, pour chaque document transcrit, nous extrayons la liste *listeNPancrage* de tous les NP qui sont présents dans cette transcription automatique. Pour cela, nous ne pouvons pas utiliser *TreeTagger* car la transcription automatique ne contient que les mots en minuscules. Nous avons étiqueté chaque mot du vocabulaire standard en termes de NP et non-NP en utilisant le dictionnaire de noms communs BDLex de Calmès *et al*. (1998). Pour certains mots, il faut faire un choix, car ils peuvent être à la fois un nom propre et un nom commun : par exemple « *sucre* » est habituellement un nom commun sauf dans « *Pain de Sucre à Rio* ». Nous avons fait le choix de privilégier la précision au rappel et donc nous considérons « *sucre* » comme un nom commun.

L'objectif est d'utiliser la *listeNPancrage* comme un point d'ancrage pour chercher de nouveaux noms propres dans le corpus diachronique.

2.2.2. *Extraction du contexte à partir des* documents sélectionnés

Seuls les documents qui correspondent à la même période de temps que le document à transcrire sont pris en compte. Comme dans la méthode de référence, un étiquetage morphosyntaxique de ces documents à l'aide de *TreeTagger* est effectué.

Afin de limiter l'augmentation excessive du vocabulaire, pour chaque NP de la *listeNPancrage* nous parcourons tous les *documents sélectionnés.* Pour chaque occurrence de ce NP dans un *document sélectionné*, nous nous limitons aux mots situés dans une fenêtre de taille fixe centrée sur ce NP. Dans cette fenêtre, nous sélectionnons tous les mots qui ont été étiquetés comme NP par le *TreeTagger* et qui

sont des nouveaux NP, c'est-à-dire ceux qui ne sont pas dans le vocabulaire standard.

2.2.3. *Augmentation du vocabulaire*

Parmi ces nouveaux NP, seuls ceux qui ont une fréquence d'apparition supérieure à un seuil donné sont retenus et ajoutés au vocabulaire standard. En utilisant cette méthodologie, nous nous attendons à extraire une liste réduite (par rapport à la méthode de référence) des NP potentiellement manquants. Comme pour la méthode de référence, un vocabulaire augmenté est construit pour chaque document à transcrire. Les prononciations de ces nouveaux NP sont générées de la même façon que pour la méthode de référence.

Par rapport à la méthode de référence, les NP du document à transcrire, reconnus par le système RAP, sont utilisés dans cette méthode pour guider notre filtrage.

2.3. *Méthode fondée sur l'information mutuelle*

Dans cette section, nous proposons une autre méthode de filtrage des nouveaux NP fondée sur l'information mutuelle, c'est-à-dire sur la dépendance statistique au sens probabiliste entre les NP. Par rapport à la méthode précédente, seule l'étape décrite en 2.3.1. est modifiée.

2.3.1. *Extraction du contexte à partir des* documents sélectionnés

La liste *listeNPancrage* est obtenue de la même façon que précédemment. Afin de limiter l'augmentation excessive du vocabulaire, nous proposons d'utiliser l'information mutuelle. Nous calculons cette information entre les NP de la *listeNPancrage* (qui appartiennent au vocabulaire standard du système de reconnaissance) et les NP des documents du sous-corpus synchronique. Seuls les documents qui correspondent à la même période de temps que le document à transcrire sont pris en compte. Si deux NP ont une information mutuelle élevée, cela augmente la probabilité qu'ils apparaissent tous les deux dans le document à transcrire.

L'information mutuelle de deux variables aléatoires est une grandeur qui mesure la dépendance mutuelle des deux variables aléatoires. Formellement, l'information mutuelle de deux variables aléatoires discrètes *X* et *Y* est définie comme :

$$I(X;Y) = \sum_{x,y} p(X = x, Y = y)\, log\left(\frac{p(X = x, Y = y)}{p(X = x)p(Y = y)}\right)$$

Dans notre cas, *X* et *Y* représentent des noms propres et *x = 1*, si *x* est présent dans le document, et *x = 0* sinon.

Enfin, on calcule l'information mutuelle entre toutes les combinaisons de la variable *X* (*X* est un NP de la *listeNPancrange*) et la variable *Y* (*Y* est un nouveau

NP de la *listeNouveauxNP*). Pour un nouveau NP *Y*, l'information mutuelle résultante est calculée de la façon suivante :

$$I(Y) = \max_X I(X;Y)$$

Notre hypothèse est la suivante : plus la probabilité de dépendance statistique de deux noms propres dans le corpus diachronique est grande, plus la probabilité de leur apparition dans le document à transcrire est élevée. Les nouveaux NP dont l'information mutuelle est supérieure à un seuil sont retenus (noté *seuilMI* dans les tableaux, cf. section 4.3). Afin de sélectionner les NP les plus pertinents, nous éliminons ceux qui ont une faible fréquence d'apparition dans les *documents sélectionnés*. Les prononciations de ces NP sont générées comme précédemment.

En utilisant cette méthodologie, nous nous attendons à extraire une liste réduite (par rapport à la méthode de référence) des NP potentiellement manquants.

2.4. *Méthode fondée sur le modèle vectoriel avec comme distance la similarité cosinus*

Dans cette méthode nous souhaitons prendre en compte des informations lexicales supplémentaires pour modéliser le contexte : nous allons utiliser non seulement les noms propres (comme dans les méthodes précédentes) mais aussi les verbes, les adjectifs et les noms communs. Ces mots sont extraits à l'aide de *TreeTagger*. Ils sont lemmatisés car nous sommes intéressés par l'information sémantique. Les autres mots sont retirés.

