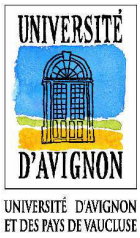

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 536 «Sciences et Agrosciences »
Laboratoire d'Informatique (EA 4128)

Modèles de langage ad hoc pour la reconnaissance automatique de la parole

par
Stanislas OGER

Soutenue publiquement le 30 novembre 2011 devant un jury composé de :

M.	Kamel Smaïli	Professeur, LORIA, Nancy	Rapporteur
M.	Laurent Besacier	Professeur, LIG, Grenoble	Rapporteur
M.	Yannick Estève	Professeur, LIUM, Le Mans	Examinateur
M.	Alexandre Allauzen	Maître de Conférence, LIMSI, Paris	Examinateur
M.	Pascal Nocéra	Maître de Conférence, LIA, Avignon	Examinateur
M.	Georges Linarès	Professeur, LIA, Avignon	Directeur de thèse

Laboratoire d'Informatique d'Avignon

Remerciements

Je remercie Kamel Smaïli et Laurent Besacier d'avoir accepté d'être rapporteurs de ma thèse. Je remercie également Yannick Estève, Alexandre Allauzen et Pascal Nocéra de participer au jury en tant qu'examinateurs.

J'ai effectué ma thèse au sein du Laboratoire Informatique d'Avignon et cette expérience a été très enrichissante. Je tiens à remercier tout particulièrement mon directeur de thèse, Georges Linarès, pour son soutien et son engagement dans mes travaux, tout en me laissant une très grande liberté de recherche. Je remercie également Benoît Favre, Frederic Béchet et Pascal Nocéra pour leur aide précieuse lors du démarrage de ma thèse.

Ces années au Laboratoire Informatique d'Avignon ont été très enrichissantes sur le plan humain, aussi je tiens à remercier les personnes que j'y ai croisé pour leur contribution (dans le désordre) : Nicolas F., Alain, Benjamin L., Benjamin M., Claire, Christian, Pierre G., Remy, Driss, Raphaël, Hugo, Vladimir, Lauriane, Christophe S., Fabrice, Christophe L., Nimaan, Gilles, Pierre C., Thierry V., Philou, Marie-jean, Juliette, Mickaël, Marc, Jean-Pierre, Anthony, Frederic D., Gregory, Corinne, Nathalie, Loïc, Florian P., Florian V., Thierry P., Mohamed.

Résumé

Les trois piliers d'un système de reconnaissance automatique de la parole sont le lexique, le modèle de langage et le modèle acoustique. Le lexique fournit l'ensemble des mots qu'il est possible de transcrire, associés à leur prononciation. Le modèle acoustique donne une indication sur la manière dont sont réalisés les unités acoustiques et le modèle de langage apporte la connaissance de la manière dont les mots s'enchaînent. Dans les systèmes de reconnaissance automatique de la parole markoviens, les modèles acoustiques et linguistiques sont de nature statistique. Leur estimation nécessite de gros volumes de données sélectionnées, normalisées et annotées.

A l'heure actuelle, les données disponibles sur le Web constituent de loin le plus gros corpus textuel disponible pour les langues française et anglaise. Ces données peuvent potentiellement servir à la construction du lexique et à l'estimation et l'adaptation du modèle de langage. Le travail présenté ici consiste à proposer de nouvelles approches permettant de tirer parti de cette ressource.

Ce document est organisé en deux parties. La première traite de l'utilisation des données présentes sur le Web pour mettre à jour dynamiquement le lexique du moteur de reconnaissance automatique de la parole. L'approche proposée consiste à augmenter dynamiquement et localement le lexique du moteur de reconnaissance automatique de la parole lorsque des mots inconnus apparaissent dans le flux de parole. Les nouveaux mots sont extraits du Web grâce à la formulation automatique de requêtes soumises à un moteur de recherche. La phonétisation de ces mots est obtenue grâce à un phonétiseur automatique.

La seconde partie présente une nouvelle manière de considérer l'information que représente le Web et des éléments de la théorie des possibilités sont utilisés pour la modéliser. Un modèle de langage possibiliste est alors proposé. Il fournit une estimation de la possibilité d'une séquence de mots à partir de connaissances relatives à l'existence de séquences de mots sur le Web. Un modèle probabiliste Web reposant sur le compte de documents fourni par un moteur de recherche Web est également présenté. Plusieurs approches permettant de combiner ces modèles avec des modèles probabilistes classiques estimés sur corpus sont proposées. Les résultats montrent que combiner les modèles probabilistes et possibilistes donne de meilleurs résultats que les modèles probabilistes classiques. De plus, les modèles estimés à partir des données Web donnent de meilleurs résultats que ceux estimés sur corpus.

Abstract

The three pillars of an automatic speech recognition system are the lexicon, the language model and the acoustic model. The lexicon provides all the words that can be transcribed, associated with their pronunciation. The acoustic model provides an indication of how the phone units are pronounced, and the language model brings the knowledge of how words are linked. In modern automatic speech recognition systems, the acoustic and language models are statistical. Their estimation requires large volumes of data selected, standardized and annotated.

At present, the Web is by far the largest textual corpus available for English and French languages. The data it holds can potentially be used to build the vocabulary and the estimation and adaptation of language model. The work presented here is to propose new approaches to take advantage of this resource in the context of language modeling.

The document is organized into two parts. The first deals with the use of the Web data to dynamically update the lexicon of the automatic speech recognition system. The proposed approach consists on increasing dynamically and locally the lexicon only when unknown words appear in the speech. New words are extracted from the Web through the formulation of queries submitted to Web search engines. The phonetization of the words is obtained by an automatic grapheme-to-phoneme transcriber.

The second part of the document presents a new way of handling the information contained on the Web by relying on possibility theory concepts. A Web-based possibilistic language model is proposed. It provides an estition of the possibility of a word sequence from knowledge of the existence of its sub-sequences on the Web. A probabilistic Web-based language model is also proposed. It relies on Web document counts to estimate n-gram probabilities. Several approaches for combining these models with classical models are proposed. The results show that combining probabilistic and possibilistic models gives better results than classical probabilistic models alone. In addition, the models estimated from Web data perform better than those estimated on corpus.

Table des matières

Introduction

Avec l'apparition des systèmes de reconnaissance automatique de la parole s'appuyant largement sur des modèles statistiques, est né un besoin de corpus d'entraînement sans cesse plus volumineux. Pour estimer correctement ce type de modèles, il est nécessaire de disposer d'une base d'exemples aussi conséquente que leur complexité l'exige.

En plus de la nécessité de disposer de gros volumes de données, les modèles estimés *a priori* doivent être adaptés dès qu'intervient une modification du domaine des documents à transcrire. En d'autres termes, il n'existe pas de modèle universel.

La collecte, la normalisation et éventuellement l'étiquetage des corpus aussi bien pour l'estimation que pour l'adaptation des modèles sont des tâches longues et très coûteuses. Pour cette raison, beaucoup d'efforts ont été fournis pour trouver des algorithmes moins gourmands en corpus et qui s'adaptent automatiquement au contenu des documents à transcrire.

Un système de reconnaissance automatique de la parole s'appuie généralement sur trois modèles : le modèle de langage, le modèle acoustique et le lexique. L'estimation et l'adaptation du modèle acoustique nécessitent des enregistrements audio pour lesquels une transcription est disponible. L'estimation du modèle de langage nécessite généralement un gros volume de données textuelles. Le lexique est habituellement constitué des mots les plus fréquents des corpus ayant servi à l'estimation du modèle de langage. La phonétisation des mots est renseignée manuellement ou de manière automatique, au prix d'éventuelles erreurs.

Les transcriptions produites par un système de reconnaissance automatique de la parole sont composées uniquement de mots du lexique. Si un mot n'est pas dans le lexique, il ne pourra pas être transcrit correctement. Ces mots sont appelés des mots hors-vocabulaires. La taille du lexique d'un système de reconnaissance automatique de la parole est limitée en raison de la complexité qu'elle induit lors du décodage mais également car la confusion introduite provoque une diminution des performances, comme le montre Rosenfeld (1995). Les enregistrements de parole dont le lexique n'est pas contraint pourront donc contenir des mots hors-vocabulaires.

Lorsque la tâche de transcription concerne des documents dont le contenu évolue au cours du temps, comme par exemple la transcription de journaux d'information, les mots hors-vocabulaires posent un réel problème. Le lexique du système de reconnais-

sance automatique de la parole nécessite alors d'être adapté régulièrement pour suivre l'évolution de l'actualité.

Lorsque le système de reconnaissance automatique de la parole est utilisé pour obtenir le contenu linguistique de vidéos ou d'enregistrements afin de les indexer, chaque mot hors-vocabulaire qui s'y trouve représente un risque de manquer le document lors d'une recherche faisant intervenir ce terme. Il faut dans ce cas adapter le lexique du système de reconnaissance automatique de la parole au contenu des documents à transcrire pour minimiser le taux de mots hors-vocabulaires.

Pour adapter le lexique, il faut disposer d'une source d'information contenant les mots manquants. Il peut s'agir d'un corpus suffisamment volumineux pour couvrir le contenu lexical des documents à transcrire ou un corpus plus réduit mais dont on est certain qu'il corresponde au contenu des documents.

Les données du Web constituent un corpus très intéressant comme source de mots pour l'adaptation lexicale car il possède deux caractéristiques inédites. La première est qu'il est de loin le plus gros corpus textuel disponible et regroupant des données très diversifiées. La seconde est qu'il est en perpétuelle évolution, car de nouvelles données y sont ajoutées en permanence. Il constitue alors un corpus très intéressant pour l'adaptation du lexique de reconnaissance automatique de la parole puisqu'il représente la plus vaste source de mots disponible et qui est, de plus, mise à jour de manière continue et automatique.

Dans la première partie de cette thèse, nous proposons une approche pour adapter localement et dynamiquement le lexique du modèle de langage à l'aide d'informations extraites du Web. Le contenu des documents à transcrire est extrait à l'aide d'une première passe de reconnaissance automatique de la parole. Des descripteurs sont ensuite extraits du contexte d'apparition des mots HV et permettent de les caractériser. Ils servent à élaborer des requêtes soumises à un moteur de recherche Web. Le lexique est alors adapté avec des nouveaux mots présents dans les documents retournés.

Parallèlement au lexique, les modèles de langages nécessitent de gros volumes de données textuelles pour être estimés correctement. Comme le montrent Banko et Brill (2001), dans certaines situations il est plus judicieux d'augmenter la taille du corpus d'entraînement plutôt que d'optimiser les paramètres des algorithmes d'apprentissage ou de mieux nettoyer les données. Ils montrent sur une tâche de désambiguïsation de mots que les performances du système augmentent avec la quantité de données d'entraînement suivant une loi log-linéaire, même en utilisant des algorithmes d'apprentissage très simples.

Là encore, on peut tirer parti des caractéristiques du Web. Il constitue une source de données textuelle dont la taille est sans commune mesure avec les corpus utilisés traditionnellement dans le domaine du traitement automatique de la langue naturelle. Son utilisation permettrait d'augmenter considérablement la taille des corpus d'entraînement utilisés dans les tâches de traitement automatique de la langue naturelle. On pourrait également s'appuyer sur son caractère évolutif pour proposer des modèles auto-adaptatifs dans des domaines liés à l'actualité.

L'utilisation des données du Web comme corpus d'entraînement a suscité beaucoup d'intérêts depuis quelques années. On a vu se développer des ateliers *Web as Corpus* en marge de nombreuses conférences, comme par exemple lors de *Corpus Linguistics Conference* 2005 ou encore *Language Resources and Evaluation Conference* 2008. Des revues reconnues sortent régulièrement des numéros spéciaux concernant l'utilisation du Web dans le domaine du traitement automatique de la langue naturelle, comme par exemple *Computational Linguistics* en 2003.

Les données accessibles depuis le Web ont été utilisées dans de nombreux domaines du traitement automatique de la langue naturelle. Par exemple dans le domaine de la traduction automatique on peut citer les travaux de Grefenstette (1999), de Resnik (1999) ou encore de Cao et Li (2002). Les travaux de Dumais et al. (2002) ou Soricut et Brill (2006) ont exploré l'utilisation du Web pour des problème de réponse automatique à des questions formulées en langage naturel. Rigau et al. (2002) ou Zahariev (2004) ont également utilisé le Web sur une tâche de désambiguïsation de mots. Dans presque tous les domaines du traitement automatique de la langue naturelle, cette ressource a été étudiée.

En ce qui concerne l'utilisation du Web pour la modélisation du langage dans le domaine de la reconnaissance automatique de la parole, la majorité des travaux ont consisté à récupérer des documents textuels sur le Web et à utiliser les techniques classiques d'estimation et d'adaptation pour en tirer parti.

Dans la seconde partie de cette thèse, nous proposons une nouvelle manière de considérer les informations disponibles sur le Web. Plutôt que de le voir comme une collection de documents textuels, nous l'utilisons comme un corpus ouvert dont l'accès au contenu se fait à l'aide de moteurs de recherche. Les documents ne sont pas utilisés directement, les scores linguistiques sont dérivés des statistiques des moteurs de recherche. Nous développons dans ce cadre un modèle de langage Web utilisant des éléments de la théorie des possibilistes pour mieux tirer parti de cette information. Nous proposerons différentes stratégies pour combiner ce modèle avec des modèles plus conventionnels.

Cette thèse est donc axée sur la définition d'un modèle de langage Web *ad-hoc* qui tire parti de l'évolutivité du Web en le considérant comme un corpus ouvert. L'accès aux données se fait *via* un moteur de recherche. Dans une première partie, le lexique est dynamiquement enrichi de nouveaux mots qui sont extraits du Web à l'aide de requêtes soumises à un moteur de recherche. Dans la seconde partie, un modèle de langage est estimé dynamiquement à partir des statistiques du moteur de recherche Web.

Ces travaux ont été réalisés dans le cadre de deux projets ANR. Le premier est le projet SIGMUND[1], soutenu par le Réseau National de Recherche et d'Innovation en Audiovisuel et Multimédia (RIAM) qui s'est déroulé de 2006 à 2009. Le but du projet est de développer des techniques de navigation et de surveillance de flux audio.

Le second est le projet est AVISON[2], soutenu également par le RIAM, qui

1. Projet numéro ANR-05-RIAM-0903
2. Projet numéro ANR-07-RIAM-0903

vise à développer des outils de navigation dans le fond documentaire de l'Institut de Recherche contre les Cancer de l'Appareil Digestif et le European Institute of TeleSurgery (IRCAD-EITS). Ce fond est composé de films d'opérations, d'avis d'experts et de cours dans le domaine de la chirurgie robotisée. L'objectif du projet AVISON est de proposer une plateforme pour l'indexation de la base audiovisuelle ainsi que des solutions pour son exploitation pédagogique dans un contexte multilingue.

Ce document est organisé en quatre parties. La première présente les systèmes de reconnaissance automatique de la parole Markoviens et plus particulièrement les modèles de langages qui y sont utilisés. La seconde partie est une étude des caractéristiques du Web du point de vue de son utilisation pour la modélisation du langage. Dans la troisième partie, est décrite l'approche que nous proposons pour adapter automatiquement et dynamiquement le lexique du moteur de reconnaissance automatique de la parole à partir des données du Web. Enfin, la dernière partie présente une manière originale de considérer les informations disponibles sur le Web dans le cadre de la reconnaissance automatique de la parole. Un modèle de langage Web s'appuyant sur des éléments de la théorie des possibilités y est proposé. Quelques conclusions et perspectives terminent le document.

Première partie

Principes des systèmes de reconnaissance automatique de la parole Markoviens

Chapitre 1

Les systèmes de reconnaissance automatique de la parole

Sommaire

Dans ce chapitre nous présentons les systèmes de reconnaissance automatique de la parole statistiques Markoviens, qui sont à la base des travaux présentés dans cette thèse. Le fonctionnement de tels systèmes sera présenté dans son ensemble. La partie concernant la modélisation du langage est présentée brièvement et sera développée dans le chapitre suivant.

1.1 Introduction

Un système de reconnaissance automatique de la parole a pour objectif de transcrire un message oral en un message orthographique de même contenu sémantique. L'idée de base est qu'il doit produire une transcription la plus parfaite possible du contenu linguistique du message audio. Ce type de système est d'un grand intérêt pour de nombreuses applications qui vont de la commande vocale au dialogue homme-machine.

1.1.1 Applications

Les systèmes de reconnaissance automatique de la parole sont souvent utilisés pour réaliser l'indexation de bases de données de vidéos ou d'émissions radio. En effet, grâce à de tels systèmes, les documents sont transcrit sous forme de texte, ce qui permet une représentation informatique plus aisée. Il est ainsi possible d'y effectuer des recherches en utilisant les techniques développées pour l'indexation de documents textes.

Ils sont également de plus en plus utilisés pour faire communiquer l'homme et la machine. C'est ce que l'on appelle l'interface homme-machine. La parole est certainement le moyen de communication le plus intuitif pour l'être humain. A contrario, il est plus aisé pour une machine de traiter des informations et des commandes sous la forme textuelle. Ainsi un système de reconnaissance automatique de la parole peut faire l'intermédiaire entre l'homme et la machine, permettant ainsi aux deux de communiquer de la manière la plus appropriée.

Les systèmes de reconnaissance automatique de la parole sont également utilisés dans le domaine du handicap. Par exemple, pour écrire, il faut avoir le plein usage de ses mains. Lorsque l'on ne dispose pas de ces facultés, mais que l'on est doué de parole, il est possible d'utiliser un système de reconnaissance automatique de la parole pour retranscrire ce que l'on dit.

La précédente application est également utilisée pour effectuer du sous-titrage. En effet, il est fréquent que la transcription de discours soit effectuée par un opérateur, appelé perroquet, qui répète le discours à un système de reconnaissance automatique de la parole qui se charge de la transcription écrite. Ce procédé est beaucoup plus rapide que de la saisie manuelle.

Il existe beaucoup d'autres application des systèmes de reconnaissance automatique de la parole, mais nous n'allons pas toutes les citer ici.

1.1.2 Historique

Avant toute chose, commençons par définir ce qu'est un système de reconnaissance automatique de la parole. Il existe plusieurs types de systèmes de reconnaissance automatique de la parole, qui se sont développés au fur et à mesure des avancées théoriques et technologiques dans le domaine de l'informatique.

Les premiers systèmes permettaient de reconnaître des mots isolés. C'est à dire que chaque mot devait être prononcé séparément pour que le système puisse les reconnaître. Ces systèmes ont été les premiers à être étudiés car ils reposent sur des techniques nécessitant peu de ressources de calcul. Une des manière les plus simples pour résoudre ce problème est de procéder par comparaison entre le signal observé et une réalisation de référence de chaque mot. La réalisation la plus proche du signal correspond alors au mot prononcé (Myers, 1980).

Par la suite est apparu la reconnaissance automatique de la parole continue. Contrairement aux systèmes précédents, les mots présents dans le message audio ne sont pas séparés. C'est en fait la manière dont les êtres humains s'expriment naturellement. En effet, en analysant le signal acoustique produit par un être humain énonçant une phrase, on s'apperçoit qu'il n'y a pas de pause entre les mots. La production de la parole chez l'humain consiste en un flux ininterrompu de sons, quelque soit les mots sous-jacents. Cette tâche constitue un réel défi car il est nécessaire d'analyser le message audio dans son ensemble.

Les premières études concernant la parole continue traitaient de la parole lue. Ce type de parole possède plusieurs caractéristiques qui font qu'elle est plus aisée à transcrire. Tout d'abord les conditions sont contrôlées. En effet, l'enregistrement est en général effectué dans un studio avec du matériel d'enregistrement adapté. De plus, la parole lue est plutôt stable, il n'y a pas de perturbation du signal dues par exemple à l'émotion, à des hésitations ou a des modifications intempestives de la position du locuteur par rapport au micro. Finalement le texte lu est en général issu d'ouvrages textuels disponibles. De ce fait, ils sont beaucoup plus rigoureux concernant la grammaire que ne l'est la parole spontanée. La structure du message est également différente. Les phrases spontanées sont en général plus courtes que les phrases écrites (Biber, 1988).

Par la suite, les systèmes et les algorithmes qui ont été développés pour la parole lue ont commencé à être adaptés à la parole spontanée, comme le décrivent par exemple

Gauvain et al. (2002). Ce changement est en parti dû au fait que la puissance des ordinateurs avait augmenté, permettant de développer de nouvelles approches plus gourmandes en ressources. De plus, l'apparition de campagnes d'évaluation sur ce type de données a offert à la communauté scientifique une référence commune pour comparer les systèmes et donc pouvoir plus facilement évaluer l'apport des travaux de chacun. On peut notamment citer la campagne HUB4 en 1996 (Stern, 1997) pour l'anglais, financée par l'*Defense Advanced Research Projects Agency* (DARPA), par le biais du *National Institute of Standards and Technology* (NIST) ou les campagnes RT qui ont suivi [1]. Pour le français on peut citer la campagne ESTER (Gravier et al., 2004), financée par le Ministère de l'Enseignement Supérieur et de la Recherche français et la Direction Générale de l'Armement par l'intermédiaire du projet Technolangue, qui a eu lieu en 2005, puis ESTER-2 qui s'est déroulée en 2008.

La parole spontanée est la parole que l'on rencontre le plus souvent. Il s'agit de messages non préparés, énoncés par les locuteurs de manière naturelle. Les données de prédilection pour cette tâche sont les enregistrements de conversations et d'interviews. Ces données possèdent plusieurs caractéristiques qui les rendent plus difficiles à transcrire. Tout d'abord la parole est souvent bruitée. En effet, le matériel d'acquisition peut être différent d'un locuteur à l'autre, il arrive qu'il y ait un fond musical à la parole, les mêmes locuteurs peuvent être enregistrés dans différentes conditions, il peut y avoir plusieurs locuteurs qui parlent en même temps, ils peuvent hésiter ou bredouiller, etc. De plus, différents styles de parole peuvent se côtoyer : de la parole préparée, presque lue, à la parole très spontanée. Le style est également très différent du style employé dans des documents écrits. Ce dernier point pose problème pour la modélisation du langage qui repose souvent sur l'analyse de grandes quantités de texte qui est en grande parti non issus de parole (Adda-Decker, 2006).

Depuis plusieurs années, on voit la recherche s'orienter vers une tâche encore plus ardue : la transcription de réunions. Comme pour la précédente évolution, l'apparition de campagnes d'évaluation motive les équipes du monde entier à effectuer des travaux dans cette voie. On peut citer par exemple la campagne NIST RT09 [2], pour l'anglais.

1.1.3 Formulation théorique du problème

Le développement rapide qu'a connu la reconnaissance automatique de la parole continue est, pour beaucoup, dû à la formulation statistique du problème que nous a fourni Jelinek (1976). Cette formulation est issue de la théorie de l'information proposée par Shannon et Weaver (1959).

A partir d'une observation acoustique X, la reconnaissance automatique de la parole consiste à trouver la séquence de mots $\hat{W}$ la plus probable parmi l'ensemble des séquences de mots possibles. Chaque mot constituant W proviennent d'un ensemble fixe et fini de mots : le vocabulaire V. Ce qui peut être formulé ainsi :

1. http://www.itl.nist.gov/iad/mig/tests/rt/2002/index.html
2. http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html

$$\hat{W} = \arg\max_{W} P(W|X) \tag{1.1}$$

où $P(W|X)$ est la probabilité de la séquence de mots W étant donné l'observation acoustique X.

Comme il n'est pas possible en pratique d'estimer $P(W|X)$, on peut utiliser le théorème de Bayes pour transformer la formule 1.1 en :

$$\hat{W} = \arg\max_{W} \frac{P(X|W) \times P(W)}{P(X)} \tag{1.2}$$

avec :

- $P(X|W)$: la probabilité d'observer le signal X lorsque la séquence de mots W est prononcée. Cette probabilité est donnée par le modèle acoustique.
- $P(W)$: la probabilité *a priori* de la séquence de mots W. Cette probabilité est fournie par le modèle de langage.
- $P(X)$: la probabilité *a priori* du signal X.