Dans cette méthode nous proposons d'utiliser la notion de modèle vectoriel (Singhal, 2001). C'est une méthode algébrique de représentation du contenu d'un document textuel dans laquelle les documents sont généralement représentés par des *vecteurs de mots*. La proximité entre les documents est souvent calculée en utilisant la similarité cosinus. Donc nous allons représenter les documents à transcrire et les *documents sélectionnés* sous la forme d'un *vecteur de mots* et utiliser la similarité cosinus entre les modèles vectoriels pour extraite les NP pertinents.

2.4.1. *Génération des vecteurs de mots de chaque document à transcrire*

Comme dans la méthode précédente, chaque document est transcrit en utilisant le système de RAP et le vocabulaire standard. Puis, chaque document à transcrire est représenté par l'histogramme des occurrences des mots le composant, c'est-à-dire par un vecteur de mots (sac de mots, BOW[3]). Comme cela a été dit précédemment, seuls les verbes, les adjectifs, les noms propres et les noms communs lemmatisés sont pris en compte.

3. *BOW : Bag Of Word*

2.4.2. *Extraction du contexte à partir des* documents sélectionnés

Chaque *document sélectionné* est également représenté par son *vecteur de mots* de la même façon que précédemment. Puis, nous constituons la liste des nouveaux NP (*listeNouveauxNP*) en choisissant dans les *documents sélectionnés* les mots qui ont été étiquetés comme « *nom propre* » et qui ne sont pas dans le vocabulaire standard. Cette liste est constituée de la même façon que dans les méthodes précédentes.

Pour chaque NP appartenant à cette *listeNouveauxNP*, nous calculons un *vecteurNP*. Pour cela, tout d'abord, un lexique commun est construit : il contient la liste des mots (verbes, adjectifs, noms communs et noms propres) qui apparaissent au moins une fois dans les *documents sélectionnés* ou dans le document à transcrire. Ensuite, tous les *vecteurs de mots* sont projetés sur ce lexique commun. Finalement, le *vecteurNP* est calculé comme étant la somme des *vecteurs de mots* des *documents sélectionnés* dans lesquels ce nouveau NP apparaît.

Pour chaque NP de la *listeNouveauxNP*, nous calculons la similarité cosinus entre son *vecteurNP* et le *vecteur de mots* (projeté sur le lexique commun) du document à transcrire. Les nouveaux NP dont la similarité cosinus est supérieure à un seuil sont sélectionnés (noté *seuilCos* dans les tableaux, cf. section 4.4).

2.4.3. *Augmentation du vocabulaire*

Les NP sélectionnés sont ajoutés au vocabulaire de chaque document à transcrire. Les prononciations de ces nouveaux NP sont générées comme précédemment.

Par rapport aux méthodes précédentes, la méthode cosinus prend en compte une information contextuelle lexicale plus large en utilisant en plus les noms propres, les verbes, les adjectifs et les noms communs présents dans les *documents sélectionnés* et dans le document à transcrire.

3. Expériences

3.1. *Corpus de développement et de test*

Pour ajuster les paramètres des méthodes proposées, sept documents audio du corpus de développement d'ESTER2 sont utilisés comme corpus de développement. La méthodologie proposée est validée sur un corpus de test composé de treize documents audio du corpus de test d'ESTER2 (voir le tableau 1). L'objectif de cette campagne était d'évaluer la transcription automatique d'émissions de radio en français (Galliano *et al.*, 2009). La campagne ciblait une large variété d'émissions : bulletins d'information, reportages, débats, etc.

Développement	Test
Dev1 2007/07/07 rfi	**Test1** 2007/12/18 fr-inter
Dev2 2007/07/10 rfi	**Test2** 2007/12/20 fr-inter
Dev3 2007/07/10 fr-inter	**Test3** 2007/12/21 fr-inter
Dev4 2007/07/11 fr-inter	**Test4** 2008/01/17 fr-inter
Dev5 2007/07/12 fr-inter	**Test5** 2008/01/18 fr-inter
Dev6 2007/07/16 fr-inter	**Test6** 2008/01/18 rfi
Dev7 2007/07/23 fr-inter	**Test7** 2008/01/22 rfi
	Test8 2008/01/22 rfi
	Test9 2008/01/23 rfi
	Test10 2008/01/24 fr-inter
	Test11 2008/01/24 rfi
	Test12 2008/01/25 rfi
	Test13 2008/01/28 rfi

Tableau 1. *Dates des documents de développement et de test*

Le tableau 2 présente les occurrences de tous les NP (appartenant au vocabulaire standard et hors vocabulaire) dans chaque document de développement par rapport au vocabulaire standard. Cela conduit à un taux de NP OOV d'environ 1,3 % (401/31681). Le tableau 3 présente les mêmes informations pour le corpus de test.

Fichiers	Nombre d'occ. de mots	Nombre de NP appartenant au vocabulaire	Nombre d'occ de NP appartenant au vocabulaire	NP OOV	Nombre d'occ. NP OOV
Dev1	5 473	171	333	61	93
Dev2	3 020	103	200	32	39
Dev3	3 891	109	211	32	57
Dev4	3 745	109	225	31	52
Dev5	3 749	101	198	30	53
Dev6	3 757	37	64	14	57
Dev7	8 046	64	217	15	50
Moyenne	4 525,9	99,1	164,0	30,7	57,3

Tableau 2. *Couverture des noms propres du corpus de développement*

Fichiers	Nombre d'occ. de mots	Nombre de NP appartenant au vocabulaire	Nombre d'occ. de NP appartenant au vocabulaire	NP OOV	Nombre d'occ. de NP OOV
Test1	4 254	122	282	23	34
Test2	4 027	117	213	33	66
Test3	4 464	122	252	32	56
Test4	10 655	63	128	17	42
Test5	9 008	121	268	44	95
Test6	1 752	82	130	25	30
Test7	1 426	48	71	14	33
Test8	1 866	80	130	28	36
Test9	1 931	92	129	24	32
Test10	7 170	74	391	21	67
Test11	1 954	72	99	29	50
Test12	2 001	86	118	22	31
Test13	1 813	86	125	26	34
Moyenne	4 024,7	89,6	179,7	26,0	46,6

Tableau 3. *Couverture des noms propres du corpus de test*

3.2. *Corpus diachronique*

Comme corpus diachronique, nous avons utilisé le corpus GigaWord, qui contient des documents produits par l'Agence France-Presse (AFP) et l'Associated Press Worldstream (APW). Ce corpus français est une archive de dépêches de presse : pour l'AFP de mai 1994 à décembre 2008, pour l'APW de novembre 1994 à décembre 2008. Le choix de GigaWord a été motivé par le fait qu'il est contemporain du corpus ESTER2, qu'il est rédigé dans le même style (journalistique) et qu'il traite de domaines similaire (politique, sports, etc.). De plus, sa granularité temporelle est fine car journalière.

3.3. *Système de transcription*

Le système ANTS (*Automatic News Transcription System*, (Illina *et al.*, 2004)) utilisé pour ces expériences est fondé sur des modèles HMM dépendants du contexte, appris sur un corpus audio de deux cents heures d'émissions de radio. Le moteur de reconnaissance est Julius (Lee *et al.*, 2009).

Le lexique de notre système ANTS contient 122 k mots. Afin d'augmenter artificiellement le taux de mots hors vocabulaire, nous avons retiré de ce vocabulaire 223 noms propres qui étaient présents dans les corpus de développement et de test et qui n'étaient pas présents dans le corpus d'apprentissage d'ESTER2. Ce corpus d'apprentissage correspond à une période temporelle antérieure à la période des documents de développement et de test. Dans la suite de l'article nous appellerons ce vocabulaire le vocabulaire standard. Le lexique phonétique de ce vocabulaire contient 260 k prononciations pour les 122 k mots. Le modèle de langage quadrigramme pour ce vocabulaire standard est estimé en utilisant la boîte à outils SRILM (Stolcke, 2002) et la méthode de lissage Kneser-Ney sur des corpus de textes d'environ 1,8 milliard de mots. Ces corpus de textes sont issus d'articles de journaux (*Le Monde*), de transcriptions d'émissions de radio, des données collectées sur Internet et de l'intégralité du corpus diachronique.

4. Résultats expérimentaux

Les résultats de la sélection des nouveaux NP sont présentés en termes de rappel : le pourcentage des NP OOV retrouvés par rapport au nombre total des NP OOV des documents à transcrire. Nous nous plaçons dans le cadre de la reconnaissance automatique de la parole. Dans ce cadre, le fait qu'un NP présent dans le document à reconnaître ne soit pas dans le vocabulaire du système de reconnaissance va produire une erreur importante car ce NP ne pourra pas être reconnu. En revanche, ajouter au vocabulaire du système de reconnaissance un NP qui n'est pas présent (prononcé) dans le fichier de test, aura peu d'influence sur la phrase reconnue (si l'on ajoute trop de mots, il y a un risque d'augmenter la confusion entre les mots et donc de provoquer des erreurs). Le rappel est plus important que la précision dans notre cas, de fait nous ne présentons que le rappel.

Pour les expériences de reconnaissance, le taux d'erreur de mots (*Word Error Rate*, WER) et le taux d'erreur de NP (*Proper Name Error Rate*, PNER) sont calculés. Le PNER est calculé de la même façon que le WER mais en prenant en compte uniquement les noms propres.

Nous appelons « NP sélectionnés » les noms propres que nous avons récupérés à partir des *documents sélectionnés* en utilisant notre méthode et qui ne sont pas dans notre vocabulaire.

Nous appelons « NP OOV retrouvés » les noms propres OOV de la liste NP sélectionnés qui sont présents dans le document à transcrire.

Afin d'étudier si la période temporelle joue un rôle important, nous avons étudié trois intervalles de temps pour les *documents sélectionnés* : le même jour que le document à transcrire (noté « 1 jour » dans les tableaux) ; trois jours avant et trois jours après la date du document à transcrire (noté « 1 semaine » dans les

tableaux) ; le mois courant du document à transcrire (noté « 1 mois » dans les tableaux).

Pour confirmer notre hypothèse sur l'importance de choisir des *documents sélectionnés* correspondant à la même période temporelle que les documents à transcrire, nous avons mis en place un protocole expérimental contrastif : au lieu d'utiliser des *documents sélectionnés* de la même période que le document à transcrire, nous utilisons des *documents sélectionnés* dont la date est située dix mois après la date du document à transcrire. Notre but ici est de sélectionner des documents qui sont très éloignés temporellement des documents à transcrire. Nous avons choisi de prendre des documents dix mois après la date du document à transcrire pour se mettre dans des conditions plus difficiles : des documents rédigés dix mois avant la date du document à transcrire ont moins de chance de contenir des noms propres pertinents.