Cette formulation se justifie en partie grâce à la modélisation *source/canal* du processus de production de la parole, comme on peut le voir sur la figure 1.1 :

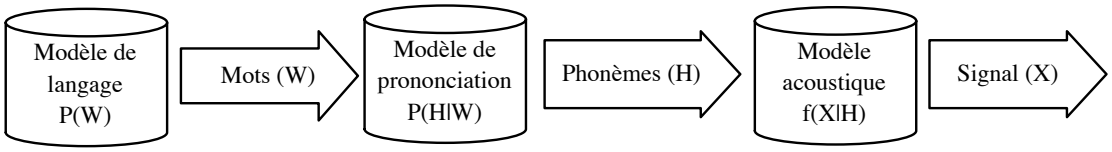

FIGURE 1.1 – *Représentation schématique de la modélisation source/canal du processus de production de la parole.*

La suite de mots W est tout d'abord générée par un processus linguistique, c'est le choix du message linguistique par le locuteur. Ce processus est modélisé par un modèle de langage, ici représenté par la mesure de la probabilité d'une séquence de mots w : $P(W)$. Un processus de prononciation transforme le message W en une séquence de phonèmes H. C'est le modèle de prononciation qui le modélise à l'aide de la mesure de probabilité $P(H|W)$, la probabilité de la séquence de phonèmes H étant donné la séquence de mots W. Finalement, ces phonèmes sont encodés par le canal acoustique $f(X|H)$ et produit le signal acoustique X.

On remarque qu'avec cette modélisation, la probabilité $P(X|W)$ doit être décomposée en un produit de probabilités : $P(H|W)$ et $P(X|H)$, ce qui donne la formule :

$$\hat{W} = \arg\max_{W} \frac{P(X|H) \times P(H|W) \times P(W)}{P(X)} \tag{1.3}$$

Cette dernière formulation est utilisée par quelques systèmes de reconnaissance automatique de la parole, mais généralement la probabilité de prononciation d'un mot n'est pas prise en compte.

On remarque que le terme $P(X)$ n'influence pas la décision puisqu'il ne dépend pas de W et est donc constant pour une recherche de l'arg max donné. On l'élimine donc de la formule pour obtenir :

$$\hat{W} = \arg\max_{W} P(X|H) \times P(H|W) \times P(W) \tag{1.4}$$

On remarque que pour qu'un mot puisse être reconnu par le processus de reconnaissance automatique de la parole, c'est à dire qu'il fasse parti de l'hypothèse $\hat{W}$, il faut qu'il soit présent dans le vocabulaire V servant à construire W. De plus, le modèle de langage et le modèle de prononciation doivent lui attribuer une probabilité non nulle.

1.1.4 Vue d'ensemble du système de reconnaissance automatique de la parole

Comme nous venons de le voir, la formulation théorique de Jelinek (1976) du problème de la reconnaissance automatique de la parole nécessite deux modèles distincts :

- le modèle de langage, qui fournit la probabilité *a priori* d'une séquence de mots
- le modèle acoustique, qui fournit la probabilité d'une séquence de mots étant donné une observation acoustique

Voici une vue schématique du processus de reconnaissance automatique de la parole décrit plus haut :

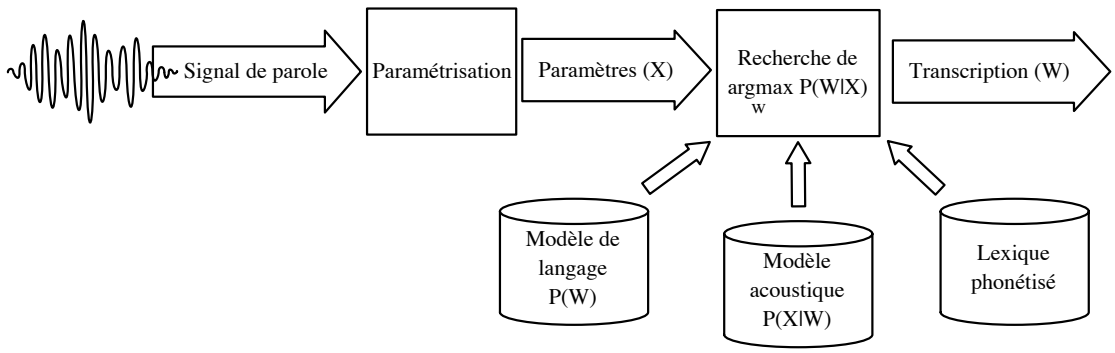

FIGURE 1.2 – *Rerpésentation schématique du processus de décodage.*

Cependant, un système de reconnaissance automatique de la parole inclut d'autres éléments en plus du décodage que nous venons de voir. En effet, il est nécessaire d'analyser et de normaliser les documents audio avant leur traitement. Par exemple, afin de pouvoir manipuler le signal acoustique de manière plus appropriée, les modèles acoustiques utilisent généralement une représentation de ce signal sous la forme d'une suite de vecteurs de paramètres. Chaque vecteur couvre une certaine portion du signal. Pour pouvoir décoder un signal de parole, il est donc nécessaire de le représenter sous cette forme vectorielle. La plupart des systèmes actuels suivent donc les étapes suivantes :

1. Paramétrisation : transformation d'un signal audio en une succession de vecteur de paramètres acoustiques
2. Segmentation : découpage du signal en segments *homogènes*
3. Identification de la parole : identification des segments contenant de la parole et élimination des autres (musique, silence, etc.)
4. Décodage des segments de parole : le processus de décodage proprement dit, c'est à dire la recherche de la séquence de mots de probabilité maximale

En plus de ces étapes, beaucoup de systèmes ajoutent désormais des passes de raffinement successives de la transcription. Elles servent en général à incorporer de nouvelles sources d'information qu'il aurait été difficile d'inclure directement dans le processus initial, comme un modèle de langage d'ordre plus élevé ou des connaissances sémantiques et syntaxiques.

Le système utilisé au cours des travaux présentés ici possède une passe d'adaptation non supervisée des modèles acoustiques au locuteur. La succession d'étapes est alors :

1. Paramétrisation : transformation d'un signal audio en une succession de vecteur de paramètres acoustiques
2. Segmentation : découpage du signal en segments *homogènes*
3. Identification de la parole : identification des segments contenant de la parole et élimination des autres (musique, silence, etc.)
4. Identification des locuteurs : identification des segments prononcés par les même locuteurs
5. Première passe de décodage : le processus de décodage est effectué une première fois, produisant une première hypothèse de transcription
6. Adaptation des modèles acoustiques : les modèles acoustiques sont adaptés pour chaque locuteur à l'aide de la première hypothèse de transcription
7. Seconde passe de décodage : le processus de décodage est effectué une seconde fois, mais avec les modèles acoustiques adaptés aux locuteurs, la transcription produite est donc théoriquement meilleure que la première.

La suite de ce chapitre sera consacrée à la partie qui nous intéresse : le moteur de reconnaissance automatique de la parole. Nous allons présenter plus en détails les différents éléments qui lui sont directement liés et que nous avons présenté sur la figure 1.2.

1.2 Paramétrisation

1.2.1 Introduction

Le signal de parole contient de nombreuses informations en plus du message linguistique. On y trouve par exemple des informations sur le locuteur, sur les conditions d'enregistrement, sur le canal de transmission, etc. La plupart de ces informations ne

sont pas directement utiles pour la tâche de reconnaissance automatique de la parole et peuvent avoir un rôle néfaste.

Le but de la paramétrisation est de transformer le signal de parole en une représentation où les informations inutiles ou nuisibles sont éliminées. Cette étape est réalisée en extrayant du signal vocal les paramètres pertinents à l'aide de méthodes d'analyses issues, pour la plupart, du domaine du traitement du signal.

1.2.2 Principe

Le signal de parole varie au cours du temps. Pour effectuer une paramétrisation avec le minimum de perte possible, il est nécessaire de l'effectuer à intervalles suffisamment réguliers pour capturer chaque variation du signal. Il est généralement admis que le signal de parole peut être considéré comme stationnaire sur une fenêtre de l'ordre de 10 millisecondes. C'est pour cela que la paramétrisation est généralement effectuée sur des fenêtres glissantes de 30 millisecondes décalées de 10 millisecondes.

Pour réduire les discontinuités dans le signal et ainsi améliorer la qualité de l'analyse, les fenêtres de 30 millisecondes sont en général pondérées par une fenêtre temporelle. On utilise souvent la fenêtre de Hamming qui permet d'atténuer le signal en bordure de fenêtre, dont la formule est :

$$h(t) = \begin{cases} 0.54 - 0.46 \cos 2\pi \dfrac{t}{T-1} & \text{si } t \in [0, T-1] \\ 0 & \text{sinon.} \end{cases} \tag{1.5}$$

où T est la taille de la fenêtre en nombre d'échantillons.

De plus, les sons aigus sont toujours plus faibles en énergie que les sons graves. C'est pour cela qu'on applique généralement au signal un filtre de préaccentuation pour les rehausser.

1.2.3 Méthodes utilisées

Les fenêtres présentées précédemment sont utilisées par quasiment toutes les approches de paramétrisation. C'est la manière d'extraire les paramètres ainsi que leur nature qui fait la différence entre les approches.

De nombreuses méthodes d'analyse ont été testées. Celles qui fonctionnent le mieux sont celles basées sur des modèles de perception. Le signal de parole a des particularités et les méthodes d'analyse qui les exploitent semblent les plus performantes. Ce sont ces méthodes qui sont le plus largement utilisées à l'heure actuelle. Nous allons présenter brièvement les plus utilisées.

Le modèle de prédiction linéaire

L'analyse de prédiction linéaire (*Linear Predictive Coding* , LPC), comme l'expliquent Markel et Gray (1974), est basée sur le fait qu'un échantillon à un instant t, $s(t)$, d'un signal vocal peut être approximé par une combinaison linéaire des T échantillons précédents :

$$s(t) \approx \sum_{i=1}^{T} \alpha_i s(t-i) \tag{1.6}$$

L'analyse LPC fournit alors les paramètres de prédiction α_i pour chaque trame. Ils peuvent être utilisés directement pour représenter le signal en reconnaissance automatique de la parole.

L'analyse cepstrale

Le signal de parole est issu de la convolution d'une source, la fréquence fondamentale F_0, et d'un conduit, les fréquences formantiques. Pour déconvoluer ce signal, il est plus pratique de le transposer dans un espace où la convolution est remplacée par une somme. C'est ce que permet l'analyse cepstrale par le passage dans le domaine log-spectral (Bogert et al., 1963).

Pour obtenir le cepstre d'un signal $s(n)$, on commence par calculer son spectre $S(f)$ grâce à une transformée de Fourier, puis on applique une transformée de Fourier inverse au logarithme de ce spectre. Les coefficients cepstraux sont donnés par :

$$c(n) = 1/N \sum_{j=0}^{N-1} \log(|S(j)|) e^{\frac{2i\pi jn}{N}} \tag{1.7}$$

Le résultat de cette analyse n'est pas utilisée directement en reconnaissance automatique de la parole mais les techniques de pramétrisation en dérivent souvent. On peut citer par exemple la méthode de paramétrisation *Mel Frequency Cepstrum Coefficients* (MFCC), qui a été proposée par Davis et Mermelstein (1980). Le principe est de calculer des coefficients c_i en utilisant une échelle de Mel. C'est une échelle de fréquences perceptive qui modélise, à l'aide d'un ensemble de filtres passe-bande, la réponse en fréquence du système auditif humain. Les M coefficients cepstraux sont alors calculés à l'aide d'une transformée en cosine discrète (*Discrete Cosine Transform*, DCT) :

$$c_i = \sum_{j=0}^{M} S(j) cos(i(j-1/2)\pi/N_f) \tag{1.8}$$

où N_f désigne le nombre de filtres.

L'analyse perceptive

Partant du constat que l'appareil auditif humain est très performant pour reconnaître la parole même dans des conditions difficile, des travaux ont été menés dans le but de s'inspirer de ce système pour améliorer les méthodes d'analyse de la parole. L'idée est de mettre à profit les connaissances que nous avons du processus perceptif humain pour modéliser de manière plus fine les processus auditif et s'en servir pour mettre en oeuvre des méthodes d'analyse plus robustes.

Par exemple la méthode d'analyse *Perceptual Linear Predictive* (PLP), proposée par Hermansky (1990) vise à introduire des connaissances issues de la psycho-acoustique dans l'estimation des modèles auto-régressifs semblables à ceux utilisés dans l'analyse de prédiction linéaire.

Techniquement, il s'agit d'ajouter à une analyse LPC une résolution en fréquence non linéaire à l'aide de bandes critiques sur une échelle de Bark, une préaccentuation du signal non linéaire selon une courbe isotonique et une compression en racine cubique du spectre résultant pour simuler la loi de perception humaine en puissance sonore. Les coefficients PLP sont ensuite calculés comme le sont les coefficients LPC.

1.2.4 Le vecteur de paramètres

Quelque soit la méthode utilisée, à chaque trame extraite du signal est associée un ensemble de paramètres. Le nombre de paramètres à extraire est déterminé par la méthode d'analyse utilisée et la précision de représentation que l'on souhaite. Il est habituel de conserver les douze premiers coefficients pour les méthodes d'analyse MFCC et PLP. On ajoute généralement un paramètre qui est le logarithme de l'énergie normalisée.

Le vecteur de paramètres représentant une trame est généralement composé de ces paramètres de base et de leur dérivées premières et secondes pour modéliser l'évolution temporelle du signal. Pour treize coefficients de base, on obtient donc un vecteur de 39 paramètres.

1.3 Modèles acoustiques

1.3.1 Introduction

La modélisation acoustique consiste en l'estimation de la probabilité $P(H|X)$ de la formule 1.3. Il s'agit de la probabilité de la séquence d'unités acoustiques élémentaires H étant donné le signal observé. Dans la plupart des cas les unités acoustiques élémentaires utilisées sont des phonèmes. Ces unités acoustiques n'ont pas de frontières apparentes dans le signal. Pour les modéliser dans le cadre de la reconnaissance automatique de la parole, il est généralement utilisé un ensemble de modèles de Markov cachés (*Hid-*

den Markov Model, HMM) (Rabiner, 1989). Chaque unité acoustique est représentée par un modèle de Markov caché.

1.3.2 Les modèles de Markov cachés

On modélise généralement chaque phonème de la langue à reconnaître par un modèle de Markov caché unidirectionnel à trois états émetteurs. Le premier état émetteur représente l'attaque du phonème, celui du milieu représente la partie centrale du phonème et le dernier modélise la queue du phonème. Un tel modèle est représenté sur la figure 1.3.

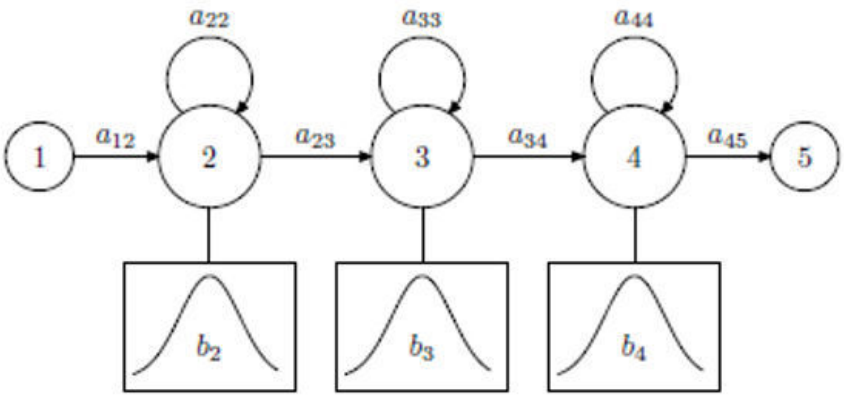

FIGURE 1.3 – *Représentation d'un modèle de Markov caché à 5 états dont 3 sont émetteurs.*

Un modèle de Markov caché est défini par les éléments suivants :
- Le nombre d'états qu'il contient : N. Avec un état i pour $i \in \{1, \ldots, N\}$.
- Les probabilités de transition entre états, sous la forme d'une matrice de dimension $N \times N$: A. $a_{i,j}$ est la probabilité de passage de l'état i à l'état j.
- Pour chaque état i, la probabilité qu'il s'agisse de l'état initial : π_i. Ces probabilités constituent l'ensemble Π.
- Pour chaque état i, un modèle d'émission fournissant la probabilité d'émettre une observation acoustique donné sachant que le processus markovien est dans cet état : $b_i(x)$. Ces lois d'émission forment l'ensemble B. Comme les vecteurs acoustiques prennent des valeurs continues, il s'agit d'une densité de probabilité

La probabilité qu'une suite de vecteurs acoustiques $X = (x_1, \ldots, x_T)$ aient été générés par la suite d'états $E = (s_0, \ldots, s_T)$ est données par la densité de probabilité conjointe suivante :

$$f(X, S) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}, s_t} \times b_{s_t}(x_t) \tag{1.9}$$

La distribution de probabilités d'émission d'une observation acoustique x_t sachant

que le processus markovien est dans l'état i est généralement approché par un mélange de K lois normales (communément appelé mixture de gaussiennes) :

$$b_i(x_t) = \sum_{k=1}^{K} c_{ik} \mathcal{N}(x_t | \mu_{ik}, \Sigma_{ik}) \tag{1.10}$$

avec :

- c_{ik} le poids de la loi k dans le mélange de l'état i, avec $\sum_{k=1}^{K} c_{ik} = 1$.
- Σ_{ik} la matrice de covariance de la loi k dans l'état i
- μ_{ik} la moyenne de la loi k dans l'état i

1.3.3 Estimation des paramètres des modèles de Markov cachés

Un HMM est donc représenté par l'ensemble des paramètres suivants : $\Lambda = \{\Pi, A, B\}$. L'estimation de ces paramètres est généralement effectué de manière empirique à l'aide d'un gros volume de données annotées. Il existe plusieurs techniques d'estimation de ces paramètres.

Estimation par maximum de vraisemblance

Une méthode d'estimation des paramètres d'un HMM est basée sur le principe du maximum de vraisemblance. Il s'agit de déterminer l'ensemble de paramètres du HMM qui maximisent la vraisemblance des données d'apprentissage.

Etant donné un ensemble d'observations $X = (x_1, \ldots, x_T)$ suivant une loi $f(x_i, \theta)$, de paramètres θ, on cherche à déterminer les paramètres θ maximisant la log-vraisemblance donnée par :

$$L(X; \theta) = \sum_{i=1}^{T} \log f(x_i, \theta) \tag{1.11}$$

L'estimation des paramètres θ n'est pas directe. Il faut généralement passer par une procédure récursive. L'algorithme *Expectation Maximization* (EM), proposé par Dempster et al. (1977), est le plus souvent utilisé. Ce sont les formules de réestimation de Baum-Welch (Baum et al., 1970) qui fournissent les valeurs des paramètres.

L'algorithme consiste à alterner des phases d'estimation de l'espérance (*Expectation*) et de maximisation (*Maximization*). Lors des étapes d'espérance, on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées. Lors des étapes de Maximization, on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

Estimation par discrimination maximale

Au lieu de maximiser, comme précédemment, la vraisemblance des données, il est possible d'utiliser d'autres critères d'optimisation pour l'estimation des paramètres des modèles.

L'approche *Maximum Mutual Information* (MMI) propose par exemple de maximiser l'information mutuelle entre les données d'entraînement et les HMM. Contrairement à l'approche de maximum de vraisemblance où les paramètres des modèles sont estimés individuellement, dans l'approche MMIE on va chercher à maximiser la vraisemblance d'un modèle tout en minimisant la vraisemblance des autres sur les mêmes données. Cette approche peut donc être considérée comme discriminante. Elle abouti sur un jeu de paramètres qui maximisent la dépendance statistique entre chaque modèle et ses données.

Cette approche a été proposée par Bahl et al. (1986) pour l'estimation des paramètres des modèles de Markov cachés dans une tâche de reconnaissance de mots isolés. Plus tard, Valtchev et al. (1997) proposeront l'adaptation de cette approche pour les systèmes de reconnaissance automatique de la parole continue grand vocabulaire.

Autres approches

Il est naturellement possible d'utiliser d'autres fonctions objectif lors de l'optimisation des paramètres des modèles. Avec l'augmentation du volume de données d'entraînement disponibles et l'augmentation des puissances de calcul des ordinateurs, des critères d'optimisation de plus haut niveau sont apparus.

Par exemple l'approche *Minimum Phone Error* (MPE) proposée par Povey et Woodland (2002) propose de minimiser les erreurs faites sur les phonèmes. L'algorithme d'estimation qu'ils proposent est semblable à celui de l'estimation MMI.

D'une manière similaire, Heigold et al. (2005) ont par exemple proposé de minimiser directement l'erreur commise sur les mots : le *Minimum Word Error* (MWE).

1.3.4 Adaptation des modèles acoustiques

Les modèles acoustiques sont habituellement estimés *a priori* sur les données d'entraînement du système de reconnaissance automatique de la parole. Si les conditions acoustiques des documents à transcrire changent, il faut théoriquement estimer de nouveaux modèles pour ces données. Cependant, lorsque peu de nouvelles données sont disponibles ou que le lourd processus de ré-estimation des modèles ne peut être effectué, il est possible d'adapter les anciens modèles aux nouvelles données.

C'est par exemple la solution retenue pour réaliser l'adaptation non supervisée des modèles acoustiques aux locuteurs dans les systèmes de reconnaissance automatique

de la parole récents que nous avons présenté dans la section 1.1.4. Dans une telle situation le volume des données d'adaptation est en général insuffisant pour estimer de nouveaux modèles.

Il existe plusieurs approches pour adapter les modèles acoustiques. Nous allons présenter les deux plus répandues.

Adaptation par maximum *a posteriori*

La méthode d'estimation par maximum *a posteriori* (MAP) a été introduite dans le cadre de la reconnaissance automatique de la parole par Gauvain et Lee (1994). Elle consiste à maximiser la probabilité d'un jeu de paramètres θ étant donné un ensemble d'observations acoustiques X :

$$\hat{\theta} = \arg\max_{\theta} P(\theta|X) \tag{1.12}$$

Comme le calcul de $P(\theta|X)$ est difficile, on utilise la formule de Bayes pour inverser les dépendances et obtenir la formule suivante :

$$\hat{\theta} = \arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)} \tag{1.13}$$

Or comme il s'agit d'une maximisation et que le dénominateur de la formule ne dépend pas de θ, on peut la simplifier ainsi :

$$\hat{\theta} = \arg\max_{\theta} P(X|\theta)P(\theta) \tag{1.14}$$

Ici $P(\theta)$ représente la probabilité *a priori* du jeu de paramètres θ. On peut noter que si $P(\theta)$ possède une distribution uniforme, alors cette méthode est identique à la méthode d'estimation par maximum de vraisemblance présentée dans la section 1.3.3. Par contre, comme le montrent Gauvain et Lee (1994), l'approche de maximum *a posteriori* nécessite moins de données d'entraînement par rapport à l'approche de maximum de vraisemblance pour des performances équivalentes.

Cette approche est souvent utilisée pour adapter des modèles acoustiques existants à de nouvelles données. On part du jeu de paramètres initial des modèles et on effectue quelques itérations de cette approche pour obtenir des paramètres qui modélisent mieux les données.

Adaptation par régression linéaire

Une autre approche permettant d'adapter les paramètres des modèles à de nouvelles données est la régression linéaire. Le critère à maximiser peut être la vraisem-

blance, comme lors de l'estimation des modèles initiaux. Cette approche a été proposée par Gales (1998) et communément appelée *Maximum Likelihood Linear Regression* (MLLR).

L'idée de base de cette approche est d'estimer un nouveau jeu de paramètres qui maximisent la vraisemblance des données d'adaptation par une transformation linéaire des paramètres initiaux des modèles. Cette transformation linéaire est obtenue grâce à la méthode de régression linéaire.