Nous construisons un vocabulaire augmenté spécifique à chaque document à transcrire, à chaque période choisie et à chaque méthode. Le vocabulaire augmenté contient tous les mots du vocabulaire standard et les NP sélectionnés par la méthode et la période choisies. Il nous faut donc estimer les probabilités n-grammes pour ces NP sélectionnées. Dans ce but, nous avons choisi de réestimer totalement le modèle de langage pour chaque vocabulaire augmenté en utilisant l'ensemble du corpus de texte (cf. section 3.3). Nous utilisons le même outil (SRILM) et la même méthode de lissage (Kneser-Ney). La meilleure façon d'intégrer les nouveaux noms propres dans le modèle de langage n'entre pas dans le champ d'application de cet article.

Les NP d'ancrage sont extraits à partir de la transcription automatique générée par notre système ANTS en utilisant le vocabulaire standard. Pour cela, ce vocabulaire a été étiqueté en termes de NP et de non-NP. Environ 76 % de NP du vocabulaire standard sont bien reconnus dans les transcriptions automatiques du corpus de développement. Cela donne une idée de la qualité de l'ancrage pour nos méthodes.

Dans tous les tableaux, le meilleur résultat est indiqué en caractères gras.

4.1. *Méthode de référence*

En utilisant l'outil d'étiquetage morphosyntaxique *TreeTagger*, nous avons extrait 160 k NP différents à partir d'une année du corpus diachronique. Parmi ces 160 k NP, 119 k ne sont pas dans notre vocabulaire standard (voir le tableau 4). Parmi ces 119 k, en moyenne seulement 24 NP OOV sont présents dans un fichier du corpus de développement. Si on voulait ajouter tous les NP du corpus diachronique qui ne sont pas dans le vocabulaire standard, il faudrait ajouter 667 k NP (à la fois dans le lexique et dans le modèle de langage). Ce système de reconnaissance serait probablement inutilisable car il nécessiterait une énorme quantité de mémoire vide et serait très lent. Cela montre qu'il est nécessaire de

filtrer la liste des NP pour avoir un meilleur compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire.

Période temporelle	**Méthode**	**Nombre moyen de NP sélectionnés par fichier de développement**	**Nombre moyen de NP OOV retrouvés par fichier de développement**	**Rappel (%)**
1 jour	Méthode de référence	532,9	9,9	32,1
	Méthode de référence, exp. contrastive	1 191,0	3,1	10,2
1 semaine	Méthode de référence	2 928,4	11,3	36,7
	Méthode de référence, exp. contrastive	5 459,0	7,7	25,1
1 mois	Méthode de référence	1 2452,0	17,6	57,1
	Méthode de référence, exp. contrastive	15 445,0	13,9	45,1
1 année	Méthode de référence	118 797,0	24,0	**78,1**

Tableau 4. *Résultats de la méthode de référence sur le corpus de développement*

Comme nous construisons un vocabulaire augmenté pour chaque fichier du corpus de développement à transcrire, les résultats présentés dans le tableau 4 sont moyennés sur les sept fichiers de développement.

Le tableau 4 montre que l'utilisation des *documents sélectionnés* dont la date est proche de celle du document à transcrire permet de réduire très fortement le nombre de nouveaux noms propres ajoutés tout en conservant un rappel intéressant. Par exemple, en passant d'une période temporelle d'un mois à un jour, on réduit le nombre de nouveaux NP ajoutés d'un facteur supérieur à 24 alors que le rappel n'est réduit que d'un facteur 1,7. Ce résultat confirme l'idée que l'utilisation de l'information temporelle réduit la liste des nouveaux candidats NP pour l'enrichissement du vocabulaire tout en conservant un bon taux de rappel. Les résultats des expériences contrastives confirment ce fait. Le rappel chute très significativement si on utilise des *documents sélectionnés* dont la date est éloignée de celle du document à transcrire.

Dans la suite de l'article, nous étudierons trois périodes temporelles (un jour, une semaine un mois) car la période d'une année ne semble pas être intéressante dans le cadre des bulletins d'information.

Afin de sélectionner des noms propres plus pertinents, nous avons introduit un seuil d'occurrence dans les méthodes présentées par la suite : seuls les noms propres dont le nombre d'occurrences dans les *documents sélectionnés* dépasse ce seuil sont retenus. Ce seuil dépend de la durée de la période considérée (noté « occ. » dans les tableaux).

4.2. *Méthode fondée sur des cooccurrences dans des fenêtres*

Le tableau 5 présente les résultats pour la méthode des cooccurrences dans des fenêtres sur le corpus de développement pour les trois périodes de temps considérées (à la suite d'expériences préliminaires, la taille de la fenêtre a été fixée à cent mots).

Par rapport à la méthode de référence, la méthode fondée sur des cooccurrences permet de réduire de façon notable le nombre de noms propres sélectionnés au prix d'une faible baisse du rappel. Par exemple, pour une période d'une journée, le nombre de mots sélectionnés est divisé par 2, tandis que le rappel ne baisse que de 1 %. Pour la période d'un mois, le filtrage permet de diviser le nombre de NP par 5, en perdant seulement 5 points de rappel. Cela montre l'efficacité du filtrage proposé.