Cette approche a pour inconvénient la complexité du calcul de la transformation. Pour diminuer cette complexité, on pratique souvent une régression par classe, qui permet de réduire le nombre de paramètres à adapter.

1.3.5 Adaptation des paramètres

Il existe également des techniques qui permettent d'adapter les nouvelles données observées aux anciens modèles acoustiques. L'avantage de ces approches est qu'il n'y a pas besoin d'adapter les modèles.

Régression linéaire des paramètres

L'adaptation *feature Maximum Likelihood Linear Regression* (fMLLR) proposée par Gales (1998) est l'une de ces approches. L'idée est d'effectuer une adaptation MLLR des vecteurs de paramètres issus des nouveaux documents à transcrire pour les rapprocher de ceux issus des données d'entraînement des modèles acoustiques initiaux. Ainsi, une fois les transformations MLLR appliquées aux vecteurs d'observation, ils peuvent être décodés avec les modèles acoustiques initiaux.

Normalisation du conduit vocal

La technique appelée *Vocal Tract Lenght Normalization* (VTLN) en est un autre exemple. Elle a été proposée par Zhan et Waibel (1997) et consiste à modifier les vecteurs d'observation pour effectuer une normalisation du conduit vocal. En effet, cet aspect diffère d'un locuteur à l'autre et ajoute du bruit dans les paramètres. Ces variations sont éliminées par des filtres de fréquences.

1.3.6 Choix des unités acoustiques

Traditionnellement, l'unité acoustique de référence des systèmes de reconnaissance automatique de la parole est le phonème. Chaque phonème peut être modélisé par un modèle de Markov caché.

Les modèles acoustiques des systèmes de reconnaissance automatique de la parole récents modélisent les phonèmes en contexte. C'est ce que l'on appelle des allophones.

Cette technique permet une modélisation acoustique plus discriminante et donc garantit en général de meilleurs performances.

En pratique, la modélisation des phonèmes en contexte pose des problèmes d'estimation. En effet, le nombre de modèles augmente exponentiellement avec la taille du contexte. Par exemple avec 33 unités phonétiques de base il faut théoriquement estimer $33^3 = 35937$ modèles si l'on veut modéliser des triphones. En plus de ce problème de complexité, il existe un problème de disponibilité des données d'entraînement. En effet, pour un corpus d'entraînement de taille fixe, plus il y a de modèles à estimer, moins il y a de données pour chaque modèle. Il arrive même souvent que certains contextes ne soient pas présents dans le corpus d'entraînement.

Pour résoudre ce problème et réduire la complexité des modèles, on utilise souvent des modèles contextuels avec partage d'états. Le principe est de regrouper les états des modèles qui sont proches pour pouvoir les estimer sur un plus gros volume de données plutôt que d'estimer des modèles séparés sur moins de données. Une solution proposée par Young et al. (1994) consiste à utiliser un arbre de décision phonétique pour regrouper les états qui peuvent être considérés comme contextuellements équivalents. Ainsi toutes les observations acoustiques affectées initialement aux différents états d'un même groupe seront utilisées pour estimer l'état représentatif du groupe.

1.4 Le lexique

1.4.1 Introduction

Le lexique d'un système de reconnaissance automatique de la parole, aussi appelé vocabulaire, est le composant qui fait le lien entre modélisation acoustique et modélisation linguistique. Il s'agit d'une liste de mots auxquels sont associées leurs différentes prononciations. Chaque mot du lexique est présent dans le modèle de langage et chaque unité phonétique utilisée dans la prononciation des mots est présente dans le modèle acoustique.

La prononciation d'un mot est fournie sous la forme d'une suite de sons élémentaires de la langue considérée. Ces unités sont celles que modélisent les modèles acoustiques. Les représentations phonétiques sont renseignées manuellement par des experts ou produites par un système de conversion graphème-phonème.

1.4.2 Probabilité des variantes

A chaque mot est associé à une ou plusieurs représentations phonétiques qui correspondent souvent à des variations régionales. Par exemple le mot *pneu* peut être prononcé /pnø/ ou /pənø/ suivant l'accent régional du locuteur. Ces variantes peuvent être probabilisées de différentes manières. Cette probabilité est le terme $P(H|W)$ de

l'équation 1.4. Une des méthodes les plus courantes est d'utiliser une distribution unigramme des différentes prononciations pour chaque mot. La probabilité de la prononciation h d'un mot W est alors :

$$P(h|W) \approx \frac{F(h)}{\sum_{i \in H_W} F(i)} \tag{1.15}$$

avec $F(h)$ la fréquence de la prononciation h dans le corpus d'entraînement et H_W l'ensemble des prononciation du mot W.

On remarque cependant que beaucoup de systèmes affectent une probabilité uniforme aux variantes de prononciation.

1.4.3 Choix des mots et des phonèmes

L'ensemble des symboles utilisés pour représenter la phonétisation des mots est déterminé en fonction de la langue et de la nature des modéles acoustiques du système. Il s'agit des plus petites unités modélisées par le système. Les choix qui sont faits à ce niveau ont donc des répercutions sur l'ensemble du sysTème et principalement sur le modèle acoustique et sur le lexique.

L'ensemble des mots qui composent le lexique est déterminé par les choix effectués lors de la construction du modèle de langage.

Si un mot n'est pas présent dans le lexique, il ne pourra en aucune manière faire parti d'une hypothèse du système de reconnaissance automatique de la parole. Les mots présents dans les documents à transcrire qui ne sont pas présents dans le lexique du système de reconnaissance automatique de la parole qui effectue cette tâche sont appelés mots hors-vocabulaires.

1.4.4 Couverture lexicale

La notion de couverture lexicale est très importante en modélisation du langage. Elle permet d'estimer l'adéquation entre le lexique et un corpus donné en fournissant une mesurant du nombre de mots inconnus que le corpus contient par rapport au lexique. Le calcul de la couverture lexicale d'un vocabulaire V sur un corpus C est le suivant :

$$Couv_C = \frac{|V \cap C|}{|C|} \tag{1.16}$$

Il s'agit du nombre de mots hors-vocabulaires rencontrés dans le corpus C, normalisé par la taille de ce corpus. Plus la couverture lexicale est faible, moins le vocabulaire est adapté au corpus. Plus la couverture est proche de 1, plus le lexique est adapté au

corpus. Cependant, avec cette métrique, même avec une couverture totale il est possible que des mots du lexique ne se trouvent pas dans le corpus. Le lexique permet donc de reconnaître tous les mots du corpus mais reste partiellement inadapté.

Pour contourner le problème énoncé précédemment, il est parfois utilisé une autre métrique de couverture dont la formule est :

$$Couv_C = \frac{|V \cap C|}{|V|} \tag{1.17}$$

La normalisation est faite sur la taille du lexique au lieu de la taille du corpus. De cette manière, si des mots du lexique ne sont pas présents dans le corpus, ils contribueront à faire baisser la couverture. Par contre un vocabulaire peut très bien couvrir totalement un corpus alors que certains mots du corpus ne sont pas présents dans ce lexique.

Il existe donc deux manières de calculer la couverture. Chacune est adaptée à une situation précise. Par exemple, pour mesurer l'adéquation d'un lexique à un document à transcrire, c'est la première méthode qu'il faut employer car l'objectif est de posséder dans le lexique l'intégralité des mots présents dans le document. Par contre, si il est question de trouver le corpus le plus adapté à l'estimation de probabilités n-grammes pour un lexique donné, c'est plutôt la seconde métrique qu'il faut utiliser car ce que l'on souhaite c'est pouvoir recueillir des statistiques sur l'intégralité des mots du lexique.

1.5 Le modèle de langage

1.5.1 Introduction

Le modèle de langage est l'élément du système de reconnaissance automatique de la parole qui fournit la probabilité $P(W)$ de l'équation 1.4. Un modèle de langage a donc pour but d'estimer la probabilité *a priori* de toutes les séquences de mots qu'il est possible de construire à partir du lexique. Pour ce faire, il peut s'appuyer sur différentes sources d'informations, comme par exemple des règles syntaxiques ou sémantiques, ou encore des statistiques issues de gros volumes de données. Nous nous concentrerons ici sur les modèles de langages statistiques.

Les séquences de mots dont le modèle de langage fournit la probabilité sont composées de mots appartenant au lexique du système de reconnaissance automatique de la parole définit *a priori*, à partir d'observations et de statistiques collectées sur un corpus textuel de grande taille. Ce type de modèles est très utilisé dans divers domaines du traitement automatique du langage, comme par exemple en traduction automatique (Brown et al., 1990) ou en recherche d'information (Ponte et Croft, 1998).

En reconnaissance automatique de la parole il s'agit d'un élément clef puisque c'est lui qui introduit les contraintes linguistiques au sein du processus de décodage.

Les travaux présentés dans cette thèse ont pour objectif d'améliorer la modélisation du langage. Le chapitre 2 est intégralement dédié à cet aspect de la reconnaissance automatique de la parole, plus de détails y seront fournis.

1.5.2 Modélisation statistique du langage

Le but d'un modèle de langage statistique est d'estimer la probabilité *a priori* d'une séquence de M mots W : $P(W)$. Cette probabilité peut être décomposée en un produit de probabilités conditionnelles :

$$P(W) = \prod_{i=1}^{M} P(w_i|w_1, w_2, \ldots, w_{i-1}) \tag{1.18}$$

Cette formulation suppose qu'un mot w_i peut être prédit uniquement à partir de l'historique des mots qui le précèdent.

1.5.3 Les modèles de langage n-grammes

En reconnaissance automatique de la parole statistique, les modèles de langage utilisés quasiment exclusivement sont des modèles de langage n-grammes. Ces modèles ont été proposés par Jelinek (1976). Depuis, de nombreux travaux ont eu pour but de proposer de meilleurs modèles mais les gains obtenus sont en général marginaux. Ce type de modèles constitue donc l'état de l'art en modélisation du langage.

Principe

L'idée à la base des modèles de langage n-gramme est que la probabilité d'apparition d'un mot peut être estimée à partir des $n-1$ mots le précédant. On peut ainsi faire une approximation sur le contexte utilisé pour le calcul des probabilités conditionnelles de la formule 1.18 :

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx P(w_i|h_i^n) \tag{1.19}$$

avec h_i^n l'historique de taille $n-1$ du mot w_i.

La probabilité $P(W)$ d'une séquence de M mots peut alors s'écrire :

$$P(W) \approx \prod_{i=1}^{M} P(w_i|h_i^n) \tag{1.20}$$

Cette formulation réduit le nombre de dépendances permettant de prédire l'apparition d'un mot, ce qui rend l'estimation de la probabilité associée réalisable avec un corpus de taille raisonnable. Le paramètre n du modèle sert à régler la taille de l'historique

prise en compte pour la prédiction. Plus n est grand, plus précise est la modélisation, mais plus grand est le corpus d'entraînement nécessaire à une estimation correcte des probabilités.

Estimation

L'estimation d'un tel modèle se fait en général en utilisant le principe du maximum de vraisemblance. Ce maximum est atteint si l'on utilise la fréquence relative du mot :

$$P(w_i|h_i^n) \approx \frac{tf(h_i^n, w_i)}{tf(h_i^n)} \tag{1.21}$$

avec $tf(W)$ la fréquence de la chaîne de mots W dans le corpus d'entraînement du modèle de langage.

Le problème de cette estimation est que les chaînes de n mots non observées dans le corpus d'entraînement auront une probabilité nulle alors qu'il se peut qu'elles soient possibles mais juste absentes de ce corpus. Pour remédier à ce problème, on applique généralement une technique de décompte sur les fréquences des n-grammes observés dans le corpus d'entraînement afin de redistribuer la masse de probabilité associée aux événements non observés dans ce corpus. La redistribution est souvent effectuée à l'aide d'un modèle de langage d'ordre n inférieur :

$$\hat{P}(w_i|h_i^n) = \begin{cases} P(w_i|h_i^n) & \text{si } (h_i^n, w_i) \in C_n \\ \alpha(h_i^n, w_i) \times P(w_i|h_i^{n-1}) & \text{si } (h_i^n, w_i) \notin C_n \end{cases} \tag{1.22}$$

avec C_n l'ensemble des séquences de n mots du corpus d'entraînement et $\alpha(h_i^n, w_i)$ la masse de probabilités à redistribuer sur les événements non vus pour le n-gramme $\alpha(h_i^n, w_i)$.

1.6 Le décodeur

1.6.1 Introduction

Le décodeur est le composant logiciel d'un système de reconnaissance automatique de la parole qui va produire une ou plusieurs hypothèses de transcription à partir des observations acoustiques. Il va chercher la suite de mots ayant la plus forte probabilité étant donné le modèle de langage, le modèle acoustique et le lexique utilisé.

Une solution naïve à ce problème consiste à explorer toutes les hypothèses qu'il est possible de produire à partir des mots du lexique. Cependant, pour un système de reconnaissance automatique de la parole grand vocabulaire actuel qui comporte entre 50 000 et 150 000 mots, cette solution n'est pas réalisable techniquement. Pour résoudre

ce problème, on utilise un algorithme de décodage qui va permettre de trouver une solution optimale ou, en tout cas, proche de l'optimalité sans avoir besoin d'explorer l'ensemble des hypothèses.

Nous allons présenter ici les caractéristiques d'un tel décodeur. Nous étudierons plus particulièrement celui employé dans le système de reconnaissance automatique de la parole utilisé dans les travaux présentés ici : le système SPEERAL. Pour plus de détails sur ce système, une description plus complète est fournie par Nocera et al. (2002a), Nocera et al. (2002b) et Nocéra et al. (2004).

1.6.2 Algorithme de décodage

La plupart des algorithmes de décodage ont pour but d'explorer le graphe de mots que représente l'ensemble des hypothèses possibles à partir des mots du lexique. Il n'est pas possible de construire *a priori* un tel graphe, c'est pour cela qu'il est habituellement construit dynamiquement au fur et à mesure de l'avancée de l'algorithme d'exploration.

Toute la finesse de l'algorithme est de trouver le chemin du graphe allant du noeud de départ jusqu'au noeud d'arrivée ayant la probabilité acoustique et linguistique la plus élevée tout en explorant un minimum de chemins alternatifs. Il existe deux grandes familles d'algorithmes de décodage : les algorithmes synchrones et asynchrones.

Algorithmes synchrones

Il s'agit des algorithmes les plus utilisés dans les systèmes de reconnaissance automatique de la parole. Le plus répandu dans cette famille est l'algorithme de Viterbi en faisceau (*beam search*). Le principe de ces algorithmes est d'explorer le graphe d'hypothèses de manière synchronisée avec le signal de parole.

L'inconvénient de tels algorithmes est qu'il est très difficile d'intégrer des informations contextuelles dans la fonction de coût utilisée. Dans le cadre de la reconnaissance automatique de la parole il est essentiel de pouvoir utiliser un score linguistique contextuel fourni par le modèle de langage n-gramme à la fonction de coût. Il est possible de modifier l'algorithme pour intégrer cette information en créant artificiellement des chemins du graphe dépendant de l'historique. Cette modification est très lourde, c'est pour cela que les systèmes de reconnaissance automatique de la parole basés sur ce type d'algorithmes effectuent en général un décodage initial avec des modèles de langage bigrammes.

Algorithmes asynchrones

L'idée de ces algorithmes est d'explorer en profondeur le graphe. Les hypothèses ayant le score le plus élevé sont explorées en premier. L'algorithme est dit asynchrone

car il peut, après avoir longuement exploré un chemin du graphe, décider de revenir en arrière et de prendre un autre chemin.

Ce type d'approche a été décrit par Jelinek (1969). Une implémentation utilisant l'algorithme A∗ a été proposée par Paul (1992) et est la plus couramment utilisée. C'est d'ailleurs cette implémentation qu'utilise le système de reconnaissance automatique de la parole SPEERAL utilisé lors de ces travaux.

1.6.3 Représentation des hypothèses

Lors du processus de décodage, un système de reconnaissance automatique de la parole génère un ensemble d'hypothèses explorées sous la forme de séquences de mots. On représente habituellement ces hypothèses par un graphe acyclique dirigé appelé treillis de mots. Dans un tel graphe, chaque noeud correspond à un instant donné, et chaque arc est une hypothèse de mot pondérée par un score linguistique et un score acoustique.

Un exemple de treillis de mot est représenté sur la figure 1.4. On voit très clairement que le treillis de mots est une représentation compacte des hypothèses puisque les chemins aboutissant aux mêmes mots sont mutualisés.

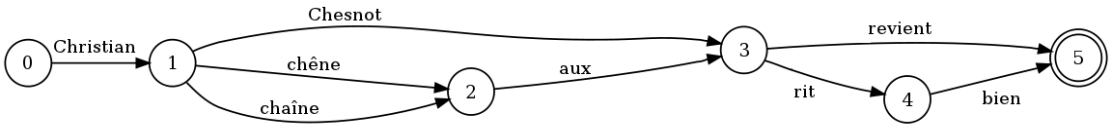

FIGURE 1.4 – *Exemple de représentation des hypothèses en treillis de mots.*

Il existe un certain nombre d'algorithmes éprouvés pour parcourir un tel graphe à la recherche d'une solution optimale qui ont des caractéristiques différentes, comme l'algorithme A∗ ou Viterbi.

1.6.4 Passes successives de décodage

Le mode de représentation en graphe des hypothèses explorées par le processus de décodage est très utile car il offre une représentation simple et complète des hypothèses de transcription. Il est alors possible de raffiner ce graphe avec de nouvelles informations par passes successives. La complexité est ainsi réduite par rapport à un processus de décodage unique qui intégrerai d'emblée l'ensemble des informations.

Il est par exemple de plus en plus courant d'utiliser ce principe pour étendre par étapes l'ordre du modèle de langage. Une première passe de décodage est effectuée avec un modèle de langage d'un ordre donné, par exemple des bigrammes. Le treillis de mots issu de ce processus est ensuite fourni à une seconde passe de décodage qui y remplacera les probabilité linguistiques bigrammes par celle d'un ordre supérieur. La première passe de décodage est rapide et permet de sélectionner un sous-ensemble

d'hypothèses à approfondir. La seconde passe peut donc intégrer des probabilités linguistiques plus précises et donc plus lourdes à calculer car le nombre d'hypothèses est réduit.

1.6.5 Décodage *N-best*

Le principe du décodage *N-best* est de fournir non pas la meilleure hypothèse du système de reconnaissance automatique de la parole, mais les N meilleurs hypothèses. Il existe plusieurs techniques pour y parvenir.

Directement lors du décodage initial

Il est possible de modifier l'algorithme du décodeur initial pour qu'il trouve les N meilleures hypothèses.

Avec les décodeurs reposant sur l'algorithme de Viterbi *beam search*, la recherche des N meilleurs hypothèses est aisée, comme on le voit dans les travaux de Schwartz et Chow (1990).

Pour les décodeurs asynchrones basés sur l'algorithme A$*$, l'ajout de cette fonctionnalité est tout aussi évidente. Il suffit par exemple de modifier le critère d'arrêt de l'algorithme. En temps normal, l'algorithme s'arrête lorsque l'hypothèse se trouvant en haut de la pile arrive au noeud terminal du graphe de décodage. Si on modifie ce critère pour que l'algorithme s'arrête quand N hypothèses dépilées arrivent au noeud terminal, alors ces N hypothèses constituent les *N-best* recherchées.

A partir d'un treillis de mots

La solution la plus pratique lorsque le décodeur du système de reconnaissance automatique de la parole ne comporte pas cette fonctionnalité est d'utiliser un des nombreux algorithmes qui ont été conçu pour trouver les N meilleurs chemins dans un graphe dirigé acyclique, comme l'est le treillis de mots. On peut par exemple citer l'algorithme proposé par Soong et Huang (1991).

Pour avoir un apperçu des principaux algorithmes pour effectuer cette tâche, on peut se référer aux travaux de Schwartz et Austin (1991).

1.7 Evaluation d'un système de reconnaissance automatique de la parole

Pour évaluer les performances d'un système de reconnaissance automatique de la parole, on compare généralement la transcription fournie par le système avec la transcription de référence, produite manuellement.

1.7.1 Le taux d'erreur sur les mots

La mesure la plus largement utilisée est celle du taux d'erreur sur les mots, communément appelée WER (*Word Error Rate*). Cette mesure est inspirée de concepts d'alignement dynamique. Trois types d'erreurs peuvent être commises lorsque l'on aligne deux séquences semblables :

- des substitutions : un élément d'une séquence est remplacé par un autre dans la seconde séquence
- des insertions : un élément a été inséré dans une séquence
- des suppressions : un élément d'une séquence est absent de l'autre séquence

Les deux séquences sont ici la transcription de référence et la transcription automatique. Le calcul du taux d'erreur mots est donc très simple. Il s'agit du rapport entre le nombre d'erreurs de n'importe quel type commises et le nombre total de mots d'une séquence :

$$WER = \frac{\#substitutions + \#insertions + \#suppressions}{N} \tag{1.23}$$

avec N le nombre de mots de la transcription de référence.

Cette mesure s'exprime donc en pourcentage. On peut remarquer qu'il est possible d'avoir un taux d'erreur mots supérieur à 100%, notamment à cause des insertions.

La méthode de comptabilisation des erreurs découle directement de la méthode d'alignement utilisé. Il s'agit d'un algorithme de programmation dynamique matriciel, comme l'algorithme *Dynamic Time Warping* par exemple. Il produit un alignement entre les séquences de mots qui minimise le nombre d'erreurs des trois types commises. Cette métrique d'évaluation est généralement appelée distance d'édition.

1.7.2 Autres mesures

Le premier inconvénients du taux d'erreur mots est que d'une part il accorde la même importance aux erreurs des trois types. Or, dans certaines circonstances, il est plus grave de commettre certaines erreurs que d'autres.

Le second problème de cette mesure est qu'elle ne tient pas compte de la nature des erreurs ni des caractéristiques linguistiques des mots erronés. Il semble pourtant qu'une erreur d'accord en nombre est moins grave qu'un mot oublié lors de la transcription d'une émission de radio.

De plus, les erreurs commises sur certains mots sont plus préjudiciables que d'autres suivant l'application visée pour les transcriptions. Par exemple, s'il s'agit de l'indexation de documents audio, la substitution d'un mot porteur de sens par un autre mot porteur de sens mais d'une autre thématique est certainement l'erreur la plus grave. A

l'opposé, toutes les erreurs commises sur les mots outils de la langue (articles, conjonctions, etc.) par un tel système n'ont aucune importance puisque le moteur d'indexation n'en tient habituellement pas compte.

Pour pallier ce problème, d'autres métriques ont été proposées pour évaluer la qualité d'un système de reconnaissance automatique de la parole de manière plus appropriée à l'application visée pour la transcription.

On peut citer par exemple les travaux de San-Segundo et al. (2001) et de Sarikaya et al. (2005a) qui proposent une mesure de la fidélité sémantique des transcriptions pour un système de dialogue.

La mesure proposée par Senay et al. (2011) évalue la capacité d'indexation d'une transcription automatique dans le cadre d'un système de recherche de documents audiovisuels.

Chapitre 2

La modélisation du langage

Sommaire

Ce chapitre présente plus en détails les modèles de langages les plus utilisés utilisés en reconnaissance automatique de la parole.

2.1 Introduction

Le modèle de langage est l'élément d'un système de reconnaissance automatique de la parole qui fournit la probabilité $P(W)$ de l'équation 1.4.

Un modèle de langage a donc pour but d'estimer la probabilité *a priori* de toutes les séquences de mots qu'il est possible de construire à partir du lexique. Pour ce faire, il peut s'appuyer sur différentes sources d'informations, comme par exemple des règles syntaxiques ou sémantiques, ou encore des statistiques issues de gros volumes de données. Nous nous concentrerons ici sur les modèles de langages statistiques.

Ce type de modèle de langage a pour but d'estimer la probabilité de séquences de mots appartenant à un lexique définit *a priori*, à partir d'observations et de statistiques collectées sur un corpus textuel de grande taille. Ce type de modèles est très utilisé dans divers domaines du traitement automatique du langage, comme par exemple en traduction automatique (Brown et al., 1990) ou en recherche d'information (Ponte et Croft, 1998).

En reconnaissance automatique de la parole, il s'agit d'un élément clef puisque c'est lui qui introduit les contraintes linguistiques au sein du processus de décodage.

2.2 Modélisation statistique du langage

Le but d'un modèle de langage statistique est d'estimer la probabilité *a priori* $P(W)$ de la séquence de mots W. Les M mots qui composent W appartiennent au lexique V :

$$W = (w_1, w_2, \ldots, w_M) \text{ avec } w_i \in V \tag{2.1}$$

On peut décomposer cette probabilité sous la forme d'un produit de probabilités conditionnelles, ce qui permet alors d'écrire :

$$P(W) = \prod_{i=1}^{M} P(w_i | w_1, w_2, \ldots, w_{i-1}) \tag{2.2}$$

Cette formulation suppose qu'un mot w_i peut être prédit uniquement à partir de l'historique des mots qui le précèdent. Le modèle de langage statistique est donc un ensemble de distributions de probabilités conditionnelles qui prédisent l'apparition d'un mot étant donné son historique.

2.3 Modèle de langage n-gramme

La formule 2.2 suppose l'estimation de la probabilité d'un mot sachant l'intégralité de son historique. En pratique il n'est pas possible d'estimer cette probabilité avec

un historique aussi long. En effet, le volume des corpus d'entraînement, même gros, ne permettent pas d'observer suffisamment ces séquences. D'ailleurs, la plupart des phrases qu'un moteur de reconnaissance automatique de la parole continue grand vocabulaire doit décoder n'ont jamais été rencontrées dans le corpus d'entraînement du modèle de langage.

Les modèles de langages n-grammes ont été introduits en reconnaissance automatique de la parole par Jelinek (1976) précisément pour résoudre ce problème. Puisqu'il n'est pas possible d'estimer des probabilités sur l'intégralité de l'historique des mots, alors une coupure sur les dépendances de la formule 2.2 est effectuée. Ainsi, le langage est modélisé comme une source markovienne d'ordre $n-1$:

$$P(w_i|w_1, w2, \ldots, w_{i-1}) \approx P(w_i|w_{i-n+1}, \ldots, w_{i-1}) \tag{2.3}$$

et la formule donnant la probabilité d'une séquence de mots W devient :

$$P(W) \approx \prod_{i=1}^{N} P(w_i|w_{i-n+1}, \ldots, w_{i-1}) \tag{2.4}$$

Le paramètre n du modèle sert à régler la taille de l'historique pris en compte pour la prédiction. Plus n est grand, plus précise est la modélisation, mais plus grand est le corpus d'entraînement nécessaire à une estimation correcte des probabilités.

La coupure dans l'historique dont dépend la probabilité d'apparition d'un mot a deux impacts majeurs sur la modélisation du langage.

L'aspect positif de cette coupure est qu'elle réduit la complexité du modèle et rend ainsi son estimation réalisable avec un corpus de taille raisonnable.

L'impact négatif est qu'étant donné que l'historique est réduit, la modélisation est moins précise. Ceci se vérifie en pratique car l'augmentation de l'ordre n des modèles engendre généralement de meilleurs performances en reconnaissance automatique de la parole, à condition que le corpus d'entraînement soit de taille suffisante.

2.4 Modèle de langage n-gramme de classes

Les modèles de langage n-gramme de classe ont été proposés par Brown et al. (1992a). L'idée qui a motivé ces travaux est de mutualiser les paramètres semblables du modèle de langage pour pouvoir l'estimer plus facilement. Le principe est de regrouper les mots en classes lexicales et de construire un modèle de langage sur ces classes. La prédiction d'un mot ne dépend plus des mots qui le précèdent, mais des classes des mots qui le précèdent ainsi que de sa propre classe.

2.4.1 Formulation

Le modèle n-gramme de classes fonctionne de la même manière qu'un modèle n-gramme de mots, mais la probabilité d'un mot sachant son historique est décomposé de la manière suivante :

$$P(w_i|w_{i-n+1},\ldots,w_{i-1}) = P(w_i|C(w_i)) \times P(C(w_i)|C(w_{i-n+1}),\ldots,C(w_{i-1})) \quad (2.5)$$

où $C(w_i)$ est la classe lexicale du mot w_i. On remarque que le second terme de l'équation peut être vu comme un modèle de langage n-gramme classique mais où les mots sont remplacés par leurs classes.

On pose généralement l'hypothèse simplificatrice selon laquelle un mot n'appartient qu'à une seule classe. La pertinence de cette modélisation dépend fortement du choix de l'ensemble des classes $C(\cdot)$.

2.4.2 Choix des classes

Le choix de l'ensemble des classes utilisées est déterminant pour les performances du modèle.

Une solution qui a été proposée par Jelinek (1990) est de se servir des catégories morphosyntaxiques des mots (par exemple *adverbe*, *article*, *nom commun*, etc.). De cette manière chaque classe modélise le comportement d'une catégorie de mot. Cette approche est naturelle car les règles de syntaxe et de grammaire qui régissent la production de la langue s'appuient elles aussi sur ces catégories.

Plus tard, des approches non supervisées ont été proposées permettant de trouver automatiquement un jeu de classes optimal. Par exemple Brown et al. (1992a) proposent une méthode non supervisée permettant de découvrir un nombre arbitraire de classes qui maximisent la vraisemblance du corpus d'entraînement pour un modèle bigramme.

Les travaux de Niesler et al. (1998) montrent que parmi ces deux approches, celle reposant sur une découverte non supervisée des classes donne de meilleurs résultats dans la plupart des cas.

2.5 Autres modèles de langage

Les modèles de langage n-grammes sont très performants mais ont néanmoins quelques lacunes. Tout d'abord, on reproche souvent à ce type de modèle de ne pas prendre en compte les dépendances entre des mots éloignés de plus de n mots, où n est l'ordre du modèle. Par exemple lorsque le sujet avec lequel doit être accordé un mot se trouve à plus de n mots de ce dernier, le modèle de langage n-gramme n'en tiendra pas compte.

De plus, ce type de modèle ne prends pas non plus en compte explicitement les informations thématiques locales. En effet, la probabilité d'un mot varie fortement en fonction de la thématique dans laquelle il apparaît. On peut considérer que cette information est modélisée implicitement par le contexte de taille n de chaque mot, mais, dans la plupart des cas, cette taille est insuffisante pour identifier précisément la thématique.

Beaucoup de modèles ont été proposés dans le but de combler ces lacunes. Au mieux, ces modèles sont plus performants que les modèles n-grammes dans des cas très particuliers, comme lorsqu'il y a très peu de données d'entraînement ou lorsque les données disponibles sont très bruitées.

Ces approches peuvent être scindées en deux groupes : celles reposant sur les principes d'un modèle de langage n-gramme ou en utilisant un, et celles proposant un nouveau paradigme. Quoi qu'il en soit, la plupart des approches qui arrivent à être meilleures que les modèles n-grammes sont en fait le résultat de la combinaison d'un modèle n-gramme avec d'une autres sources d'information.

Nous allons présenter quelques approches qui ont été utilisées en reconnaissance automatique de la parole.

2.5.1 Reposant sur un modèle n-gramme

Un grand nombre d'approches dépassant les modèles de langage n-grammes sont basées sur les même principes que ces derniers. On trouve par exemple le modèle *multigram* ou le modèle *cache* qui permettent d'intégrer de nouvelles informations dans le calcul des probabilités n-grammes.

Modèle *multigram*

Le modèle de langage *multigram* a été proposé par Deligne et Bimbot (1995). Il repose sur le constat que l'ordre des modèles n-grammes est de taille fixe quelque soit le mot. L'idée proposée ici est de modéliser la probabilité d'un mot étant donné un contexte de taille variable, dépendant du mot.

Techniquement, il s'agit d'une sorte de modèle de langage n-gramme avec n variant suivant le contexte. Ils montrent un gain par rapport aux modèles de langage n-grammes conventionnels sur une tâche particulière.

Modèle de langage *cache*

L'idée sur laquelle repose le modèle cache proposé par Kuhn et De Mori (1990) est qu'un mot qui apparaît à un endroit donné d'un texte a de grandes chances d'apparaître à nouveau à proximité de cet endroit. L'idée de ce type de modèles est donc de renforcer

la probabilité des mots qui sont apparus dans une certaine fenêtre temporelle précédent le mot dont on cherche la probabilité.

La probabilité d'un mot étant donné sont historique de taille n devient donc, avec un cache sur une fenêtre de taille m :

$$P(w_i|w_1,\ldots,w_{i-1}) \approx \lambda P_c(w_i|w_{i-m+1},\ldots,w_{i-1}) + (1-\lambda)P(w_i|w_{i-n+1},\ldots,w_{i-1}) \quad (2.6)$$

où $0 \leq \lambda \leq 1$ est le poids accordé au cache par rapport au modèle initial. Il s'agit d'une interpolation linéaire entre le modèle de langage initial P et le modèle cache P_c.

Le terme P_c de l'équation précédente est une distribution de probabilités qui modélise l'importance des mots présents dans la fenêtre du cache. Cette distribution a d'abord été formulée sur le principe du maximum de vraisemblance :

$$P_c(w_i|w_{i-m+1},\ldots,w_{i-1}) = \frac{tf(w_i)}{m} \quad (2.7)$$

où m est la taille du cache et $tf(w_i)$ est la fréquence du mot w_i dans ce cache.

L'inconvénient majeur de cette formulation est que toutes les occurrences des mots ont le même poids, quelque soit leur position dans le cache. Il semble pourtant que les mots les plus récents aient un poids plus important que les autres. C'est pour prendre en compte cet aspect que Clarkson et Robinson (1997) ont proposé la formulation suivante :

$$P_c(w_i|w_{i-m+1},\ldots,w_{i-1}) = \beta \sum_{j=1}^{i-1} I(w_i = w_j) e^{-\alpha(i-j)} \quad (2.8)$$

où $I(E) \in 0;1$ est une fonction auxiliaire qui vaut 1 si E est vrai et 0 sinon. $\alpha(\delta)$ est une fonction qui associe un facteur au décalage de mots δ, et β est un facteur de normalisation qui garanti que la distribution somme à 1.

Un inconvénient de cette formulation est que tous les mots présents dans le cache sont mis en valeur quelque soit leur type. Cependant, les mots outils de la langue qui ont une fréquence très élevée auront de grandes chances de figurer dans ce cache, alors qu'il n'est pas utile de les mettre en valeur puisqu'ils bénéficient déjà d'une probabilité élevée dans le modèle de langage. C'est pour cela que Rosenfeld (1994) propose de filtrer les mots du cache pour ne mettre en valeur que les mots peu fréquents.

Modèle de langage *trigger*

Le modèle *trigger* est une généralisation du modèle *cache* présenté précédemment. Il a été proposé par Lau et al. (1993) puis amélioré par Singh-Miller et Collins (2007). Le

constat qui motive cette approche est que les mots du champ lexical associé à une thématique ont plus de chance d'apparaître que les autres mots du lexique aux moments où cette thématique est abordée.

Le principe de ce modèle est donc de renforcer les probabilités des mots associés à ceux qui viennent d'être observés.

Un ensemble de thèmes est constitué *a priori* à partir du corpus d'entraînement à l'aide, par exemple, d'une analyse sémantique latente (*Latent Semantic Analysis*, LSA) comme le proposent Bellegarda et al. (1996). Un ensemble de couples de mots fortement relatés est construit. Chaque couple est ensuite associé à une thématique. Ces couples servent alors de déclencheurs (*triggers* en anglais). Lorsqu'ils apparaissent dans l'historique d'observation, ils déclenchent le renforcement des probabilités associées à cette thématique. Ce renforcement est effectué à l'aide d'une interpolation entre un modèle de langage correspondant à la thématique à laquelle ils sont associés et le modèle de langage initial.

2.5.2 Autres approches

On trouve dans la littérature un certain nombre de nouvelles approches pour la modélisation du langage. Certaines proposent de nouveaux paradigmes et dépassent l'approche classique n-gramme.

Modèle de langage connexioniste

Le modèle de langage connexioniste fournit un cadre pour l'estimation des probabilités des événements non observés par les modèles de langage n-grammes classiques. L'estimation de ces probabilités suppose la connaissance d'un lien entre les événements observés et ceux non observés mais rencontrés lors de l'utilisation du modèle. Ce lien est une forme de généralisation de l'observation initiale.

Un modèle de langage n-gramme utilise un espace de représentation discret composé des indices des mots. Cet espace permet difficilement une généralisation. Pour pallier ce problème, le modèle connexioniste proposé par Bengio et al. (2006) utilise une estimation des n-grammes dans un espace continu.

Techniquement, chaque mot est projeté dans un espace à N dimensions et c'est à partir de cette projection que la probabilité contextuelle est estimée. Ils utilisent un réseau de neurones pour estimer conjointement la projection et les distributions de probabilités.

Cette approche a fait ses preuves sur des tâches où la quantité de données d'apprentissage est limitée.

Modèle de langage syntaxique

Les travaux de Chelba et Jelinek (2000) concernent un modèle de langage basé sur les structures syntaxiques de la langue. Ils partent du constat que l'historique de taille fixe du modèle de langage n-gramme n'est pas satisfaisant parce qu'il ne prends pas en compte la nature des mots. En effet, tous les mots d'un historique ne sont peut-être pas pertinents pour prédire l'apparition d'un mot.

Ils proposent alors de filtrer les mots de l'historique n-gramme en fonction de leur catégorie morphosyntaxique. Ainsi, ils construisent un modèle dont l'historique de prédiction ne contient que les mots les plus pertinents pour la prédiction.

2.6 Estimation des modèles de langage n-grammes

Comme nous l'avons vu, un modèle de langage n-gramme associe à chaque mot w_i du vocabulaire V une probabilité conditionnelle $P(w_i|h_i^n)$ étant donné son historique h_i^n de taille $n-1$.

L'estimation de ces probabilités est effectuée à partir d'un corpus d'entraînement. Tous les mots de ce corpus qui ne figurent pas dans le vocabulaire V sont remplacés par un mot spécial de ce vocabulaire représentant l'ensemble des mots inconnus. On utilise habituellement l'étiquette <*UNK*>. Ce mot spécial modélise en fait une classe de mots : tous ceux qui ne sont pas dans le lexique.

L'estimation des probabilités conditionnelles peut alors être effectuée de différentes manières :

1. Pour maximiser la vraisemblance du corpus d'entraînement
2. Par une estimation Bayésienne
3. Par maximum *a posteriori*
4. etc.

La méthode la plus largement utilisée est celle du maximum de vraisemblance. Il a été prouvé que l'estimateur de probabilité maximisant la vraisemblance est la fréquence normalisée :

$$P(w_i|h_i^n) \approx \frac{tf(h_i^n, w_i)}{tf(h_i^n)} \tag{2.9}$$

avec $tf(W)$ la fréquence de la chaîne de mots W dans le corpus d'entraînement du modèle de langage.

Le problème de cette estimation est que les chaines de n mots non observées dans le corpus d'entraînement auront une probabilité nulle alors qu'il se peut qu'elles soient possibles mais simplement absentes de ce corpus. En effet, il n'existe pas de corpus

assez grand pour contenir l'intégralité des chaines de mots possibles à partir d'un vocabulaire de taille raisonnable.

2.7 Lissage des modèles de langage n-grammes

Bien que la formulation de la probabilité d'un mot du modèle n-gramme admette une coupure de l'historique de dépendance, l'estimation d'un tel modèle est toujours problématique. Même avec un historique réduit à quelques mots, beaucoup de n-grammes possibles sont absents des corpus d'entraînement et le nombre d'occurrence de beaucoup de n-grammes présents n'est pas assez élevé pour être significatif.

2.7.1 Phénomène de pénurie de n-grammes

Ce phénomène peut s'expliquer par la loi proposée par Zipf (1949), dite loi de Zipf. Cette loi prédit la fréquence d'un mot dans une langue naturelle en fonction de son rang dans la liste des mots triés par fréquence décroissante. Cette fréquence serait proportionnelle à l'inverse du rang du mot :

$$f \approx \frac{K}{r} \tag{2.10}$$

avec K la constante de proportionnalité qui est la plupart du temps proche de 1.

Il se trouve que cette loi s'applique également aux séquences de mots n-grammes, comme le montrent Manning et al. (1999). De ce fait, la plupart des n-grammes avec $n > 1$ observés dans un corpus auront une fréquence proche de 1, ce qui est peu significatif du point de vue statistique.

Pour pallier ce problème d'estimation on utilise généralement des techniques de lissage.

2.7.2 Principe du lissage

Le lissage est une technique qui permet d'estimer la probabilité d'événements non observés dans le corpus d'entraînement à partir de ceux qui s'y trouvent. Ainsi, tous les n-grammes qu'il est possible de construire à partir du lexique seront associés à une probabilité non nulle.

La plupart des techniques de lissage fonctionnent en deux étapes. Lors de l'estimation des probabilités des événements observés, une technique de décompte est appliquée aux fréquences d'occurrence de n-grammes afin de prélever une masse de probabilité à ces événements. Cette masse est ensuite redistribuée aux événements non observés.

Les deux aspects du lissage que sont le décompte et la redistribution sont assez indépendants. Beaucoup de méthodes ont été proposées. L'article de Chen et Goodman (1999) propose un apperçu de toutes ces approches.

2.7.3 Techniques de décompte

Les deux techniques de décompte les plus utilisées sont le décompte de Good-Turing, inspiré de la formule de Good (1953), et le décompte absolu.

Décompte de Good-Turing

Le décompte de Good-Turing part du constat que, si dans un gros corpus, une séquence de mots est peu observée, alors l'estimation de sa probabilité est peut-être sur-évaluée. En effet, il se peut que cette occurrence soit le fruit du hasard et que sa probabilité réelle soit bien inférieure à $\frac{1}{|corpus|}$.

Les travaux de Good montrent que pour obtenir une estimation plus juste de la fréquence réelle d'un n-gramme, il faudrait décompter de sa fréquence une certaine quantité. Il propose de modifier la calcul de la fréquence d'un n-gramme apparaissant r fois ainsi :

$$tf_{gt}(r) = (r+1)\frac{n_{r+1}}{n_r} \tag{2.11}$$

avec n_r le nombre de n-grammes qui apparaissent r fois dans le corpus d'entraînement :

$$n_r = |\{h_i^n, w_i | tf(h_i^n, w_i) = r\}| \tag{2.12}$$

En modélisation du langage naturel, la fréquence obtenue après le décompte est quasiment toujours inférieure à la fréquence d'origine. Une masse de probabilité est donc dégagée des événements observés et peut être redistribuée aux événements non observés. On peut noter qu'en général la masse de probabilité ainsi extraite est égale à $\frac{n_1}{|corpus|}$.

Décompte absolu

La technique du décompte absolu (*absolute discounting*) a été introduite par Ney et Essen. Il s'agit d'une technique plus simple que la précédente. Au lieu de calculer la quantité à décompter en fonction de la fréquence du n-gramme, cette quantité est fixe pour toutes les fréquences. La fréquence r d'un n-gramme devient alors :

$$tf_{ad}(r) = r - D \tag{2.13}$$

Ney et al. (1994) arrivent à la conclusion que la valeur optimale de D est :

$$D = \frac{n_1}{n_1 + 2n_2} \tag{2.14}$$

avec n_r défini à l'équation 2.12.

Cette formulation est partiellement en accord avec le décompte de Good-Turing car il a été démontré empiriquement par Church et Gale (1991) que le décompte de Good-Turing était constant pour les fréquences $r \geq 3$.

Décompte absolu modifié

En poussant ce raisonnement un peu plus loin, Chen et Goodman (1999) proposent de traiter le cas des n-grammes observés une fois et deux fois de manière séparée. Ils remplacent donc le décompte absolu par trois décomptes, D_1, D_2 et D_{3+}, correspondant aux décomptes appliqués aux n-grammes observés respectivement une fois, deux fois et trois fois ou plus.

De manière analogue au cas d'une constante unique, une estimation optimale des constantes est proposée :

$$\begin{cases} D_1 & = 1 - 2 \times \dfrac{n_1}{n_1 + 2n_2} \times \dfrac{n_2}{n_1} \\ D_2 & = 2 - 3 \times \dfrac{n_1}{n_1 + 2n_2} \times \dfrac{n_3}{n_2} \\ D_{3+} & = 3 - 4 \times \dfrac{n_1}{n_1 + 2n_2} \times \dfrac{n_4}{n_3} \end{cases} \tag{2.15}$$

Cette approche a prouvé son efficacité dans le cadre de l'étude menée par Chen et Goodman (1999).

2.7.4 Technique de redistribution

Quelque-soit la technique de décompte utilisée, il faut ensuite redistribuer la masse de probabilité extraite sur les événements non observés. Deux techniques sont majoritairement utilisées.

La première est une interpolation linéaire entre la distribution qui a subi le décompte et une distribution d'ordre inférieur qui peut être elle aussi lissée.

La seconde technique est le repli. Il s'agit d'utiliser la probabilité issue du décompte lorsque le n-gramme évalué a été observé dans le corpus d'entraînement et de se replier vers une distribution d'ordre inférieur lorsque le n-gramme est inconnu.

La grande différence entre ces deux approches est que l'interpolation inclue systématiquement la distribution d'ordre inférieure dans le calcul de probabilité alors que l'autre ne l'utilise qu'en cas de nécessité.

Interpolation linéaire

L'interpolation linéaire a été l'une des premières techniques de lissage utilisées. Elle s'appuye sur l'interpolation de modèles proposée par Jelinek et Mercer (1980). Cette approche consistait à interpoler systématiquement le modèle d'ordre n avec le modèle d'ordre $n-1$. Il faut noter que cette approche n'utilise pas de méthode de décompte. La formulation de cette interpolation est fournie par Brown et al. (1992b) :

$$P_{lissee}(w_i|h_i^n) = \lambda_{h_i^n} P(w_i|h_i^n) + (1 - \lambda_{h_i^n}) P_{lissee}(w_i|h_i^{n-1}) \tag{2.16}$$

Le modèle d'ordre inférieur avec lequel est interpolé le modèle initial est lui aussi lissé grâce à la même technique d'interpolation. On interpole généralement le dernier modèle de cette récursion avec une distribution uniforme $P(w_i) = \frac{1}{|V|}$.

L'ensemble des coefficients d'interpolation $\lambda_{h_i^n}$ peuvent être estimés pour maximiser la probabilité d'un corpus de développement en utilisant l'algorithme de Baum-Welch. Cependant, il semble qu'estimer une valeur différente de λ pour chaque historique ne soit pas très pertinent, de même que lui affecter une valeur constante quelque soit l'historique (Ristad, 1998). Par contre, regrouper les historiques en classes selon leur fréquence et estimer les coefficients λ pour chaque classe, comme le proposent Bahl et al. (1983), semble être la meilleure solution.

Une autre approche de l'interpolation est de l'utiliser pour redistribuer la masse de probabilités décomptés lors de l'estimation du modèle initial. C'est par exemple ce que proposent Ney et Essen. Un tel modèle peut être formulé de manière suivante :

$$P_{lissee}(w_i|h_i^n) = P_{decompte}(w_i|h_i^n) + \alpha_{h_i^n} P_{lissee}(w_i|h_i^{n-1}) \tag{2.17}$$

où $\alpha_{h_i^n}$ est le facteur de normalisation et de redistribution en fonction de l'historique.