Période temporelle	**Méthode**	**Nombre moyen de NP sélectionnés par fichier de développement**	**Nombre moyen de NP OOV retrouvés par fichier de développement**	**Rappel (%)**
1 jour (occ. > 0)	Fenêtres	245,3	9,6	31,2
	Fenêtres, exp. contrastive	383,7	2,1	7,0
1 semaine (occ. > 1)	Fenêtres	834,7	10,1	33,0
	Fenêtres, exp. contrastive	1 145,4	5,6	18,1
1 mois (occ. > 2)	Fenêtres	2 356,6	15,4	**50,2**
	Fenêtres, exp.	2 275,1	9,9	32,1

Tableau 5. *Résultats de la méthode des cooccurrences dans des fenêtres sur le corpus de développement. Taille de fenêtre : 100 mots.*

On peut noter que l'écart de performance en termes de rappel entre les périodes d'un jour et d'une semaine est faible car certains événements ne font l'objet d'articles que pendant une journée.

La très forte baisse du rappel pour les expériences contrastives (par exemple, pour un jour 7,0 % *versus* 31,2 %) confirme notre hypothèse concernant l'importance d'utiliser des *documents sélectionnés* correspondant à la date du document à transcrire.

4.3. *Méthode fondée sur l'information mutuelle*

Le tableau 6 montre, pour le corpus de développement, les résultats de la méthode fondée sur l'information mutuelle en utilisant différentes périodes de temps et différentes valeurs de seuils.

Période temporelle	***SeuilMI***	**Nombre moyen de NP sélectionnés par fichier de développement**	**Nombre moyen de NP OOV retrouvés par fichier de développement**	**Rappel (%)**
1 jour (occ. > 0)	0,01	244,7	9,7	31,6
	0,005	335,6	9,9	32,1
	0,001	438,4	10,0	32,6
1 semaine (occ. > 1)	0,01	261,4	8,1	26,5
	0,005	581,4	9,9	32,1
	0,001	1 196,9	10,6	34,4
1 mois (occ. > 2)	0,01	97,3	5,7	18,6
	0,005	197,4	8,6	27,9
	0,001	1 862,7	**14,1**	**46,1**

Tableau 6. *Résultats de la méthode MI sur le corpus de développement en fonction de différentes valeurs de seuils.*

Comme précédemment, utiliser uniquement les *documents sélectionnés* d'un seul jour s'avère être suffisant pour obtenir un rappel supérieur à 30 %.

Plus le *seuilMI* est petit, plus on a de mots sélectionnés et plus le rappel augmente. Pour les expériences de reconnaissance (cf. section 4.6) nous fixerons le *seuilMI* à 0,001 pour toutes les périodes temporelles.

Le tableau 7 montre une nette dégradation du rappel pour les expériences contrastives, comme pour la méthode fondée sur les fenêtres.

Période temporelle	Méthode	Nombre moyen de NP sélectionnés par fichier de développement	Nombre moyen de NP OOV retrouvés par fichier de développement	Rappel (%)
1 jour (occ. > 0)	MI	438,4	10,0	32,6
	MI, exp. contrastive	738,9	3,0	9,8
1 semaine (occ. > 1)	MI	1 196,9	10,6	34,4
	MI, exp. contrastive	1 623,4	6,1	20,0
1mois (occ. > 2)	MI	1 862,7	**14,1**	**46,1**
	MI, exp. contrastive	1 707,0	7,9	25,6

Tableau 7. *Résultats de la méthode MI sur le corpus de développement.* SeuilMI *de 0,001*.

4.4. *Méthode fondée sur le modèle vectoriel avec comme distance la similarité cosinus*

Les résultats pour la méthode fondée sur la similarité cosinus sont présentés dans le tableau 8. Le meilleur rappel est obtenu pour la période d'un mois et un *seuilCos* de 0,025 (47,4 % de rappel), mais ce résultat est remporté au prix d'un nombre important de noms propres sélectionnés (3 777,4 en moyenne par fichier de développement).

Pour les expériences de reconnaissance présentées section 4.6, nous avons fixé le *seuilCos* à 0,025 pour les périodes d'un jour et d'une semaine. Nous avons choisi un seuil plus élevé (0,05) pour un mois car il permet d'obtenir un meilleur compromis entre le nombre de noms propres sélectionnés et le rappel.

Les expériences contrastives présentées tableau 9 confirment les conclusions tirées des expériences contrastives précédentes.

Période temporelle	***SeuilCos***	**Nombre moyen de NP sélectionnés par fichier de développement**	**Nombre moyen de NP OOV retrouvés par fichier de développement**	**Rappel (%)**
1 jour (occ. > 0)	0,025	358,7	9,9	32,1
	0,05	208,1	9,7	31,6
	0,075	107,9	8,6	27,9
1 semaine (occ. > 1)	0,025	1 188,3	10,0	32,6
	0,05	746,0	9,7	31,6
	0,075	351,3	8,6	29,7
1 mois (occ. > 2)	0,025	3 777,4	**14,6**	**47,4**
	0,05	2 611,1	12,7	41,4
	0,075	1 209,1	11,1	36,3

Tableau 8. *Résultats pour la méthode fondée sur la similarité cosinus. Corpus de développement.*

Période temporelle	**Méthode**	**Nombre moyen de NP sélectionnés par fichier de développement**	**Nombre moyen de NP OOV retrouvés par fichier de développement**	**Rappel (%)**
1 jour (occ. > 0)	Cosinus	358,7	9,9	32,1
	Cosinus, exp. contrastive	1 017,3	3,3	10,7
1 semaine (occ. > 1)	Cosinus	1 188,3	10,0	32,6
	Cosinus, exp. contrastive	2 457,4	6,9	22,3
1 mois (occ. > 2)	Cosinus	2 611,1	**12,7**	**41,4**
	Cosinus, exp. contrastive	3 242,1	9,1	29,8

Tableau 9. *Résultats pour la méthode fondée sur la similarité cosinus. Corpus de développement.* SeuilCos *de 0,025 pour un jour et une semaine, 0,05 pour un mois.*

4.5. *Comparaison des trois méthodes proposées*

On peut noter que l'écart de performance en termes de rappel entre les périodes d'un jour et d'une semaine est faible. Cela peut s'expliquer par le fait que beaucoup d'événements sont ponctuels et ne font l'objet d'articles que pendant une seule journée. Cette remarque reste valable pour les trois méthodes proposées.