Backoff

Une autre approche qui a été proposée par Katz (1987) consiste à ne se servir d'un modèle d'ordre inférieur que lorsque le n-gramme dont on cherche la probabilité n'a pas été observé dans le corpus d'entraînement. Cette technique est couramment appelée *backoff* et peut être formulée ainsi :

$$P_{lissee}(w_i|h_i^n) = \begin{cases} P_{decompte}(w_i|h_i^n) & \text{si } tf(h_i^n, w_i) > 0 \\ \alpha_{h_i^n} P_{lissee}(w_i|h_i^{n-1}) & \text{sinon} \end{cases} \tag{2.18}$$

où $\alpha_{h_i^n}$ est un coefficient de normalisation et de redistribution de la masse de probabilité décomptée à la distribution initiale.

L'avantage d'une telle approche est que la distribution sur laquelle le repli est effectué peut être optimisée pour les événements non observés étant donné qu'elle n'est pas interpolée avec les événements observés.

2.7.5 Lissage de Kneser-Ney Modifié

La technique de lissage qui semble la plus performante actuellement et qui est la plus largement utilisée est le lissage de Kneser-Ney Modifié.

Kneser et Ney (1995a) ont proposé une méthode de lissage qui repose sur une technique de décompte absolu et une redistribution par *backoff*. Chen et Goodman (1999) ont ensuite proposé une modification de la technique de décompte de cette approche pour traiter le cas particulier des n-grammes peu observés. Cette approche est ainsi couramment appelée lissage de Kneser-Ney Modifié.

Chen et Goodman (1999) ont montré que cette approche offre les meilleurs performances parmi toutes celles qu'ils ont testé.

2.8 Comparaison des modèles

Il est indispensable de disposer d'une manière de comparer les modèles de langage pour pouvoir les exploiter pleinement. L'évaluation peut se faire en comparant les taux d'erreurs mots qu'ils permettent d'obtenir sur une tâche de RAP, mais cette tâche est longue et multifactorielle.

Étant donné que le but du modèle de langage est de déterminer la probabilité de séquences de mots, on peut mesurer l'adéquation d'un tel modèle avec un corpus en mesurant la vraisemblance de celui-ci avec le modèle. Pour un modèle M et un corpus C, elle est donnée par la formule :

$$V(C|M) = \frac{1}{N} \sum_{i=1}^{N} log_2 P_M(w_i|h_i) \tag{2.19}$$

avec $P_M(w_i|h_i)$ la probabilité du mot w_i sachant sont historique h_i.

On calcule généralement la perplexité, qui peut être interprétée comme le facteur de branchement moyen du langage avec le modèle :

$$ppl_M(C) = 2^{V(C|M)} \tag{2.20}$$

Plus la perplexité est grande, moins le modèle de langage contient d'informations. Une distribution uniforme à une perplexité égale à la taille du vocabulaire.

Cette manière de comparer les modèles est rapide, c'est pour cette raison qu'elle est très répandue. Il existe cependant certaines restrictions à sont utilisation. La valeur absolue de la perplexité n'a que peu de valeur. Afin de pouvoir être comparées, elles doivent avoir été calculées sur le même corpus. De plus, bien qu'une certaine corrélation existe entre la perplexité et le taux d'erreur mots, une amélioration de la perplexité n'implique pas nécessairement une réduction du taux d'erreur mots, comme on peut le voir dans les travaux de Clarkson et Robinson (1999). Ce phénomène est notamment dû au fait que le processus de reconnaissance automatique de la parole fait intervenir plusieurs modèles.

2.9 Combinaison de sources d'informations

La combinaison de sources d'informations est une opération fréquente en modélisation du langage. Elle consiste à utiliser les informations provenant de différentes sources textuelles pour créer un modèle de langage unique.

Par exemple, lorsque l'on dispose de plusieurs corpus d'entraînement plus ou moins adaptés à la tâche pour laquelle on souhaite un modèle de langage, il est nécessaire de combiner les informations qu'ils contiennent afin que les plus pertinentes aient un poids plus élevé dans le modèle de langage final. Cette étape permet notamment de pondérer la contribution de chaque source dans le modèle final.

Il existe deux principales approches pour réaliser cette opération. La première consiste à estimer un modèle de langage indépendant pour chaque source d'information et à les combiner ensuite pour obtenir le modèle final. La combinaison est généralement une interpolation linéaire. La seconde approche consiste à estimer un modèle directement à partir des diverses sources d'information en les représentant sous la forme de contraintes. Le modèle est alors optimisé pour maximiser un critère, qui est souvent le maximum d'entropie.

2.9.1 Interpolation de modèles de langage

Il est en pratique assez rare de ne disposer que d'un corpus d'entraînement pour l'estimation d'un modèle de langage pour une tâche donnée. On dispose en général d'un corpus générique de grande taille et d'un corpus spécifique à la tâche de plus petite taille.

Dans une telle situation, la technique la plus simple et la plus rapide pour obtenir un modèle unique tenant compte des informations présentes dans tous les corpus est d'interpoler des modèles estimés sur chacun des corpus. L'interpolation la plus utilisée reste l'interpolation linéaire.

Si l'on considère un ensemble de M modèles de langage représentés par leur distribution P_m avec $m = 1, \ldots, M$, le modèle résultat de leur interpolation est définit comme :

$$P_{interpol}(w_i|h_i^n) = \sum_{m=1}^{M} \lambda_m P_m(w_i|h_i^n), \quad \text{avec} \sum_{m=1}^{M} \lambda_m = 1 \tag{2.21}$$

Les coefficients λ_m servent à pondérer l'importance de chaque modèle dans le modèle final. L'estimation de ces coefficients peut être effectuée pour maximiser un critère grâce à l'algorithme *Expectation Maximization* de Dempster et al. (1977). Le critère le plus utilisé reste le maximum de vraisemblance, obtenu en minimisant la perplexité du modèle sur un corpus représentatif des données pour lesquelles le modèle est estimé.

2.9.2 Estimation conjointe

Lorsque l'on dispose de plusieurs sources d'informations, il est également possible de construire directement un modèle de langage prenant en compte toutes ces informations. Chaque source est alors représentée par un ensemble de contraintes que la distribution finale doit respecter. L'intersection de toutes ces contraintes constitue alors un ensemble de distributions possibles. En l'absence de toute autre information, le choix de la distribution à considérer pour le modèle final doit se faire de la manière la plus neutre possible.

Dans le domaine des statistiques, ne pas prendre de décision revient à choisir le modèle dont la distribution est la plus uniforme. Plus une distribution est uniforme, plus son entropie est élevée. Pour trouver la distribution la plus uniforme parmi celles se trouvant à l'intersection des contraintes, il suffit de chercher celle dont l'entropie est maximale. L'algorithme *Generalized Iterative Scaling* (GIS) proposé par Darroch et Ratcliff (1972) est en général utilisé pour l'estimation de tels modèles.

Les travaux de Berger et al. (1996) présentent cette approche pour la modélisation du langage dans le cadre de la traduction automatique. Della Pietra et al. (1992) proposent une approche équivalente appliquée à la reconnaissance automatique de la parole. Cette dernière est une adaptation *Minimum Discriminant Information* (MDI), décrite à la section 2.10.3, qui est équivalente à l'approche par maximum d'entropie.

2.10 Adaptation des modèles de langage

Un modèle de langage est généralement estimé à partir d'un corpus d'entraînement représentatif de la tâche pour laquelle le modèle sera utilisé. Pour mesurer les performances du système, on utilise un corpus de test qui est en fait un ensemble de documents qui proviennent directement de cette tâche et pour lesquels on possède une référence. Il est parfois utilisé un corpus de développement, très représentatif de la tâche, qui permet d'optimiser certains paramètres du modèle de langage.

Dans certaines situations, le modèle de langage ne correspond pas parfaitement à la tâche applicative. Cela peut être parce que la quantité de données disponibles initialement pour construire le modèle est insuffisante ou que les données sont trop éloignées des conditions de test. Il peut également arriver que les données à traiter dans la tâche aient évoluées. Il arrive aussi que les données à traiter soient hétérogènes du point de vue linguistique (changement de thématique ou de style fréquents) et donc un modèle de langage statique est inapproprié. Il est alors nécessaire d'adapter le modèle à ces situations.

On peut dégager de ces situations trois configurations caractéristiques :

1. La tâche est multi-domaines, les conditions linguistiques changent donc brutalement en fonction des données à traiter.
2. Les données d'entraînement sont éloignées des données de test et on ne dispose pas de données d'adaptation
3. Les données d'entraînement sont éloignées des données de test mais on dispose de nouvelles données représentatives des conditions de test.

Les modèles adaptatifs sont les mieux adaptés à la première situation. Pour les deux autres cas, il est nécessaire de réaliser une adaptation explicite du modèle de langage initial. Cette adaptation sera supervisée ou non supervisée suivant la disponibilité de données représentatives des données de test.

2.10.1 Les modèles adaptatifs

Les modèles adaptatifs sont une famille de modèles qui s'adaptent dynamiquement au contenu linguistique des documents traités. En reconnaissance automatique de la parole, ces modèles modifient dynamiquement le score linguistique accordé aux mots en fonction de ce qui a déjà été reconnu.

Les gains en terme de taux d'erreur mots de ce type de modèles sont en général marginaux. L'utilisation de corpus d'entraînement de grandes tailles et l'amélioration des techniques permettant de mieux en tirer parti ont compensés les gains que permettent ces modèles.

Modèles *cache* et *trigger*

Les modèle *cache* et *trigger* peuvent être vu comme un modèle adaptatif dans la mesure où ils modifient dynamiquement les probabilités d'occurrence des mots en fonction de ceux qui ont déjà été traités. Le fonctionnement de ces modèles a été décrit respectivement aux sections 2.5.1 et 2.5.1.

Mélange dynamique de modèles

Le mélange dynamique de modèles a été proposée par Gotoh et Renals (1999) et repose sur un principe semblable à celui du modèle *trigger*. L'idée est d'adapter dynamiquement un modèle de langage issu d'une interpolation linéaire en modifiant le poids de chaque sous-modèle dans le mélange.

Le corpus d'entraînement est divisé en un nombre arbitraire de sous-corpus linguistiquement homogènes. Il s'agit en général de thématiques. Le modèle final est obtenu par interpolation linéaire de ces modèles thématiques. Le caractère dynamique vient du fait que les coefficients d'interpolation sont estimés de manière à maximiser la vraisemblance d'une hypothèse du système de reconnaissance automatique de la parole.

2.10.2 Adaptation non supervisée

Lorsque les conditions de test sont différentes des conditions d'entraînement et que l'on ne dispose pas d'informations sur ces nouvelles conditions, il est toujours possible d'en extraire directement des données de test. En reconnaissance automatique de la parole, la principale source d'information pour l'adaptation non supervisée des modèles acoustiques ou linguistique est l'hypothèse produite par le système avec les modèles initiaux. Bien que contenant des erreurs dues au fait que les modèles initiaux soient partiellement inadaptés, cette hypothèse peut nous renseigner de manière souvent fiable sur le contenu thématique et linguistique des documents.

On trouve dans la littérature beaucoup d'approches permettant d'adapter le modèle de langage au contenu des documents de test de manière non supervisée. La plupart de ces approches sont présentées dans la section 4.4, nous ne présenterons donc ici qu'une vue d'ensemble de ces technique.

Ces approches peuvent être décomposées en 3 étapes distinctes :

1. Extraction d'information sur le contenu du document de test
2. Collecte de données proches de ce contenu
3. Intégration de ces nouvelles données dans le modèle initial

Extraction d'informations

Pour réaliser la première étape, l'utilisation de l'hypothèse produite par le système de reconnaissance automatique de la parole est proposée quasi-exclusivement. Le problème principal qui se pose est qu'elle est partiellement erronée. Pour pallier ce problème il est en général pratiqué une méthode d'extraction de mots-clés issu du domaine de la recherche d'information. Ces mots-clés représentent alors le contenu linguistique des documents à transcrire mais contiennent un minimum d'erreur.

On peut notamment citer les travaux de Ito et al. (2009), qui utilisent une méthode de clustering pour regrouper les mots-clés. Etant donné qu'il y a une forte probabilité pour

que les erreurs qui surviennent dans la transcription produisent des mots en désaccord avec la thématique générale du segment, ils se retrouveront donc isolés.

Collecte de données

A partir des mots-clés extraits à l'étape précédente, il est possible d'utiliser les techniques de recherche d'information pour retrouver des données en rapport avec le contenu linguistique des documents à transcrire. C'est l'approche la plus courante.

Il existe d'autres approches plus originales et qui fonctionnent parfois mieux. On peut par exemple citer les travaux de Bigi et al. (2004) qui proposent de récupérer des documents dont la distribution de probabilités n-grammes est semblable à celle observée dans la première passe de reconnaissance automatique de la parole. La métrique de comparaison des distributions peut par exemple être la divergence de Kullback-Leibler (Kullback et Leibler, 1951).

Dans certains cas, la collecte de données peut se faire directement grâce à des connaissances *a priori* d'ordre générales sur la nature des documents à traiter. C'est le cas de la transcription de journaux radio et télédiffusés où il est raisonnable de considérer que, pour chaque émission, l'édition papier du jour relate les mêmes informations. Cette source d'information est appelée source parallèle. C'est, par exemple, de cette manière que Federico et Bertoldi (2004) collectent des documents d'adaptation.

Adaptation du modèle de langage

Une fois que des données représentatives des documents à traiter sont collectées, l'adaptation du modèle de langage peut se faire de manière supervisée. Cependant, l'approche qui est la plus répandue consiste à estimer un nouveau modèle de langage sur les données collectées, si leur quantité le permet, et de l'interpoler avec le modèle initial.

2.10.3 Adaptation supervisée

Si l'on possède des données plus proches des conditions de test que les données d'entraînement du modèle de langage, il est alors possible d'exploiter ces informations dans le cadre d'un processus d'adaptation supervisée. La nature et la quantité de ces données va conditionner la méthode d'adaptation à utiliser.

Adaptation par Information de Discrimination Minimale

La technique d'adaptation des modèles de langage par information de discrimination minimale (*Minimal Discrimination Information*, MDI) a été proposée par Rao et al. (1995) puis améliorée par Federico (1999).

Cette approche consiste à estimer un modèle de langage $\hat{P}_A(w_i|h_i^n)$ sur les données d'adaptation, lequel contient habituellement moins d'observations que le modèle initial $P_B(w_i|h_i^n)$. Les contraintes apportées par $\hat{P}_A(w_i|h_i^n)$ sont alors intégrées au modèle initial tout en s'assurant que les information détenues par ce modèle sur les événements non observés dans le corpus d'adaptation soient préservées. Il s'agit alors de trouver le modèle $P_A(w_i|h_i^n)$ dont l'entropie relative avec le modèle $P_B(w_i|h_i^n)$ est minimum et qui intègre les contraintes imposées par $\hat{P}_A(w_i|h_i^n)$. Le nom de cette approche vient du fait que l'entropie relative est également appelée divergence de Kullback-Leibler (Kullback et Leibler, 1951) ou *discriminant information* (Cover et al., 1991).

Cette approche est donc particulièrement adaptée aux situations dans lesquelles peu de données d'adaptation sont diponibles, comme par exemple lors des travaux de Kneser et al. (1997) et de Federico (1999) où le corpus d'adaptation ne représentait que quelques milliers de mots. Dans une telle situation, le seul modèle de langage qui puisse être estimé de manière fiable sur les données d'adaptation est un modèle unigramme : $\hat{P}_A(w)$. L'adaptation consiste alors à trouver le modèle $P_A(w_i|h_i^n)$ qui satisfasse le couple d'équations :

$$\begin{aligned} P_A(\cdot) &= \arg\min_{P(\cdot)} \sum_{h,w \in V} P(w,h) \log \frac{P(w,h)}{P_B(w,h)}, \\ \text{avec} \quad \hat{P}_A(w) &= \sum_h P_A(h) P_A(w|h) \quad \forall w \in V \end{aligned} \tag{2.22}$$

La solution de ce système d'équations peut être trouvé en utilisant l'algorithme *Generalized Iterative Scaling* proposé par Darroch et Ratcliff (1972).

Mélange statique de modèles

Cette approche a été introduite par Kneser et Steinbiss (1993) et améliorée plus tard par Clarkson et Robinson (1997). Elle est à la base de l'adaptation par mélange dynamique, c'est pour cette raison qu'elles sont très similaires.

L'idée est de découper le corpus d'entraînement en thématiques et d'estimer des modèles de langage indépendants pour chacune d'elles, exactement comme pour le mélange dynamique. La différence est que les coefficients d'interpolation de ces modèles sont estimés de manière à maximiser la vraisemblance des données d'adaptation et non plus l'hypothèse du système de reconnaissance automatique de la parole. C'est en ce sens qu'il s'agit d'un mélange statique puisque les coefficients n'évoluent pas en fonction de l'hypothèse.

Interpolation linéaire

Lorsque les données d'adaptation sont suffisantes pour estimer un ou plusieurs modèles de langage, il est alors possible de les interpoler avec le modèle de langage

initial. Les coefficients d'interpolation sont en général estimés pour maximiser la vraisemblance d'un morceau du corpus d'adaptation représentatif des données de test.

Deuxième partie

Étude du Web comme source de données en modélisation du langage

Chapitre 3

Le Web comme source de données

Sommaire

Le Web est aujourd'hui très développé et contient une masse de données textuelles très importante. Nous évaluons ici le potentiel du Web comme source d'informations dans le cadre de la modélisation linguistique en reconnaissance automatique de la parole.

3.1 Introduction

Comme nous l'avons vu au chapitre 2, les modèles de langages des systèmes de reconnaissance automatique de la parole récents sont de nature statistiques et sont estimés sur de gros volumes de données textuelles. Il est très important que ces corpus couvrent correctement le domaine pour lequel le système de reconnaissance automatique de la parole est conçu. Si les données linguistiques que doit traiter le système évoluent au cours du temps, il est nécessaire d'adapter les modèles initiaux avec un nouveau corpus qui correspond aux nouvelles conditions.

Plusieurs difficultés découlent de ces caractéristiques. La première est qu'il est nécessaire de constituer un corpus initial qui soit de grande taille et en accord avec la

tâche de reconnaissance automatique de la parole. La seconde est qu'il faut construire de nouveaux corpus (les corpus d'adaptation) à chaque fois que le contenu linguistique de la tâche considérée évolue.

S'il existait un corpus mis à jour régulièrement, qui soit de taille suffisante et qui contienne des données de tous les contextes linguistiques possibles, les problèmes de constitution de corpus cités précédemment seraient résolus. Nous allons étudier les caractéristiques du Web, notamment sa taille, son évolution dans le temps et la couverture qu'il offre pour différents contextes linguistiques, afin de déterminer s'il correspond à ces attentes.

3.2 Le Web

3.2.1 La taille du Web

Le Web connaît une croissance très importante depuis sa création. Afin d'estimer l'évolution de la quantité de données disponible sur le réseau, on peut observer l'évolution du nombre de noms de domaine pleinement qualifié réservés sur internet. Un nom de domaine pleinement qualifié sert en général à identifier un appareil connecté à un réseau, ici internet. Il est habituellement composé du nom d'hôte de l'appareil concaténé au nom du domaine auquel il appartient. Par exemple *www.exemple.org* est un nom de domaine pleinement qualifié identifiant le serveur *www* sur le domaine *exemple.com*. Chaque nom d'hôte sert à héberger au moins un site web. Il arrive que plusieurs sites Web soient hébergés sous un même nom d'hôte, ce qui fait que notre estimation de l'augmentation du Web sera probablement en dessous de la réalité. La figure 3.1 représente l'évolution du nombre de noms d'hôtes réservés depuis 1995. On constate que l'évolution du nombre de noms d'hôte sur le Web suit une croissance exponentielle, et on peut supposer que le nombre de documents présents sur le réseau suit probablement cette évolution.

3.2.2 Accéder aux données du Web

Savoir qu'il y a des données sur internet est intéressant, mais encore faut-il un moyen d'y accéder. Pour d'accéder à un document il faut connaître son adresse internet. Nous nous intéressons dans ces travaux aux données textuelles, il nous faudrait donc un moyen d'obtenir l'adresse d'un document en fonction de son contenu textuel. Pour cela il faudrait indexer une grande partie des documents du Web, ce qui nécessiterai des moyens colossaux. Heureusement, depuis un certain nombre d'années, des entreprises se sont spécialisées dans cette tâche et proposent des moteurs de recherche.

Un moteur de recherche est un logiciel permettant de trouver l'adresse de documents contenant une certaine information. Dans le cadre de documents textuels, il s'agit de mots. La quantité de données textuelles accessibles depuis les moteurs de recherche du Web à naturellement suivi la même évolution que le Web. Cela a été rendu possible

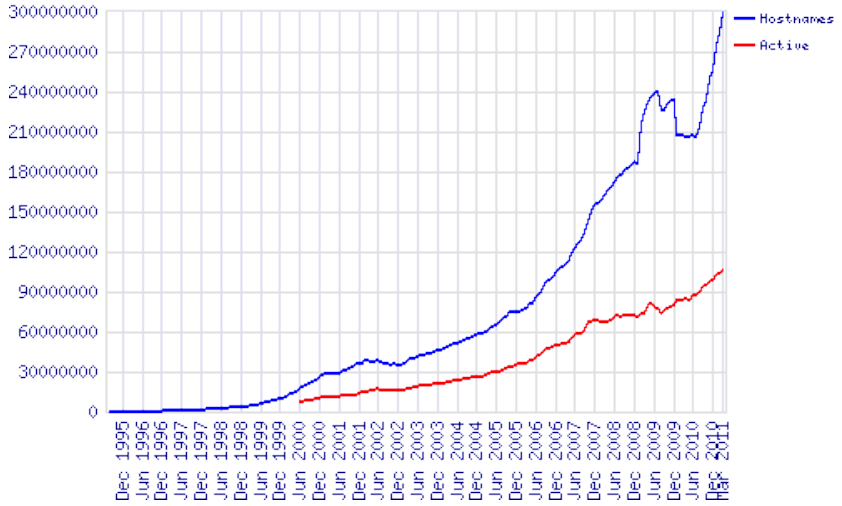

FIGURE 3.1 – *Evolution du nombre de noms d'hôtes réservés sur Internet entre 1995 et 2011, selon* `http://netcraft.com`.

par l'augmentation de la puissance de calcul et de la capacité de stockage des serveurs les hébergeant. Par exemple, le moteur de recherche Google avait en 1998 un index de 26 millions de pages, en 2000 un milliard de pages et, en 2005, il aurait franchit les 3 milliards de pages. Nous n'avons pas de chiffres officiels plus récents car ces données sont depuis considérées comme confidentielles par les sociétés proposant les moteurs de recherche.

3.3 Le Web pour la reconnaissance automatique de la parole

Nous avons vu que le Web est un réseau proposant une énorme quantité de données textuelles et qu'il existe un moyen d'y accéder facilement : les moteurs de recherche. Dans le cadre de la reconnaissance automatique de la parole, ces données peuvent être utiles pour la modélisation du langage. En effet, comme nous l'avons vu au chapitre 2, les approches qui fonctionnent le mieux sont des modèles statistiques basés sur des n-grammes de mots ou d'autres informations semblables (n-grammes d'étiquettes morphosyntaxiques, sacs de mots, etc.) et nécessitent de grandes quantités de données textuelles.

Dans ce cadre nous posons l'hypothèse fondamentale suivante : le Web est une source quasi-exhaustive de données textuelles, on peut y trouver toutes les séquences de mots correctes possibles, à une approximation près. Nous allons donc étudier le contenu textuel du Web dans cette perspective.

3.4 Mesurer la couverture du Web

Afin d'évaluer dans quelle mesure cette hypothèse est vraie, on peut mesurer la couverture du Web en terme n-grammes sur différents corpus. Ces résultats nous donneront également une idée de la pertinence du Web si on le considère comme une source de données textuelles classique, c'est à dire qu'on y récupère des fréquences de séquences de mots (les n-grammes). Afin de mesurer la couverture d'un corpus par un autre corpus, on peut utiliser la formule 3.1 :

$$Couv_n(C_a, C_b) = \frac{\sum_{i=n}^{|C_a|} \min(f^b(w_{i-n+1}^a, w_{i-n}^a, ..., w_i^a), 1)}{|C_a| - n + 1} \tag{3.1}$$

avec $Couv_n(C_a, C_b)$ la couverture du corpus C_a par le corpus C_b, comprise entre 0 et 1. $|C_a|$ est la taille du corpus C_a en nombre de mots. n est l'ordre des n-grammes pour laquelle la couverture est calculée. w_i^a est le mot i du du corpus C_a et la fonction $f^b(W)$ donne la fréquence de la séquence de mots (n-gramme) W dans le corpus C_b.

3.4.1 Mesurer la fréquence d'une séquence de mots sur le Web

La mesure de fréquence d'une séquence de mots sur le Web nécessite de pouvoir compter les occurrences de la séquence de mots dans tous les documents disponibles sur le Web. Afin d'effectuer cette opération de manière efficace, il faudrait construire un index de l'ensemble du Web. Cette tâche est très lourde et nécessite des ressources de calcul et de stockage énormes. Heureusement, certains organismes ont aussi besoin d'un index de l'ensemble du Web : les moteurs de recherche. Leur but est justement de rechercher des mots dans les documents textuels présents sur le Web. Nous proposons donc d'utiliser l'index de ces moteurs de recherche afin d'éviter de devoir construire le notre. Nous sommes cependant confrontés à deux problèmes :

1. Les termes utilisés dans les requêtes des moteurs de recherche sont initialement des mots et non des séquences de mots
2. L'index des moteurs de recherche est conçu pour retrouver des documents en fonction de leur contenu et non pour trouver des fréquences de mots ou de séquences de mots

Concernant le premier problème, la plupart des moteurs de recherches proposent maintenant des opérateurs permettant à l'utilisateur la recherche de phrases, qui sont en fait des séquences de mots. Techniquement il s'agit la plupart du temps d'encadrer la séquence de mots avec des guillemets. Nous pouvons utiliser cet opérateur en cherchant des séquences de mots n-grammes.

Le second problème est quant à lui plus délicat. Cependant, pour le calcul de la couverture lexicale, ce qui nous intéresse est de savoir si la séquence de mots est présente sur le Web ou non. On peut s'en assurer en mesurant une fréquence positive comme nous l'avons proposé dans la formule 3.1, mais cette mesure est compliquée à obtenir auprès des moteurs de recherche comme nous le verrons dans le chapitre 7. La seule

statistique que proposent certains moteurs de recherche est une estimation du nombre de documents correspondant à une requête. Si la requête est un n-gramme, alors ce nombre est une estimation de la fréquence de documents du n-gramme. Si nous mesurons ainsi une fréquence de document positive, alors c'est que le n-gramme existe sur le Web. Nous aboutissons donc à la formule 3.2 :

$$Couv_n(C_a, C_b) = \frac{\sum_{i=n}^{|C_a|} \min(df^b(w_{i-n+1}^a, w_{i-n}^a, ..., w_i^a), 1)}{|C_a| - n + 1} \tag{3.2}$$

avec $Couv_n(C_a, C_b)$ la couverture du corpus C_a par le corpus C_b, comprise entre 0 et 1. $|C_a|$ est la taille du corpus C_a en nombre de mots. n est l'ordre des n-grammes pour laquelle la couverture est calculée. w_i^a est le mot i du du corpus C_a et la fonction $df^b(W)$ donne la fréquence de la séquence de mots (n-gramme) W dans le corpus C_b.

On peut noter que l'utilisation d'un moteur de recherche ne permet pas de mesurer la couverture du Web mais d'une sous partie du Web, celle indexée par le moteur de recherche en question. La figure 3.2 représente le recouvrement qui existe entre les index des quatre plus gros moteurs de recherche Web actuels mesurée par Gulli et Signorini (2005). Nous avons, dans notre étude, utilisé les moteurs de recherche Yahoo [1] et Google [2] qui ont un taux de recouvrement d'environ 56%.

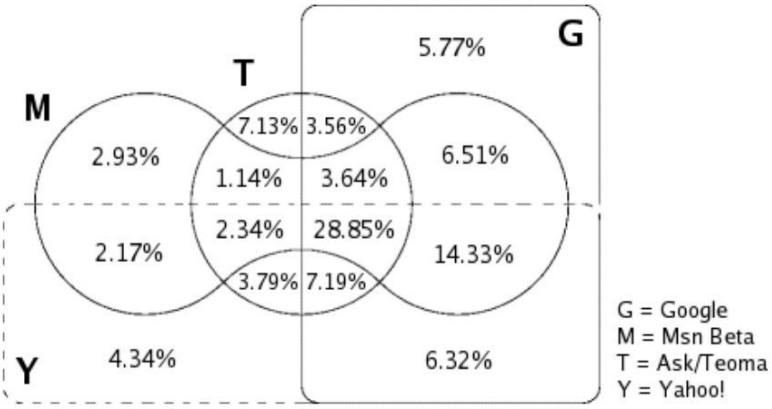

FIGURE 3.2 – *Recouvrement de l'index des principaux moteurs de recherche Web, mesuré par Gulli et Signorini (2005).*

3.4.2 Les facteurs qui influencent la couverture lexicale

La couverture lexicale d'une ressource peut être influencée par plusieurs facteurs comme par exemple la nature des données de test, les thématiques abordées, la langue, etc. Afin d'analyser plus précisément la couverture du Web, nous allons donc utiliser des corpus de différentes natures et de différents contenus.

1. http://www.yahoo.com
2. http://www.google.com

Tout d'abord, intéressons nous à la nature des documents. En effet, le Web contient beaucoup de données qui sont principalement de nature textuelle. On peut donc se poser la question de la pertinence de cette ressource lorsqu'il s'agit de l'utiliser pour faire du traitement de la langue naturelle orale. Nous allons mesurer la couverture du Web pour des corpus textuels comme des journaux mais aussi pour des corpus oraux comme des transcriptions de conversations.

Un second facteur qui peut influencer la couverture lexicale est l'ensemble des thématiques abordées dans les documents. Pour étudier le Web selon cet axe de variabilité nous allons utiliser des corpus dits *généralistes*, c'est à dire qui abordent des thématiques variées en employant un vocabulaire limité. C'est le cas par exemple des émissions d'informations (radio ou télévisées). Pour évaluer le comportement du Web sur cet axe nous allons aussi lui soumettre des corpus de spécialité, traitant d'une thématique particulière dans un domaine de spécialité, comme par un corpus du domaine de la médecine et traitant de chirurgie robotisée.

Le troisième facteur que nous allons étudier est la langue. En effet, comme on peut le voir sur la figure 3.3, les ressources disponibles sur le Web sont en quantités très inégales suivant les langues. Plus le Web contient de documents dans une langue, plus sa couverture est bonne sur des documents écrits dans celle-ci. Il est par exemple probable que la couverture linguistique du Web pour des données en anglais soit bien plus importante que pour des documents en Français car les documents en langue anglaise sont présents en quantité très largement supérieure sur le Web (57% du total) par rapport à ceux en Français (6% du total).

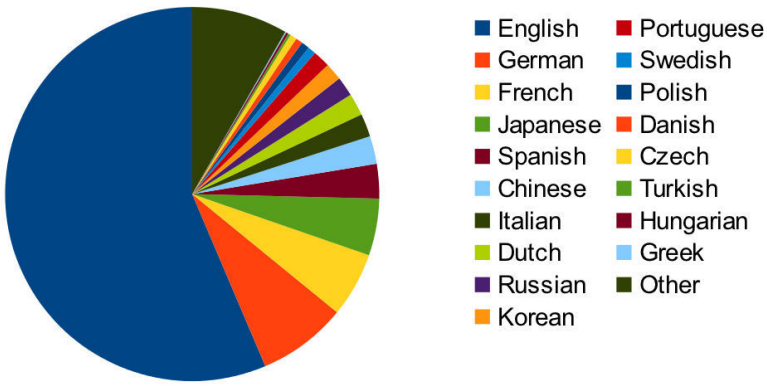

FIGURE 3.3 – *Répartition des langues sur le Web en 2002*

3.4.3 Les corpus

Pour chacune des situations qui découlent du produit cartésien des trois facteurs principaux qui influencent la couverture lexicale, nous avons choisi un corpus correspondant duquel sera extrait un ensemble de phrases qui serviront à mesurer la cou-

verture du Web dans ces situations.

Pour la langue anglaise nous avons utilisé les corpus suivants :
- Contenu généraliste
  - Texte : le corpus North American News Text Corpus (NANC) qui est un recueil de journaux d'information distribués principalement aux états unis (Graff, 1995)
  - Parole : le corpus HUB498 qui contient des transcriptions de journaux télévisés et radiodiffusés des états unis (Stern, 1997)
- Contenu de spécialité
  - Texte : les documents techniques du corpus AVISON (AVISON-tech)
  - Parole : les transcriptions des vidéos didactiques d'interventions chirurgicales du corpus AVISON (AVISON-video)

Pour la langue française nous avons utilisé les corpus suivants :
- Contenu généraliste
  - Texte : le corpus du journal Le Monde [3] (de 1987 à 2007) et Le Soir [4]
  - Parole : les transcriptions d'émissions radiodiffusées françaises du corpus ESTER (Gravier et al., 2004)

3.4.4 Résultats

Le tableau 3.1 contient les résultats de la mesure de couverture n-gramme décrite par la formule 3.1 appliquée au Web sur les corpus correspondants aux situations énoncées précédemment. Etant donné qu'il est difficile d'effectuer un gros volume de requêtes auprès des moteurs de recherche, nous avons construit un sous corpus pour chaque corpus à tester. Les sous-corpus contiennent tous environ 50k mots et sont constitués de phrases choisies aléatoirement.

Afin de mieux apprécier les résultats obtenus par, nous avons confrontés les résultats des calculs de couverture obtenus sur le Web, à ceux obtenus sur les corpus d'entraînement d'où sont extraits les corpus testés. Pour que cette seconde mesure de couverture soit la plus proche possible de ce qui est obtenu avec le Web, nous n'avons pas normalisé la taille des corpus d'entraînement. La mesure s'appuie donc sur une quantité de données qui varie d'une situation à l'autre. Par exemple les corpus de spécialité contiennent moins de données que les corpus généralistes, comme c'est d'ailleurs probablement le cas sur le Web. Le tableau 3.2 présente les tailles des différents corpus en fonction des contextes.

Contrairement aux corpus correspondants aux situations proposées, le Web est suffisamment volumineux pour contenir des documents relatifs à chacune. De ce fait, les documents relatifs à une situation peuvent contribuer à améliorer la couverture de n-grammes relatifs à une autre situation. Afin de reproduire ce comportement lors de l'expérience de couverture sur corpus, nous avons fusionné tous les corpus afin que

3. http://www.lemonde.fr
4. http://www.lesoir.be

chacun puisse contribuer à la couverture des autres situations, comme c'est le cas sur le Web.

De plus, le Web est une ressource en perpétuelle évolution. Son contenu change à chaque instant. Cette caractéristique est bénéfique lorsque l'on recherche des séquences de mots apparues récemment dans le langage ou dont la fréquence est modifiée en raison d'un événement. Dans une telle situation, le Web s'enrichira de documents représentatifs de ces nouvelles distributions contrairement aux corpus statiques.

L'inconvénient du caractère dynamique du Web est qu'il n'y a aucune garantie qu'une mesure effectuée à un instant donné soit identique a la même mesure effectuée à un autre instant. Cela pose notamment des problèmes de normalisation des mesures pour respecter le cadre dans lequel elles évoluent. Ce problème reste théorique et dans la plupart des cas, cela n'a aucun impact sur l'utilisation de la mesure. On notera également que dans un laps de temps court, le Web peut être considéré comme stable.

Les résultats de la mesure de couverture sur corpus sont reportés dans le tableau 3.3.

	Anglais				**Francais**	
	Généraliste		**Spécialisé**		**Généraliste**	
n	**Texte**	**Parole**	**Texte**	**Parole**	**Texte**	**Parole**
1	100.0	100.0	100.0	99.9	99.9	99.9
2	99.4	99.8	98.4	99.6	98.2	99.2
3	92.0	94.8	81.9	92.4	87.4	91.4
4	70.5	76.4	56.8	69.4	65.2	71.3
5	44.4	48.5	31.5	38.7	41.3	45.8
6	25.8	24.9	14.8	16.0	25.4	27.1

TABLE *3.1* – *Couverture n-gramme du Web pour différents corpus, en utilisant le moteur de recherche Yahoo*

	Anglais				**Francais**	
	Généraliste		**Spécialisé**		**Généraliste**	
	Texte	**Parole**	**Texte**	**Parole**	**Texte**	**Parole**
Corpus	NANC	HUB4-98	AVISON-tech	AVISON-video	LM+LS	ESTER
Taille	180.0	0.9	0.3	0.6	72.1	3.8

TABLE *3.2* – *Taille des corpus utilisés pour mesurer la couverture de référence, en millions de mots.*

3.4.5 Discussion

En observant ces résultats on s'apperçoit que la couverture du Web sur les corpus proposés est globalement bien meilleure que la couverture obtenue avec les corpus initiaux. Par exemple la couverture des bigrammes est quasiment de 100% pour le Web dans toutes les situations alors qu'elle est entre 70% et 90% sur les corpus. Pour les

	Anglais				Francais	
	Généraliste		Spécialisé		Généraliste	
n	Texte	Parole	Texte	Parole	Texte	Parole
1	99.2	99.8	99.3	96.6	99.4	99.8
2	80.4	86.1	77.1	66.8	87.6	91.2
3	41.5	49.2	42.0	29.8	56.1	61.6
4	15.5	23.8	18.6	8.2	27.0	30.0
5	5.8	15.5	8.7	1.8	12.3	13.6
6	2.5	13.3	4.6	0.3	6.3	6.7

TABLE 3.3 – *Couverture n-gramme de différents corpus avec eux même*

trigrammes la couverture Web est en moyenne deux fois plus élevée que sur les corpus (entre 80% et 90% pour le Web et de 30% à 60% pour les corpus). A partir de l'ordre 4 le Web offre une couverture de 5 à 10 fois supérieure à celle des corpus. Pour l'ordre 6 la couverture des corpus est quasi inexistante alors que le Web obtient une couverture d'environ 25% dans quasiment toutes les situations.

Ces chiffres montrent aussi que la couverture est globalement meilleure dans les situations de parole que dans les situations de texte. Ce résultat est d'autant plus étonnant que les corpus de parole utilisés dans ces expériences sont bien moins volumineux que ceux de texte, comme par exemple pour l'anglais généraliste où le corpus texte fait 180 millions de mots alors que le corpus de parole fait a peine un million de mots. La seule exception à ce phénomène est la mesure de couverture corpus de l'anglais de spécialité (AVISON-tech et AVISON-video) : la mesure de couverture Web se comporte normalement, c'est à dire que le texte est moins bien couvert que la parole, alors que la mesure de couverture corpus fait ressortir l'inverse. Nous pensons que cela est dû à la nature du corpus AVISON-tech, qui est un recueil de rapport techniques dont la forme est très codifiée et donc induit une réutilisation fréquente des même constructions linguistiques. Globalement, comme le laisse supposer Biber (1991), ces résultats suggèrent que les structures linguistiques employées dans les transcriptions des corpus oraux dont nous disposons sont moins diversifiées que celles des corpus écrit.

Les données du tableau 3.1 montrent que la couverture Web de données anglaises et françaises sont très proches pour des contextes équivalents (domaine et nature des données). La couverture de données en anglais est très légèrement supérieure que celle de données en français, ce qui s'expliquer par le faut que le Web contient un volume de données en anglais bien plus important que de données en français. Nous avons montré dans le graphique 3.3 que le français ne représente qu'environ 6% des données Web alors que l'anglais représente presque 60%, donc dix fois plus. Il est intéressant de constater que ce déséquilibre très important dans de quantité de données disponible n'influence pas plus la couverture.

Les données de spécialité sont presque aussi bien couvertes par le Web que les données généralistes, ce qui montre que le Web est robuste à cette variabilité. Cela est probablement dû à l'énorme quantité de données qui y est disponible. En effet, l'évolution de la couverture en fonction de la taille du corpus sur lequel on la mesure suit

une loi logarithmique : au dessus d'un certain volume de données, l'ajout de documents n'améliore plus la couverture. On peut alors supposer que la quantité de données disponible sur le Web dans ce domaine, bien que probablement très inférieure à la quantité de données généraliste, soit suffisante pour obtenir des scores de couverture finalement assez proche de la couverture du domaine généraliste. Le Web a atteint une taille critique où même les données de spécialité sont suffisamment représentées pour permettre d'obtenir une bonne couverture.

3.5 Conclusion du chapitre

Comme nous l'avons vu, quelque soit la situation, la couverture du Web en terme de n-gramme est bien meilleure que celle des corpus d'entraînement des modèles de langage. Nous avons vu que le Web contient environ 100% des unigrammes et des bigrammes testés et 90% des trigrammes pour tous les corpus testés, ce qui valide l'hypothèse que le Web contient une très grande partie les séquences de mots possibles. Pour les ordres de n-grammes plus élevés, le Web est moins exhaustif mais toujours de 5 à 10 fois plus que les corpus classiques.

Comme nous l'avons vu, le Web est un corpus intéressant de par sa taille et son évolution. Afin de pouvoir pleinement tirer parti de ces caractéristiques, il faut développer des méthodes d'exploitation qui les préservent. Une solution qui consisterait à télécharger tout ou partie des documents Web pour les utiliser briserait le caractère évolutif du Web. C'est pour cette raison que nous étudierons dans ces travaux l'utilisation d'estimateurs dynamiques, qui permettent de tenir compte de l'intégralité des données Web tout en préservant son évolutivité. Ces estimateurs font appel aux statistiques des moteurs de recherche, qui sont mises à jour de manière continue.

Nous allons étudier la manière dont nous pouvons exploiter ces données dans le cadre de la modélisation du langage en reconnaissance automatique de la parole. Nous allons explorer les deux pistes principales d'amélioration du modèle de langage : la constitution du lexique et l'estimation des scores linguistiques.

Troisième partie

Adaptation automatique du lexique

Chapitre 4

Etat de l'art : adaptation automatique du lexique

Sommaire

Dans un premier temps nous présenterons en quoi consiste l'adaptation automatique du lexique dans le cadre de la reconnaissance automatique de la parole et quels sont les problèmes sous-jacents. Nous ferons ensuite l'inventaire des approches que l'on trouve dans la littérature. Nous conclurons par une synthèse de ses approches.

4.1 Décalage entre lexique et données à transcrire

Comme nous l'avons vu, les modèles de langage statistiques sont entraînés *a priori* sur de gros corpus textuels. Les performances d'un tel modèle sont ensuite évaluées sur un corpus de test, plus petit, qui correspond aux conditions d'utilisation réelles du modèle de langage. Cette approche permet de s'assurer que le modèle est suffisamment robuste pour pouvoir traiter des données différentes de celles avec lesquelles il a été estimé. Plus ces deux corpus sont distants, en terme de couverture lexicale, syntaxique ou thématique, moins le modèle de langage sera performant.

L'adéquation du lexique du modèle de langage aux conditions d'utilisation, simulées par le corpus de test, peut se mesurer en comptant le nombre d'occurrences de mots du corpus de test qui ne sont pas dans le lexique. Ces mots sont appelés mots hors-vocabulaires. Le rapport entre le nombre de mots hors-vocabulaires et le nombre de mots dans le corpus de test est le taux de mots hors-vocabulaires, exprimé en pourcentage. Cette mesure est classiquement utilisée pour évaluer l'adéquation entre un lexique et un corpus.

La présence de mots hors-vocabulaires dans les documents à transcrire a des conséquences directes sur les performances du système de reconnaissance automatique de la parole, mais a aussi des conséquences indirectes sur les performances des systèmes qui s'appuient sur le processus de transcription.

4.1.1 Conséquences directes

Les hypothèses produites par le système pendant le décodage ne peuvent pas contenir de mots qui ne sont pas dans le vocabulaire. En conséquence, lorsqu'un tel mot survient, il se produit généralement une cassure dans la chaîne de probabilités n-grammes, qui a pour effet de polluer les scores linguistiques aux alentours du mot. C'est pour cela qu'on observe une dégradation des performances plus importantes de la transcription autour des mots hors-vocabulaires. On estime que chaque mot hors-vocabulaire engendre environ deux erreurs de transcription (Bazzi, 2002).

Le taux de mots hors-vocabulaires affecte donc directement et considérablement les performances des systèmes de reconnaissance automatique de la parole.

4.1.2 Conséquences indirectes

Comme nous l'avons vu, les mots hors-vocabulaires ne peuvent pas être transcrits par le système de reconnaissance automatique de la parole. Ce phénomène est gênant car il dégrade les performances du système et il peut aussi impacter les performances des systèmes en aval qui en dépendent.

On voit aujourd'hui se développer beaucoup de projets dont le but est la collecte massive de documents audio et vidéo pour archivage. Il doit être possible d'accéder

rapidement à des documents de ces bases suivant leur contenu. La voie privilégiée pour effectuer cette opération est d'utiliser un moteur de recherche texte. Actuellement la plupart des systèmes en production utilisent des meta-données associées aux documents, mais on voit émerger dans la littérature de plus en plus de travaux proposant d'utiliser la reconnaissance automatique de la parole pour extraire le contenu textuel des documents.

Dans un tel contexte, chaque mot que le moteur de reconnaissance automatique de la parole ne transcrira pas correctement ne pourra pas être utilisé pour l'indexation des documents. Ainsi les mots hors-vocabulaires ne pourront pas être utilisés pour retrouver les documents dans lesquels ils apparaissent. Nous avons fait une rapide étude typologique des mots hors-vocabulaires présents dans le corpus de test d'ESTER avec comme lexique les 65k mots les plus fréquents du corpus d'entraînement. Les résultats de cette étude sont présentés dans le tableau 4.1. On constate que les trois quarts des mots hors-vocabulaires sont des entités nommées et que le reste est principalement constitué de termes techniques. Ces mots se rapportent directement au contenu thématique des documents et représentent 97% des mots hors-vocabulaires. Il apparaît donc essentiel de minimiser au maximum les erreurs sur ces mots car ils sont très important pour l'indexation des documents.

Type	Proportion
Entités nommées	73%
Termes techniques	24%
Autre	3%

TABLE 4.1 – *Répartition des mots hors-vocabulaires par catégorie de mot observés sur le corpus ESTER*

4.2 Importance du phénomène

Nous avons vu que les mots hors-vocabulaires ont un impact important sur le système de reconnaissance automatique de la parole et sur les systèmes se servant de leurs sorties. Cependant, suivant le contexte, il est plus ou moins facile de construire un système initial qui permet de maintenir un taux de mots hors-vocabulaires faible dans le temps. Nous allons voir deux phénomènes qui influencent le taux de mots hors-vocabulaires.

4.2.1 L'influence de la langue

Le premier paramètre qui rentre en jeu dans l'apparition des mots hors-vocabulaires est la langue. En effet, l'unité lexicale dans un système de reconnaissance automatique de la parole est en général le mot. Ce dernier est défini dans ce contexte comme une unité sémantique mise en évidence par un séparateur qui est matérialisé à l'écrit par un espace. La taille maximale du lexique de reconnaissance automatique de la parole est

limitée par les ressources computationnelles disponibles et donc plus une langue possède de mots et plus le système sera touché par le problème des mots hors-vocabulaires.

La morphologie et la syntaxe de la langue peuvent donc influencer le nombre de mots possibles. Par exemple, dans les langues flexionnelles, un lemme change de forme en fonction du contexte et de sa fonction grammaticale. Pour un ensemble de lemmes donnés, l'ensemble des mots possible qui en dérivent est bien plus important si la langue est flexionnelle.