La figure 1 présente le taux de rappel (en %) en fonction du nombre moyen de mots sélectionnés par fichier de développement pour les trois méthodes étudiées. Pour cette courbe, nous avons choisi une période d'une semaine car c'est un bon compromis entre le nombre de NP sélectionnés et le rappel.

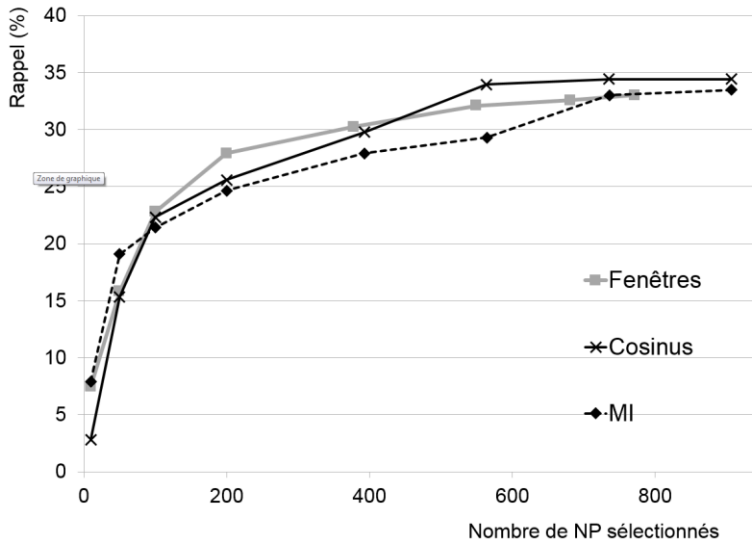

Figure 1. *Taux de rappel (%) en fonction du nombre moyen de NP sélectionnés par fichier de développement pour trois méthodes proposées (période temporelle d'une semaine)*

Nous remarquons un comportement assez similaire pour les trois méthodes : une très forte augmentation du rappel pour les 100 premiers NP sélectionnés ; la valeur maximale du rappel est obtenue à partir de 700 NP sélectionnés et ensuite la courbe est plate. La méthode MI semble être très légèrement moins performante que les deux autres méthodes.

4.6. *Résultats de reconnaissance sur le corpus de développement et de test*

Nous avons effectué la reconnaissance automatique des documents du corpus de développement et du corpus de test en utilisant les vocabulaires augmentés par les trois méthodes proposées. Nous générons un vocabulaire par fichier à transcrire, par période et par méthode. Pour la génération des prononciations des nouveaux noms propres, nous utilisons une approche automatique fondée sur des CRF. Nous avons choisi cette approche car elle a montré de très bons résultats par rapport à l'une des

meilleures approches sur l'état de l'art (Illina et al., 2011). Les CRF (Lafferty *et al.*, 2001) sont un modèle probabiliste pour l'étiquetage ou la segmentation de données structurées telles que des séquences, des arbres ou des treillis. Les CRF permettent de tenir compte de relations à long terme, de réaliser un apprentissage discriminant et de converger vers un optimal global. En utilisant cette approche, nous avons obtenu une précision et un rappel de plus de 98 % pour la phonétisation des noms communs du français (BDLex) (Illina *et al*., 2011). Dans le cadre de cet article, nous avons entraînés les CRF avec un corpus de 12 000 noms propres phonétisés.

Les résultats de reconnaissance sur les sept documents du corpus de développement sont présentés dans le tableau 10. En moyenne, le vocabulaire augmenté en utilisant les *documents sélectionnés* réduit légèrement le WER (intervalle de confiance ± 0,4 %). Nous rappelons que le taux de NP OOV dans le corpus de développement est d'environ 1,3 % et donc l'amélioration obtenue pourra difficilement dépasser ce taux. Les résultats détaillés montrent que la performance en termes de WER dépend du type de documents : pour certaines émissions de radio nous n'observons aucune amélioration (par exemple, un débat sur un thème récurrent comme le nucléaire), pour d'autres une forte amélioration est atteinte (bulletins d'information).

Les noms propres sont très souvent indispensables pour comprendre l'information contenue dans un document. En termes de PNER, une amélioration significative est obtenue pour toutes les durées et toutes les méthodes (intervalle de confiance ± 2,4 %). Les meilleurs résultats sont obtenus pour une durée d'un mois (35,0 % PNER comparé à 40,7 % pour méthode fondée sur les fenêtres), mais le nombre de NP sélectionnés est également le plus grand. Le grand nombre de noms propres ajoutés qui ne sont pas dans les documents à transcrire n'ont pas une influence négative sur le taux de reconnaissance.

Les résultats des expériences contrastives confirment notre hypothèse que la correspondance temporelle entre les documents à transcrire et les *documents sélectionnés* est importante.

En termes de PNER, l'écart entre l'expérience contrastive et l'expérience normale est maximal pour la période d'une journée.

Pour valider les approches proposées, nous avons effectué la reconnaissance automatique de treize documents de test en utilisant les paramètres qui ont été choisis sur le corpus de développement. Pour chaque vocabulaire augmenté, le modèle de langage est réestimé sur l'intégralité du corpus textuel (1,8 milliard de mots).