Un autre paramètre de la langue qui peut influencer le nombre de mots possibles est la manière dont s'orthographient les mots composés. En effet, dans plusieurs langues, dont l'allemand, les mots composés s'écrivent attachés. Par exemple, le mot « pommier » est composé en allemand du mot « Baum » (« arbre ») et du mot « Apfel » (« pomme ») et s'écrit « Apfelbaum ». Cette manière d'orthographier les compositions fait que le nombre de mots d'une telle langue est énorme.

Les langues agglutinantes fonctionnent d'une manière similaire. Chaque mot résulte de la composition de plusieurs traits qui expriment chacun une idée élémentaire. Le nombre de mots qu'il est possible de construire augmente de manière exponentielle avec le nombre de traits disponibles et il est difficile de savoir à l'avance quelles sont les combinaisons impossibles.

En conséquence, comme le montrent Geutner et al. (1998), le taux de mots hors-vocabulaires est plus élevé pour les langues à forte composition de mots ou fortement flexionnelles. Il est ainsi plus difficile de réduire le taux de mots hors-vocabulaires en augmentant la taille du lexique de reconnaissance automatique de la parole.

4.2.2 L'influence du contenu

Le second facteur qui influence le taux de mots hors-vocabulaires est bien évidemment le contenu des documents. Dans certaines situations, le contenu des documents à transcrire évolue dans le temps et des nouveaux mots sont continuellement introduits. C'est le cas par exemple des domaines de pointe où de nouvelles expressions, de nouveaux mots apparaissent ou de nouvelles entités nommées apparaissent en fonction des découvertes.

On retrouve également ce phénomène dans les journaux d'information où il est difficile de prédire les thématiques abordées. Le taux de mots hors-vocabulaires est en général stable dans ce domaine, mais il arrive parfois qu'un événement médiatique survienne et qu'un mot auparavant peu fréquent et relatif à celui-ci soit tout à coup employé continuellement. C'est par exemple le cas du mot "tsunami", qui était peu présent dans les lexiques des moteurs de reconnaissance automatique de la parole de journaux avant la catastrophe de 2004, et qui depuis s'y trouve.

4.3 Détection des mots hors-vocabulaire

Une première approche pour limiter le problème des mots hors-vocabulaires à été de les détecter. En effet, la connaissance de la position des mots hors-vocabulaires dans les transcriptions issues d'un processus de reconnaissance automatique de la parole permet d'effectuer un traitement particulier avec les mots erronés qui ont été transcrits à leur place. Par exemple si la reconnaissance automatique de la parole est couplée à un système de RI, la connaissances des mots erronés dus à des mots hors-vocabulaires permet de ne pas les utiliser pour l'indexation des documents.

Nous allons présenter les trois familles d'approches que l'on trouve dans la littérature pour réaliser cette détection.

4.3.1 Détection par *fillers* acoustiques

Les premières approches développées se situaient essentiellement au niveau acoustique du processus de reconnaissance automatique de la parole. Le principe général d'un moteur de reconnaissance automatique de la parole est de trouver la suite de mots du lexique ayant le meilleur score linguistique et dont la représentation phonétique est la plus proche de celle observée dans le signal acoustique à transcrire. Les mots hors-vocabulaires n'étant pas dans le lexique, le moteur de reconnaissance automatique de la parole va donc les transcrire par des mots du lexique. Si les mots hors-vocabulaires ont une phonétisation assez éloignée des mots du lexique, la distance entre la suite de phonèmes observée dans le signal et les mots du lexique remplaçant les mots hors-vocabulaires sera plus grande que dans le reste de la transcription.

Ces approches s'appuyent donc sur le fait que les mots hors-vocabulaires ont en général une phonétisation raisonnablement éloignée des phonétisations des mots du lexique de reconnaissance automatique de la parole. Si le lexique comportait un mot qui pouvait être choisit par le moteur de reconnaissance automatique de la parole lorsque la distance phonétique entre le signal et les autres mots du lexique est trop grande, cela permettrait de détecter les mots hors-vocabulaires.

Un tel mot générique doit pouvoir être associée à n'importe quelle suite de phonèmes si l'on veut qu'il puisse être utilisé lors du processus de décodage. La méthode privilégiée consiste à utiliser un automate qui peut absorber toutes les suites de phonèmes possibles.

Par exemple (Asadi et al., 1990) proposent d'ajouter un mot au lexique qui puisse correspondre à n'importe quelle suite de phonème. Un inconvénient majeur de leur approche est que ce mot générique est souvent décodé à la place de mots du lexique. Pour limiter ce phénomène, (Bazzi et Glass, 2000) proposent un modèle similaire mais utilisent un modèle bigramme de phonèmes pour contraindre les phonétisations possibles du mot générique. Finalement, (Boulianne et Dumouchel, 2001) vont plus loin en faisant en sorte que le mot générique ne puisse modéliser aucun mot du lexique.

L'inconvénient majeur de cette approche est qu'elle est difficile à régler. En effet, si la probabilité du mot générique est trop élevée, il va avoir tendance à remplacer aussi les mots du lexique présents dans la transcription, en commettant éventuellement des erreurs phonétiques. Si par contre il a une probabilité trop faible, il va manquer des mots hors-vocabulaires.

Il est également possible de se passer d'un mot du lexique pour lequel le modèle acoustique peut absorber n'importe quelle suite de phonèmes. Il suffit pour cela d'agir sur le modèle de langage. Par exemple (Bisani et Ney, 2005) proposent de remplacer tous les mots hors-vocabulaires du corpus d'entraînement du modèle de langage par leur représentation phonétique. Le modèle de langage estimé sur le corpus ainsi transformé est donc hybride car il comporte des mots et des phonèmes. L'utilisation d'un tel modèle de langage dans un processus de reconnaissance automatique de la parole classique fait ressortir dans les hypothèses fournies par le moteur les mots hors-vocabulaires sous la forme de leur transcription phonétique. L'étalonnage des probabilités des phonèmes est automatiquement fait par le processus classique d'estimation des probabilités n-grammes du modèle de langage. Chaque phonème est considéré comme un mot.

Rastrow et al. (2009) proposent également un modèle de langage hybride combinant des mots et des sous-mots déterminés de manière statistiques. Lors du processus de décodage, ils mesurent la confusion qui existe entre le mot de la meilleure hypothèse et les sous-mots. Cette mesure est comparée à celle observée lors du décodage de ce même mot dans le corpus d'entraînement. Si la distance entre ces deux mesures est trop importante, le système étiquette le mot comme une erreur causée par un mot hors-vocabulaire.

Etant donné que de tels modèle de langage hybrides contiennent des unités différentes (mots et phonèmes), il est nécessaire d'augmenter l'ordre du modèle pour que les plus petites unités puissent être transcrites correctement. Quand l'ordre d'un modèle n-gramme est trop élevé pour une unité donné, les performances de la reconnaissance automatique de la parole sont en général moins bonnes. Etant donné que le modèle contient des unités de deux niveaux et que l'ordre du modèle s'applique indifféremment à toutes les unités, l'ordre ne pourra être optimal pour les deux unités.

L'approche par *filler* acoustique suppose que les mots hors-vocabulaires ont une phonétisation éloignée des mots du lexique, ce qui n'est pas toujours vrai.

Finalement, cette approche n'est pas très robuste au bruit acoustique. Par exemple, lorsque quelques phonèmes dans le signal sont très perturbés, modèle de langage mis à part, le moteur de reconnaissance automatique de la parole retrouve le mot du lexique correspondant, car aucun autre mot n'est associé à une suite phonétique suffisamment proche. Par contre le mot générique, de part sa nature polymorphe, pourrait avoir une probabilité plus élevée que le mot du lexique normalement choisi car il pourra associer les phonèmes de probabilité les plus élevées aux emplacement dégradés dans le signal, contrairement au mot du lexique qui a une phonétisation figée.

4.3.2 Détection par caractérisation de mesures

Une autre manière de détecter les mots hors-vocabulaire consiste à observer le comportement de mesures effectuer lors du décodage et d'observer des variations caractéristiques de l'apparition des mots hors-vocabulaires. Cette famille d'approche consiste à caractériser le comportement du processus de transcription lorsqu'un mot hors-vocabulaire est rencontré.

distorsion acoustique

On peut supposer qu'il existe une distance phonétique assez grande entre les mots hors-vocabulaires et les mots du lexique. Au lieu d'insérer un mot générique dans le lexique pour détecter les cas où aucun mot du lexique ne correspond au signal, il est possible de comparer la transcription phonétique du signal et la phonétisation de la transcription orthographique du même signal produite par le moteur de reconnaissance automatique de la parole.

La phonétisation de la transcription orthographique est contrainte par les mots du lexique alors que la transcription phonétique est non contrainte. Cette dernière est donc censée être la suite de phonèmes réellement présente dans le signal. Si les mots hors-vocabulaires ont une phonétisation suffisamment éloignée des mots du lexique, alors la distance entre la transcription phonétique et la phonétisation de la transcription orthographique devraient l'être aussi.

Cette mesure est en général calculée dans une fenêtre glissante sur la transcription. Les pics observés dans la mesure correspondent en théorie à des mots hors-vocabulaires.

Par exemple, Lin et al. (2007) proposent d'utiliser des modèles graphiques (Lauritzen, 1996) pour combiner et aligner un treillis de mots et un treillis de phonèmes comme expliqué par Ji et al. (2006). La détection des mots hors-vocabulaires consiste à chercher les endroits où l'alignement est mauvais.

White et al. (2008) proposent d'utiliser deux systèmes de reconnaissance automatique de la parole : un système classique dit "contraint" par un modèle de langage et un lexique et un système dit "peu contraint" sans modèle de langage ni lexique. Le premier système permet d'obtenir un décodage en mots alors que le système "peu contraint" permet un décodage en phonèmes. Ils proposent de comparer les sorties des deux systèmes au niveau des mots et des phonèmes. Pour obtenir des phonèmes à partir de la sortie en mots, ils utilisent le lexique de prononciations ayant servi au décodage. Pour obtenir des mots à partir de la sortie phonétique du système *peu contraint* ils utilisent un transducteur phonèmes vers mots. Ces deux mesures sont ensuite combinées à une mesure de confiance appelée *Cmax* et proposée par Wessel et al. (2001). La combinaison se fait par maximisation de l'entropie.

Burget et al. (2008) proposent d'utiliser les probabilités *a posteriori* d'un système de reconnaissance automatique de la parole *contraint* et *peu contraint* pour la détection des

mots hors-vocabulaires. Ces probabilités sont ensuite exploitées par un réseau de neurones pour faire la détection des mots hors-vocabulaires.

L'avantage de ces approches est qu'elles ne nécessitent pas de modifications profondes du moteur de reconnaissance automatique de la parole. En effet, la plupart des systèmes produisent déjà une transcription phonétique non contrainte sur laquelle est construit de graphe de mots.

mesures de confiances

Le principe des mesures de confiance en reconnaissance automatique de la parole est de fournir un score associé à chaque mot en sortie du processus de reconnaissance automatique de la parole. Ce score est une indication sur la certitude qu'a le système dans l'hypothèse. Une telle mesure est très utile lorsque les sorties du système de reconnaissance automatique de la parole est utilisé pour une autre tâche comme l'indexation. Le score de confiance peut permettre de ne réaliser l'indexation que sur les mots dont le système est sûr, ce qui évite d'introduire du bruit dans l'index. Cette approche a aussi été utilisée pour détecter les erreurs et tenter de les corriger.

On trouve dans la littérature plusieurs travaux qui s'appuient sur des méthodes similaires aux mesures de confiance pour détecter les mots hors-vocabulaires dans les transcriptions issues de reconnaissance automatique de la parole.

Le principe général de telles approches est d'extraire un maximum d'informations pertinentes du processus de reconnaissance automatique de la parole et d'utiliser un classifieur pour étiqueter les mots comme hors-vocabulaire à partir de ces informations. Toute la difficulté consiste ici à choisir les paramètres qui seront pertinents pour la détection des mots hors-vocabulaires.

Par exemple Sun et al. (2003) proposent un détecteur de mots hors-vocabulaires offrant un faible taux de fausses alarmes comparé à un modèle *filler* existant sur leur système de dialogue. Leur système extrait, pour chaque mot, un certain nombre de paramètres du processus de décodage, comme par exemple le nombre de mots en concurrence dans le graphe d'hypothèses ou les scores linguistiques et acoustiques du mot. Ils extraient également des paramètres relatifs au contexte du mot, comme les scores acoustiques des mots précédents et suivants. Ces paramètres sont ensuite utilisés pour détecter les mots hors-vocabulaires à l'aide d'une analyse discriminante linéaire.

Cai et Zhu (2004) proposent d'utiliser un classifieur adapté à la tâche de détection des mots hors-vocabulaires. Ils adaptent une machine à vecteurs de support (Hearst et al., 1998) afin qu'elle soit plus robuste et qu'elle généralise mieux car on dispose souvent de trop peu de corpus d'entraînement pour avoir suffisamment de mots hors-vocabulaires pour l'entraînement d'un tel classifieur. Les paramètres qu'ils utilisent dans ce classifieurs sont extraits du processus de reconnaissance automatique de la parole. Ce sont des paramètres classiques, comme les différences entre les N meilleurs hypothèses proposées par le système. Les résultats montrent que leur classifieur fournit de meilleures performances que des classifieurs classiques comme les machines à

vecteurs de support ou les réseaux de neurones.

Lecouteux et al. (2009) proposent d'utiliser un algorithme de boosting (Freund et Schapire, 1995) pour la détection des mots hors-vocabulaires. Pour chaque mot, des paramètres sont extraits du processus de reconnaissance automatique de la parole, comme la probabilité linguistique, la confusion dans le treillis de mots ou le score acoustique. Il est fourni au classifieur un vecteur pour chaque mot contenant les paramètres correspondant au mot et ceux correspondant aux mots voisins. Ce processus de détection est ensuite affiné à l'aide d'une analyse sémantique latente. Cette dernière passe a pour but de réduire le taux de fausses alarmes en remettant en cause les mots hors-vocabulaires détectés mais qui sont pourtant sémantiquement cohérents.

Les approches par mesure de confiance offrent de bonnes performances mais sont réservées à des situations où le corpus d'entraînement est conséquent. En effet, les mots hors-vocabulaires sont rares et les classifieurs utilisés pour les détecter à partir des mesures de confiance nécessitent beaucoup d'exemples. De plus, le corpus utilisé ne doit pas avoir servi à l'estimation du système de reconnaissance automatique de la parole car son comportement serait différent d'un cas nouveau.

4.3.3 Combinaison de techniques

On trouve plusieurs travaux proposant de combiner les différentes approches proposées précédemment.

Par exemple Hazen et Bazzi (2001) proposent de combiner la méthode de modélisation des mots hors-vocabulaires par modèle acoustique présentée par Bazzi et Glass (2000) et la technique de détection d'erreurs de reconnaissance automatique de la parole à l'aide de mesures de confiance décrite par Timothy J. Hazen et Seneff (2002). Les auteurs montrent que la combinaison de ces approches fournit de meilleurs résultats que la meilleure des approches seule.

Les performances des systèmes de détection de mots hors-vocabulaire sont actuellement autour de 10% d'égale erreur (fausse acceptation et faux rejet).

4.4 Adaptation automatique du lexique

Détecter les mots hors-vocabulaires n'est pas une solution suffisante dans beaucoup de cas car elle ne fournit aucun moyen pour récupérer le mot hors-vocabulaire. Cet aspect est problématique quand les mots hors-vocabulaires sont des mots pertinents pour une tâche en aval du processus de reconnaissance automatique de la parole, comme l'indexation des documents par exemple.

Comme nous l'avons vu, il est très important de maintenir un taux de mots hors-vocabulaires dans les documents à transcrire le plus bas possible. Si le lexique des documents à transcrire est relativement stable dans le temps, le lexique initial du moteur de reconnaissance automatique de la parole peut suffire pour maintenir un taux de mots

hors-vocabulaires bas. Dans le cas contraire, il est nécessaire d'effectuer une adaptation du lexique du moteur de reconnaissance automatique de la parole au cours du temps, idéalement de manière automatique.

L'adaptation du lexique d'un système de reconnaissance automatique de la parole peut être vu comme une succession de problématiques distinctes :

1. Collecter un corpus d'adaptation qui soit le plus proche possible du contenu des documents à transcrire
2. Sélectionner dans ce corpus les mots qui seront introduits dans le lexique à adapter
3. Associer une représentation phonétique à chacun des mots à introduire dans le lexique
4. Attribuer une probabilité linguistique à ces mots dans le modèle de langage.

Ces problématiques peuvent être vue comme indépendantes. Pour chacune, nous allons faire un bilan des travaux de la littérature qui s'y rapportent.

4.4.1 Choix d'une source d'information

La première difficulté consiste à choisir une source d'information adéquate pour réaliser l'adaptation du lexique. En général il va s'agir de sélectionner un ensemble de documents dans un corpus. Les documents en question doivent être le plus proche linguistiquement du contenu des documents à transcrire pour minimiser le décalage lexical entre ces deux ressources.

Utilisation de sources parallèles

La recherche de documents similaires pour constituer le corpus d'adaptation nécessite une information sur le contenu du document à transcrire. Dans certains cas on dispose de sources d'informations qui peuvent directement être mises en parallèle avec les document à transcrire. C'est par exemple le cas de la transcription de journaux d'information. Les journaux papiers constituent des documents textuels qui peuvent être mis en parallèle des journaux audio du même jour. Dans ce cas de figure, il n'est pas nécessaire d'avoir d'information sur les documents à transcrire, la simple collecte des journaux papiers du jour d'émission (ou des quelques jours avoisinant) constitue un corpus d'adaptation suffisant.

Cette technique a été exploitée pour la première fois pour les journaux d'information par Federico et Bertoldi (2004). La tâche qu'ils doivent accomplir est la transcription de journaux télévisés du jour. Pour constituer un corpus d'adaptation ils proposent de récupérer sur le Web les articles d'information publiés le jour même. Ils réussissent ainsi à constituer un corpus d'environ 60k mots par jour. Plus tard, cette technique sera reprise par Allauzen et Gauvain (2005a) dans un but similaire : la transcription des journaux radiodiffusés.

Récemment, Martins et al. (2010) ont proposé une amélioration de ce procédé. Ils partent du constat que l'ensemble des articles d'information collectés un jour ne conviennent pas nécessairement à toutes les thématiques abordées dans les documents d'information à transcrire de la journée. Ils proposent donc de construire un corpus d'adaptation pour chaque thématique apparaissant dans les documents à transcrire. De cette manière l'adaptation sera plus précise et donc moins bruitée. Comme dans l'approche précédente, ils collectent quotidiennement les articles de journaux sur des sites Web d'information. Le volume de données qu'ils collectent est de 80k mots par jour. A la différence des autres approches, où ces données sont souvent utilisées directement, ils proposent de les indexer à l'aide d'un moteur de recherche d'information. La constitution du corpus d'adaptation est ensuite réalisée en deux étapes. Ils effectuent une segmentation thématique des documents à transcrire et, pour chaque thématique, ils extraient les documents les plus pertinents de la base d'articles collectés quotidiennement grâce au moteur de recherche. Les documents récupérés pour chaque thématique constituent le corpus d'adaptation utilisé pour tous les segments des documents à transcrire relatifs à celle-ci.

Les travaux de Yu et al. (2000b) sont s'inscrivent dans la même perspective. La tâche qu'ils s'étaient fixé était la transcription d'enregistrements de réunions d'entreprise. L'information dont ils disposaient *a priori* sur le contenu des documents à transcrire est qu'ils concernaient les activités de l'entreprise. Ils ont donc considéré l'utilisation de l'ensemble des documents constituant le site Web de l'entreprise comme corpus d'adaptation. Ils ont montré que cette approche permettait d'obtenir des documents contenant une bonne partie des noms propres présents dans les documents à transcrire (nom d'employés, de produits, etc.).

Lorsque l'on ne possède pas de connaissance a priori sur les documents à transcrire ou que l'on ne dispose pas de sources d'information parallèles, il faut trouver un moyen d'obtenir de l'information sur le contenu linguistique du document. On trouve dans la littérature deux tendances. La première consiste à utiliser des informations associées aux documents à transcrire, ce que l'on appelle des meta-données. La seconde consiste à extraire directement du document son contenu linguistique par le biais d'une première passe de transcription.

Extraction d'information par le biais de la reconnaissance automatique de la parole

Le but de l'information extraite des documents audio est de servir à retrouver des documents textuels similaires pour construire le corpus d'adaptation. Il faut donc un moyen de trouver les documents correspondants à l'information extraite. C'est le but de la recherche d'information, un domaine étudié depuis le début de l'informatique. Généralement la base de documents dans laquelle on souhaite chercher des documents similaires est indexée par un moteur de recherche. Les informations extraites du document vont ensuite servir à formuler une requête qui sera soumise au moteur de recherche qui va trouver les documents les plus pertinents selon un critère spécifique (Manning et al., 2008). C'est pour cette raison que beaucoup de techniques d'extraction

d'information du flux de parole présentées ici sont issues du domaine de la recherche d'information ou sont conçues pour optimiser cette recherche.

Les premiers travaux que l'on trouve sur le sujet sont ceux de Berger et Miller (1998). Ils proposent d'effectuer une première passe de transcription des documents audio pour extraire des indices sur leur contenu linguistique. Ils filtrent cette transcription avec une liste de mots-outils pour obtenir une liste de mots-clés représentatifs du contenu thématique des documents. Ils soumettent ensuite ces mots-clés à un moteur de recherche Web et utilisent les documents résultat comme corpus d'adaptation.

A la même époque, Kemp et Waibel (1998) proposent une approche similaire appliquée à la transcription de journaux radio-diffusés. La première différence avec la précédente est qu'ils ne filtrent pas la première passe de transcription, ils utilisent donc tous les mots. La seconde est qu'ils ne cherchent pas des documents dans un moteur de recherche Web généraliste, mais dans un moteur de recherche qu'ils ont construit avec des documents collectés quotidiennement sur des sites Web d'information. Le critère de recherche qu'ils utilisent est la métrique Okapi proposée par Beaulieu et al. (1997). La maîtrise des documents indexés par le moteur de recherche permet de minimiser le bruit qui pourrait se glisser dans le corpus d'adaptation en évitant que le moteur de recherche ne retourne des documents trop éloignés. L'inconvénient d'une telle approche est qu'elle ne bénéficie pas du nombre considérable de documents indexés par les moteurs de recherche Web.

Par la suite, plusieurs aspects de cette famille d'approches ont été améliorés. Une des améliorations majeures est l'utilisation de techniques d'extraction de mots-clés de la première passe de transcription.

Utilisation de mots-clés issus de la reconnaissance automatique de la parole

L'utilisation de techniques d'extraction de mots-clés de la première passe de transcription permet de capturer plus précisément le contenu thématique des documents à transcrire. L'avantage d'une telle pratique est qu'elle permet de réduire le bruit introduit par les mots moins porteurs de sens ou sans rapport avec les thématiques principales du document.

On trouve notamment les travaux de Yu et al. (2000b) qui concernent la transcription d'enregistrements de réunion. Lles mots-clés sont extraits de la première passe de reconnaissance automatique de la parole grâce à l'information mutuelle. Cette métrique est souvent utilisée en classification thématique (Yang et Pedersen, 1997). Ces mots-clés servent ensuite à construire des requêtes soumises à un moteur de recherche Web et les documents résultat constituent le corpus d'adaptation. Ils montrent que cette approche offre de meilleures performances en terme de récupération de mots hors-vocabulaires que d'utiliser l'intégralité des documents constituant le site Web de l'entreprise auquel l'enregistrement de réunion est associé.

Par la suite, beaucoup de travaux ont utilisé ce principe d'extraction de mots-clés pour obtenir une représentation du contenu des documents à transcrire. Une raison

majeure est qu'à partir de mots-clés, il est très facile de trouver des documents similaires dans une base indexée par un moteur de recherche.