Comme le montre le tableau 11, par rapport au vocabulaire standard, les trois méthodes obtiennent une amélioration notable en termes de WER (intervalle de confiance ± 0,4 %). En revanche, nous n'observons aucune différence significative entre les trois méthodes.

Vocabulaire standard	Méthode	Vocabulaire augmenté		
		1 jour	1 semaine	1 mois
WER 30,2	WER fenêtres	29,8	29,9	**29,7**
	WER fenêtres, exp. contrastive	30,3	30,1	29,9
	WER MI	29,9	29,9	29,8
	WER MI exp. contrastive	30,3	30,0	30,1
	WER cosinus	29,8	30,0	**29,7**
	WER cosinus, exp. contrastive	30,1	30,0	29,9
PNER 40,7	PNER fenêtres	37,0	37,4	**35,0**
	PNER fenêtres, exp. contrastive	40,6	38,9	37,0
	PNER MI	36,8	37,0	35,6
	PNER MI, exp. contrastive	40,3	38,3	37,8
	PNER cosinus	36,8	36,9	35,4
	PNER cosinus, exp. contrastive	40,3	38,1	36,9

Tableau 10. *Résultats de reconnaissance en termes de WER et PNER pour les trois méthodes proposées. Corpus de développement.*

Les résultats, obtenus pour les expériences contrastives, sont, comme attendu, beaucoup moins bons que les expériences normales.

En termes de PNER, une amélioration significative de plus de 6 % (en absolu) est obtenue pour les trois méthodes et la période d'un mois (intervalle de confiance ± 2 %). Il est intéressant de noter qu'en utilisant uniquement les *documents sélectionnés* du même jour que les documents à transcrire, nous obtenons une amélioration relative de 13 % du taux PNER. En revanche, il n'y a pas de différences notables entre les trois durées évaluées. Ceci est peut-être dû au fait que certains événements font l'objet d'articles de presse pendant une très courte période de temps, parfois seulement un jour.

Vocabulaire standard	Méthode	Vocabulaire augmenté		
		1 jour	1 semaine	1 mois
WER 31,8	WER fenêtres	31,3	**31,2**	31,4
	WER fenêtres, exp. contrastive	31,9	31,7	31,5
	WER MI	31,3	**31,2**	**31,2**
	WER MI, exp. contrastive	31,9	31,6	31,5
	WER cosinus	31,3	**31,2**	**31,2**
	WER cosinus, exp. contrastive	31,8	31,7	31,6
PNER 44,0	PNER fenêtres	38,3	37,7	37,7
	PNER fenêtres, exp. contrastive	44,4	41,7	40,9
	PNER MI	38,0	37,6	37,6
	PNER MI, exp. contrastive	44,1	41,8	41,0
	PNER cosinus	38,1	37,6	**37,4**
	PNER cosinus, exp. contrastive	43,8	41,3	40,2

Tableau 11. *Résultats de reconnaissance en termes de WER et PNER pour les trois méthodes proposées. Corpus de test.*

5. Conclusion

Dans le cadre de la reconnaissance automatique de la parole, notre étude a porté sur le problème de l'augmentation du vocabulaire à l'aide de *documents sélectionnés*. Nous avons proposé des méthodes qui augmentent le vocabulaire avec des noms propres en utilisant la notion de contexte lexical et temporel. Ces nouveaux NP sont ajoutés au vocabulaire du système. Des méthodes de filtrage fondées sur des cooccurrences, sur l'information mutuelle et sur le modèle vectoriel ont été définies.

Des expériences ont été menées sur des émissions de radio en utilisant des données textuelles d'agences de presse comme corpus diachronique. Les résultats valident l'hypothèse que le contexte temporel et le contexte lexical permettent de récupérer des noms propres manquants et d'obtenir une réduction significative du taux d'erreur de noms propres en ajoutant ces NP dans le vocabulaire.

Les expériences contrastives ont permis de montrer l'importance du contexte temporel. En revanche, il est délicat de conclure concernant le choix de la durée de

la période temporelle des *documents sélectionnés* à retenir. Nous pouvons noter que des gains substantiels sont atteints dès une journée.

Les trois méthodes de filtrage proposées donnent des résultats comparables en termes de WER et PNER. Une perspective intéressante pourrait être d'étudier l'intersection des ensembles de noms propres sélectionnés par ces trois méthodes et éventuellement de les combiner. Une autre perspective pourrait être aussi d'enlever du vocabulaire les mots « dépassés » temporellement (Ohtsuki *et al*., 2005a).

Nous pourrons envisager également d'exploiter des informations sémantiques contenues dans le document à transcrire : quand une date précise est identifiée, les *documents sélectionnés* temporellement proches de cette date pourraient être utilisés pour extraire de nouveaux noms propres.

Remerciements

Les auteurs tiennent à remercier l'ANR *ContNomina* SIMI-2 de l'Agence nationale de la recherche (ANR) pour son soutien.

6. Bibliographie

Allauzen A., Gauvain J.-L. « Diachronic vocabulary adaptation for broadcast news transcription », *Proc. of Interspeech*, 2005.

Allauzen A., Gauvain. J.-L. « Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés » *dans Traitement automatique des langues*, 44(1) : p. 11-31, 2003.

Bechet F., Yvon F. « Les noms propres en traitement automatique de la parole », *Revue Traitement automatique des langues*, vol. 41, n° 3, p. 672-708, 2000.

Bellegarda J. « Statistical language model adaptation: review and perspectives », *Speech Communication* Volume 42, Issue 1, p. 93-108, 2004.

Bertoldi N., Federico M. « Lexicon adaptation for broadcast news transcription », *In Adaptation-2001*, p. 187- 190, 2001.

Bigot B., Senay G., Linares G., Fredouille C., Dufour R. « Person name recognition in ASR outputs using continuous context models », *Proc. of ICASSP*, 2013.

Bigot B., Senay G., Linares G., Fredouille C., Dufour R . « Combining Acoustic Name Spotting and Continuous Context Models to improve Spoken Person Name Recognition in Speech », *Proc. of International conference of the Speech Communication Association, ISCA*, InterSpeech'13, 2013a.

Bisani, M., Ney, H., « Joint-Sequence Models for Grapheme-to-Phoneme Conversion », *Speech Communication*, 50 : p. 434-451, Elsevier, 2008.

Chen L., Gauvain J.L., Lamel L., Adda G. « Unsupervised language model adaptation for broadcast news », *in Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing – ICASSP*, p. 220-223, 2003.

Church, K., Hanks, P. « Word association norms, mutual information, and lexicography », *Proc. of the 27th Annual Meeting of the Association for Computational Linguistics*, 1989.

De Calmès M., Pérennou G. « BDLEX : a Lexicon for Spoken and Written French », *Proc. of 1st International Conference on Language Resources & Evaluation (LREC)*, Grenade, 1998.

Federico M., Bertoldi N. « Broadcast news LM adaptation over time », *Computer Speech and Language*, 18(4), 417-435, 2004.

Friburger N., Maurel D. « Textual Similarity Based on Proper Names », *Proc. of the workshop Mathematical/Formal Methods in Information Retrieval*, 2002, p. 155-167.

Galliano S., Gravier G., Chaubard L. « The ESTER2 Evaluation Campaign for the Rich Transcription of French Radio Broadcast », *Proc. of Interspeech*, 2009.

Illina I., Fohr D., Mella O., Cerisara C. « The Automatic News Transcription System: ANTS, some Real Time experiments », *Proc. ICSLP*, 2004.

Illina I., Fohr D., Linarès G. « Extension du vocabulaire d'un système de transcription avec de nouveaux noms propres en utilisant un corpus diachronique », *Proc. of JEP*, 2014.

Illina I., Fohr D., Jouvet D. « Grapheme-to-Phoneme Conversion using Conditional Random Fields », *Proc. of Interspeech*, 2011.

Jensen, K.J., Riis, S., « Self-Organizing Letter Code-Book for Text-to-Phoneme Neural Network Model », *Proc. of International Conference on Spoken Language Processing*, 3, p. 318-321, 2000.

Jouvet D., Fohr D., Illina I. « Evaluating grapheme-to-phoneme converters in automatic speech recognition context », *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

Kemp, T., Waibel, A. « Reducing the OOV Rate in Broadcast News Speech Recognition », *in Proc. of ICSLP*, p. 1839-1842, 1998.

Kobayashi A., Onoe K., Imai T., Ando A. « Time dependent language model for broadcast news transcription and its post-correction », *Proc ICSLP*, 1998.

Lafferty, J., McCallum, A., Pereira, F. « Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data », *Proc. of International Conference on Machine Learning*, p. 282-289, 2001.

Lee A., Kawahara T. « Recent Development of Open-Source Speech Recognition Engine Julius », *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2009.

Manning C., Raghavan P., Schütze H. « Introduction to Information Retrieval », *Cambridge University Press*, 2008.

Martins C., Teixeira A., Neto J. « Dynamic language modeling for European Portuguese ». *Computer Speech and Language*, 24(4), p. 750-773, 2010.

Nkairi I., Illina I., Linarès G., Fohr D. « Exploring temporal context in diachronic text documents for automatic OOV proper name retrieval », *Proc. of LTC*, 2013.

Oger S., Linarès G., Béchet F. « Local methods for on-demand out-of-vocabulary word retrieval », *Proc. of the Language Resources and Evaluation Conference (LREC)*, 2008.

Ohtsuki K, Nguyen L. « Incremental Language Modeling for Broadcast News », *Proc. of ASRU*, 2005a.

Ohtsuki K., Hiroshima N., Oku M., Imamura A. « Unsupervised vocabulary expansion for automatic transcription of broadcast news » *Proc of International conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1021-1024, 2005b.

Pagel, V., Lenzo, K., Black, A.W., « Letter-to-Sound Rules for Accented Lexicon Compression », *Proc.of International Conference on Spoken Language Processing*, 5, p. 2015-2018, 1998.

Parada C., Dredze M., Filimonov F., Jelinek F. « Contextual Information Improves OOV Detection in Speech », *Proc. of NAACL*, 2010.

Salton G., Wong A., Yang C. S. « A vector space model for automatic indexing », *Communications of the ACM*, v.18 n.11, p. 613-620, Nov. 1975.

Schmid H. « Probabilistic part-of-speech tagging using decision trees », *Proc. of ICNMLP*, 1994.

Singhal A. « Modern Information Retrieval: A Brief Overview », *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): p. 35-43, 2001.

Stolcke A. « SRILM - An Extensible Language Modeling Toolkit », *Proc. of ICSLP*, 2002.

Taylor, P. « Hidden Markov Models for Grapheme to Phoneme conversion », *Proc. of Interspeech*, p. 1973-1976, 2005.

Yu H., Tomokiyo T., Wang Z., Waibel, A. « New Developments in Automatic Meeting Transcription », *Proc. of ICSLP*, 2000.