Cette technique a été améliorée par la suite, notamment en ce qui concerne la robustesse des mots-clés extraits de la première passe de reconnaissance automatique de la parole. En effet, il existe des erreurs dans la première transcription et si les mots erronés sont utilisés comme mots-clés pour construire le corpus d'adaptation, des documents non pertinents y seront introduit.

Robustesse aux erreurs de reconnaissance automatique de la parole

Un des principes les plus utilisés pour limiter l'impact des erreurs de reconnaissance automatique de la parole dans la construction du corpus d'adaptation est, sans doute, la cohérence thématique. L'idée de base est que les erreurs de reconnaissance automatique de la parole produisent des mots qui ne sont souvent pas cohérents avec la thématique du document. De plus il est très peu probable qu'un grand nombre d'erreurs aboutissent à des mots cohérents entre eux. Donc si l'on regroupait les mots porteurs de sens d'une transcription automatique homogène par thématique, nous obtiendrions une thématique qui rassemblerait presque tous les mots de la transcription et un ensemble de thématiques avec peu de mots, qui correspondraient probablement à des mots erronés.

L'approche développée par Ito et al. (2009) s'inscrit exactement dans cette perspective. Leur approche permet d'éliminer implicitement les mots-clés qui pourraient altérer la qualité du corpus d'adaptation, tout en conservant ceux qui pourraient l'améliorer. Dans un premier temps, ils extraient les mots-clés de la première passe de reconnaissance automatique de la parole du document à transcrire grâce à la métrique TF.IDF (Spärck Jones, 1972). Ces mots-clés sont ensuite regroupés par similarité, ce qui forme des groupes de mots-clés similaires. A ce stade, les mots-clés erronés se trouvent très probablement dans des groupes différents des bon mots-clés. Chaque groupe est soumis au moteur de recherche Web comme une requête et les documents résultats sont récupérés. Pour chaque groupe de mots-clés, la similarité entre l'ensemble des documents récupérés et la première passe de reconnaissance automatique de la parole du document a transcrire est mesurée. Si cette similarité est trop faible, tous les documents récupérés grâce au groupe de mots-clés sont écartés. Grâce à cette procédure, les seules erreurs de reconnaissance automatique de la parole qui pourraient intervenir sur le contenu du corpus d'adaptation sont celles qui permettent de récupérer des documents Web de la thématique du document à transcrire, ce qui peut être vu comme un enrichissement de requête et ne pose donc pas de problème. On peut noter que la mesure de similarité entre mots-clés que les auteurs utilisent ici est adaptée à la tâche visée. La mesure de similarité entre deux mots-clés consiste à utiliser le critère Dice (Rijsbergen, 1979) avec, pour fonction de fréquence, le nombre de documents qu'un moteur de recherche Web retourne. Il s'agit simplement du rapport entre le nombre de document contenant les deux mots que l'on trouve sur le web et la somme des documents contenant l'un ou l'autre des mots.

L'approche proposée par Chen et al. (2003) attaque le problème différemment. Plutôt que d'extraire des mots-clés de la transcription puis de filtrer ceux qui sont susceptibles d'être erronés, ils font en sorte de ne pas extraire de mots-clés erronés. Ils utilisent l'hypothèse précédente de cohérence thématique et n'extraient de la première passe de transcription que des mots-clés en rapport avec la thématique du document. Cela suppose de connaître cette thématique et comme un document est une suite séquentielle de thématiques, ils réalisent au préalable une segmentation thématique de ce dernier. Pour chaque segment thématique, ils extraient des mots-clés qu'ils soumettent à un moteur de recherche Web. Ils obtiennent ainsi un corpus d'adaptation par segment thématique du document à transcrire.

Kajiura et al. (2006) proposent de détecter explicitement les erreurs de reconnaissance automatique de la parole dans la première passe de transcription afin de les filtrer. La détection d'erreurs s'appuye sur deux informations : des scores de confiance issu du processus de reconnaissance automatique de la parole et la proximité sémantique des mots-clés. Chaque mot-clés est associé à un score de confiance fournit par le moteur de reconnaissance automatique de la parole qui indique à quel point il est certain de cette hypothèse. Les mots-clés ayant un score de confiance trop faible sont écartés. De plus, d'une manière similaire à Ito et al. (2009), les mots-clés n'étant pas suffisamment proches de l'ensemble des autres sont écartés. Les mots-clés restants servent ensuite à formuler des requêtes soumises à d'un moteur de recherche Web et les documents résultats servent de corpus d'adaptation.

Un autre piste d'amélioration qui a été explorée est la manière de formuler les requêtes à partir des informations sur le contenu du document à transcrire. En effet, la manière de considérer un document comme pertinent pour les informations dont on dispose sur le document à transcrire est déterminant pour la qualité du corpus d'adaptation ainsi produit.

Une meilleure formulation des requêtes

La formulation des requêtes à partir des informations sur le contenu du document à transcrire est très importante. En effet, la manière de considérer un document comme pertinent pour les informations dont on dispose sur le document à transcrire est déterminant pour la qualité du corpus d'adaptation ainsi produit.

Bulyko et al. (2003) ont proposé de formuler des requêtes permettant de collecter des documents non seulement dans la même thématique que les document à transcrire, mais qui partagent également d'autres caractéristiques linguistiques, comme par exemple le style éditorial. Les modèles de langage n-gramme sont connus pour capturer de manière simple ces caractéristiques linguistique. Ils ont donc proposé de récupérer des documents qui partagent un grand nombre de n-grammes avec le contenu linguistique des documents ciblés par l'adaptation. Pour cela, ils proposent de fournir des requêtes n-grammes à un moteur de recherche Web. Ce dernier retourne alors des documents contenant le plus de ces n-grammes, lesquels constituent le corpus d'adaptation.

Bigi et al. (2004) proposent également de collecter des documents partageant plus que la thématique avec le contenu linguistique de la cible de l'adaptation. Ils proposent également de constituer un corpus de documents qui partagent les caractéristiques n-gramme des documents cibles. Par contre, leur approche permet de se passer de l'extraction de mots-clés ou de n-grammes et d'opérer une recherche de similarité directement au niveau du document. Le critère qu'ils proposent d'utiliser pour mesurer la similarité entre la première passe de transcription et les documents qui peuvent potentiellement servir pour l'adaptation est la distance de Kullback-Leiber (Kullback et Leibler, 1951). Cette mesure fournit la distance entre deux distributions de probabilités. Les documents sont donc représentés dans cette approche par leur distribution n-gramme. Cette approche est plus poussée que la précédente car elle permet de mesurer la similarité des distributions n-gramme entière et pas seulement de quelques n-grammes.

On peut également citer les travaux de Sarikaya et al. (2005b). Leur but est de construire un ensemble de modèles de langage thématiques à partir d'un échantillon de documents textes de chaque thématique. Les documents pour lesquels ils construisent un corpus d'adaptation sont donc des documents texte. Ils proposent d'utiliser un moteur de recherche Web pour collecter les documents d'adaptation. Comme leur but est de construire des modèles de langage n-gramme, ils adaptent la manière de formuler les requêtes à ce contexte. Au lieu de chercher des documents de la même thématique, ils vont chercher des documents qui partagent un même ensemble de n-grammes discriminants avec les documents constituant les échantillons. Pour cela, ils utilisent toutes les suites ininterrompues de mots porteurs de sens comme n-grammes signifiants de taille variable. Les mots non porteurs de sens sont identifiés grâce à une stopliste. Ces n-grammes sont ensuite utilisés comme requêtes dans un moteur de recherche Web et les documents résultats constituent le corpus d'entraînement du modèle de langage pour la thématique de laquelle sont issus ces n-grammes.

Plus tard Tsiartas et al. (2010), qui poursuivent le même but que Sarikaya et al. (2005b), ont proposé une technique du même type, mais qui semble plus performante. L'idée est non pas de récupérer des documents ayant des n-grammes en commun avec les échantillons de texte, mais plutôt de trouver des documents ayant des phrases en commun. Pour cela ils effectuent une segmentation en phrases des documents initiaux. Ils extraient ensuite les phrases une à une, en sélectionnant à chaque fois celle qui est la plus distante sémantiquement de l'ensemble des phrases déjà extraites. Ils arrêtent l'extraction de phrase en fonction de la quantité de corpus qu'ils souhaitent récupérer pour chaque thématique. Le premier avantage de cette approche est qu'elle permet de s'assurer que toutes les variations de la thématique considérée seront représentés. Le second avantage est qu'elle permet de contrôler la taille du corpus récupéré tout en conservant un équilibre entre les sous-thématiques.

Une idée similaire à celle de Tsiartas et al. (2010) a été proposée par Meng et al. (2010) dans le contexte de la constitution de corpus d'adaptation pour la reconnaissance automatique de la parole. Comme dans beaucoup d'approches, le contenu linguistique du document à transcrire est obtenu par une première passe de reconnaissance automatique de la parole. Cette transcription est ensuite segmentée en phrases et ces phrases sont utilisées comme requêtes dans un moteur de recherche Web, à l'instar de Tsiartas

et al. (2010).

Liu et al. (2007) proposent d'utiliser des phrases entières comme requête. La différence est qu'ils utilisent un seul type de phrases : celles dont le sujet est un nom. Comme les autres approches, ces phrases sont utilisées pour collecter des documents auprès d'un moteur de recherche Web. Ces documents constituent le corpus d'adaptation. Les phrases sont extraites des meta-données qui sont associées aux documents à transcrire.

Pour augmenter la précision des requêtes de document, on trouve dans la littérature différentes méthodes d'enrichissement des requêtes par combinaison de mots-clés.

Des requêtes plus précises

Afin d'améliorer l'efficacité des requêtes en y apportant plus de précision, certains auteurs proposent de combiner plusieurs mot-clés par requête. Pour que cette combinaison soit pertinente, il est nécessaire que les mots-clés appartiennent à la même thématique.

Par exemple, Lecorvé et al. (2008) disposent d'un système de segmentation thématique qui leur permet de découper le document à transcrire en segments homogènes. De chaque segment, ils extraient des mots-clés de la première passe de reconnaissance automatique de la parole grâce à la métrique TD.IDF. Ces mots-clés sont ensuite combinés aléatoirement pour former des requêtes. Ces requêtes servent ensuite à récupérer des documents auprès d'un moteur de recherche Web. Ils constatent que des requêtes de plus de 6 mots sont trop discriminantes et ne retournent pas assez de documents pour construire un modèle de langage avec. Afin de filtrer les documents inappropriés qui auraient pu se glisser dans les résultats de recherche, ils mesurent la similarité cosine des documents récupérés avec la première passe de transcription du segment associé à la requête. Au delà d'un certain seuil empirique, les documents sont rejetés. De cette manière ils obtiennent un corpus d'adaptation très ciblé et relativement peu bruité.

Ito et al. (2009) ont également étudié une meilleure manière de combiner les mots-clés. Ils proposent de regrouper les mots-clés relatifs à une thématique en groupes représentatifs des sous-thématique. Ces groupes fournissent donc des requêtes de taille variables et ciblés sur une sous-thématique. De cette manière les documents sont mieux ciblés et on a la possibilité de contrôler la quantité de documents relatifs à chaque sous-thématiques dans le corpus d'adaptation ainsi constitué.

Enfin, une dernière piste explorée pour l'optimisation de ce type d'approche est d'ajuster le nombre de documents récupérés auprès des moteurs de recherche en fonction des requêtes. De cette manière, le corpus collecté par requête sera plus précis.

Optimis ation de la collecte de documents par requête

Les documents retournés par un moteur de recherche correspondent tous à la requête et sont ordonnés par pertinence décroissante. Le nombre de documents à sélectionner pour l'adaptation détermine le degré de pertinence minimal admis pour le corpus d'adaptation. Dans la plupart des approches présentées, le nombre de documents est fixe. Des travaux ont été effectués afin d'optimiser ce facteur pour chaque requête.

Ito et al. (2009) proposent l'approche suivante : plus les documents récupérés pour une requête sont en moyenne proches du contenu du document à transcrire, plus nombreux seront les documents de cette requête qui constitueront le corpus d'adaptation. La difficulté majeure de cette approche est d'estimer la pertinence moyenne des résultats d'une requête par rapport au contenu linguistique du document à transcrire. Pour cela, ils proposent de récupérer les 100 premiers documents résultats de la requête, de les agglomérer et de mesurer la similarité cosine entre ce meta-document et la première passe de transcription du document audio. L'espace dans lequel sont projetés les documents est celui des noms présents dans le document à transcrire, avec TF.IDF comme valeur. Etant donné J requêtes, pour chaque requête j, le nombre de documents à récupérer pour cette requête N_j, se calcule très simplement. Ils faut commencer par fixer le nombre global de documents souhaités pour le corpus d'adaptation, que nous appellerons N_d. N_j est alors la proportion de N_d correspondant à la proportion de pertinence de la requête j par rapport à la pertinence globale. Si la pertinence des documents de la requête j est notée P_j, alors N_j est donné par la formule 4.1 :

$$N_j = N_d \times \frac{P_j}{\sum_{k=1}^{J} P_k} \tag{4.1}$$

Ils montrent que cette approche produit un corpus d'adaptation qui est plus pertinent pour le document à transcrire que si un nombre fixe de documents était récupéré pour chaque requête.

4.4.2 Sélection des nouveaux mots

Une fois la source de mots choisie, il faut sélectionner ceux qui sont le plus pertinents. Le but de l'adaptation du lexique est principalement de réduire le taux de mots hors-vocabulaires. Cependant, la taille du lexique ne peut croître démesurément car cela nécessite plus de ressources computationnelles pour le décodage et, au delà d'une certaine taille, la confusion acoustique atteint un seuil où le taux d'erreur mots augmente (Rosenfeld, 1995). Il est donc important de supprimer du lexique initial les mots les moins pertinents. Le choix des mots à supprimer et à ajouter est déterminant pour le taux de mots hors-vocabulaires des documents confrontés au lexique adapté, mais il peut aussi modifier considérablement les performances du système si la structure initiale du lexique est dégradée ou si trop de bruit y est introduit.

Le risque encouru lors de l'ajout de mots est principalement l'introduction de bruit dans le lexique en remplacement d'informations pertinentes. Les mots non pertinents

ajoutés ne posent *a priori* pas de problèmes de performances, mais c'est le caractère fini du lexique qui fait que ce bruit prend la place de mots plus pertinents qui pose problème. De plus, plus le lexique contient de mots, plus le risque de confusion acoustique due à deux mots de prononciations proches est élevé.

Approches par proximité contextuelle

La plupart des approches consistent à sélectionner les mots non présents dans le lexique initial et qui apparaissent dans des contextes similaires à l'hypothèse de reconnaissance automatique de la parole.

Ohtsuki et al. (2005) proposent une indexation contextuelle des mots inconnus présents dans les documents sélectionnés pour l'adaptation. Les mots à intégrer dans le lexique initial de reconnaissance automatique de la parole seront ceux qui apparaissent dans des contextes proches de celui observé dans une première passe de reconnaissance automatique de la parole. L'indexation contextuelle consiste en fait à représenter une phrase par un vecteur de coocurrences moyen dans un espace de mots. Un ensemble de mots pivots, appelés mots-concepts, est sélectionné dans le corpus d'entraînement du système de reconnaissance automatique de la parole. Ces mots-concepts doivent naturellement être présents dans le lexique initial afin de pouvoir les utiliser pour représenter la première passe de reconnaissance automatique de la parole. Une matrice de co-occurrence est ensuite construite avec ces mots, associant ainsi un vecteur de coocurrence à chacun. Une phrase est alors représentée par la moyenne des vecteurs de co-occurrences des mots-concepts qui y apparaissent. Chaque mot inconnu observé dans les documents d'adaptation est représenté par le vecteur de coocurrences moyen de la phrase dans laquelle il apparaît. Les N mots les plus proches de l'hypothèse de reconnaissance automatique de la parole dans cet espace de co-occurrences moyennes selon la distance Cosine sont sélectionnés pour être ajoutés au lexique de reconnaissance automatique de la parole.

Allauzen et Gauvain (2003) proposent une technique originale pour sélectionner les mots à insérer dans le lexique adapté. L'idée fondatrice est que les N mots les plus fréquents des documents à transcrire constituent la sélection de mots d'adaptation maximisant la couverture lexicale. Comme la distribution de mots des documents à transcrire est inconnue, ils proposent de se servir d'un corpus de développement suffisamment représentatif. Les N mots de ce corpus de développement constituent donc la sélection de mots optimale. Cependant, ce corpus de développement peut ne pas être suffisamment proche des documents à transcrire et risque donc de produire une liste de mots non optimale. C'est par exemple le cas si la distribution de mots est globalement proche de celle des documents à transcrire, mais qu'il manque un ensemble de mots relatifs à une thématique abordée dans le document audio. Par ailleurs, ils disposent d'un ensemble de corpus d'adaptation qui, étant donné leur taille et leur méthode de sélection, couvrent l'ensemble des thématiques susceptibles d'être abordées. Ils proposent alors de calculer des coefficients d'interpolation linéaire des distributions de mots des corpus d'adaptation pour se rapprocher le plus possible de la distribution de mots du

corpus de développement. En effet, si l'interpolation est suffisamment précise, la distribution de mots issue de cette interpolation sera très proche de celle du corpus de développement, et donc des documents à transcrire. De plus, de part la linéarité de l'interpolation, la distribution résultante contiendra probablement des mots relatifs à des thématiques non présentes dans le corpus de développement. Les N mots les plus fréquents de la distribution interpolée sont sélectionnés pour augmenter le lexique initial. Les auteurs montrent expérimentalement que cette approche est plus performante que la précédente, tout en nécessitant moins de connaissances *a priori*. L'inconvénient majeur de cette méthode est qu'il faut disposer d'un corpus de développement suffisamment proche des documents à transcrire.

Liu et al. (2007) proposent d'entraîner un classifieur à sélectionner les mots pertinents dans les documents d'adaptation. Leur idée est que les mots hors-vocabulaires potentiellement absents des documents à transcrire peuvent être identifiés en analysant leur distribution. Ils proposent d'extraire des documents d'adaptation un ensemble de statistiques pour chaque mot. Il s'agit de métriques courantes en recherche d'information :

1. la fréquence du mot dans le document
2. la fréquence normalisée du mot dans le document ($TF(w)$ avec w le mot en question)
3. le TF.IDF du mot ($TF(w) \times IDF(w)$)
4. le TTF (Tapered Term Frequency) du mot ($TF(w) = \log(1 + TF(w))$)
5. le TTF.IDF du mot ($TTF(w) \times IDF(w)$)

Ces paramètres sont fournis à un classifieur de type réseaux de neurones qui prend la décision de sélectionner ou non le mot pour l'adaptation. L'entraînement du classifieur est effectué avec un ensemble de documents audio pour lesquels les transcriptions de référence sont disponibles. A chaque document à transcrire, est associé un ensemble de documents d'adaptation. Chaque mot de ces documents non présent dans le lexique initial mais présent dans la transcription de référence est indiqué comme "à sélectionner" dans le classifieur. En fait, le classifieur apprends à reconnaître des mots des documents d'adaptation qui se trouvent dans la transcription de référence. Ils obtiennent expérimentalement de très bonnes performances avec cette approche.

Martins et al. (2010) ont proposé que la distribution morphosyntaxique des mots ajoutés dans le lexique adapté respecte la distribution observée dans un échantillon de référence des documents à transcrire. Ils mesurent la fréquence normalisée d'un certain nombre d'étiquettes morphosyntaxiques (noms, verbes, adjectifs et adverbes) dans un échantillon de référence. Notons $F_n(e)$ la fréquence normalisée de l'étiquette e dans l'échantillon de référence, avec $0 \leq F_n(e) \leq 1$ et $\sum_e F_n(e) = 1$. Ils étiquettent ensuite les mots inconnus présents dans le corpus d'adaptation. Pour une augmentation du vocabulaire de V mots et pour chaque étiquette morphosyntaxique e, ils sélectionnent les $N(e) = V \times F_n(e)$ mots les plus fréquents ayant pour étiquette morphosyntaxique e dans le corpus d'adaptation. Ainsi la distribution morphosyntaxique des mots ajoutés dans le lexique respecte celle des documents à transcrire. On notera

cependant que, comme l'ont observé par exemple Allauzen et Gauvain (2003), les mots hors-vocabulaires ont une distribution morphosyntaxique particulière puisqu'ils sont composés à plus de 90% de noms. Il faudrait donc en théorie ajouter au lexique adapté un ensemble de mots respectant cette distribution et non la distribution observée dans l'échantillon représentatif. Ils obtiennent cependant de bons résultats avec cette approche.

Approches fréquentistes

Le problème de la sélection des mots peut parfois être résolu simplement. On peut citer par exemple les travaux de Yu et al. (2000b). Les documents qu'ils doivent transcrire sont des enregistrements de réunions d'entreprise. Le corpus d'adaptation qu'ils utilisent est constitué de l'ensemble des documents du site Web de l'entreprise. Dans ces conditions, la majorité des mots inconnus présents dans le corpus d'adaptation sont susceptibles de figurer dans le document à transcrire. C'est pour cela que l'utilisation d'un simple seuil de fréquence suffit à former la sélection des mots à insérer dans le lexique. Ce seuil permet de limiter l'expension du lexique à une taille raisonnable.

Un autre exemple est fourni par Bertoldi et Federico (2001) et développé dans Federico et Bertoldi (2004). Ils proposent une approche simple pour sélectionner les mots à intégrer dans le lexique adapté. Les documents qu'ils transcrivent sont des journaux radiodiffusés et leur corpus d'adaptation est récupéré chaque jour sur des sites Web d'information. Pour chaque document audio à transcrire ils possèdent donc, comme corpus d'adaptation, les données collectées sur le Web le jour même et les jours précédents. Afin de suivre l'évolution de l'actualité, ils proposent d'utiliser un double critère de sélection des mots : la fréquence d'apparition des mots et la proximité temporelle de parution des documents dans lesquels ils se trouvent. Ils montrent que sélectionner en priorité les mots apparaissant dans les documents récents puis par fréquence dans l'ensemble des documents est une meilleure stratégie que de n'utiliser que l'une ou l'autre des méthodes.

Une approche semblable a été proposée par Allauzen et Gauvain (2003). La tâche de reconnaissance automatique de la parole considérée est la transcription de journaux d'actualité radio ou télé et le corpus d'adaptation est collecté chaque jour sur des site Web d'information. Ils proposent cependant une théorie pour expliquer les bonnes performances de l'approche. Il supposent l'existence de deux catégories de mots inconnus à intégrer dans le lexique d'adaptation : les mots relatifs à des thématiques récurrentes et ceux relatifs à des thématiques ponctuelles. Pour identifier ces deux groupes de mots, ils proposent une méthode proche de la précédente, basée sur les fréquences d'occurrence des mots dans les données d'adaptation. Pour identifier les mots issus des thématiques récurrentes, ils sélectionnent tous les mots présents au moins 5 fois dans les données collectées sur le Web des quatre semaines précédentes. Pour les mots qui se rapportent aux thématiques ponctuelles, ils sélectionnent tous les mots présents dans les données du jour. Cette approche a pour avantage sa simplicité de mise en oeuvre. Par contre, elle a deux inconvénient majeurs. Le premier est qu'il faut que les données d'adaptation soient très proches des données à transcrire en terme de distribution de

mots, ce qui nécessite des connaissances *a priori* des données à transcrire. Le second est que la méthode ne tient pas compte des particularités de chaque document à transcrire.

Approches classiques

Dans certains cas, la méthode de sélection des mots peut être la méthode utilisée pour construire le lexique du modèle de langage initial. En effet, lorsque les documents d'adaptation sont utilisés pour construire un nouveau modèle de langage qui sera interpolé avec le modèle initial, la sélection du lexique de ce modèle se fait de manière classique. Les lexiques des deux modèles sont ensuite fusionnés pour que l'interpolation puisse se faire. C'est le cas généralement dans les travaux qui se focalisent sur l'adaptation du modèle de langage à un domaine de spécialité. On peut citer par exemple les travaux de Tsiartas et al. (2010), ceux de Ito et al. (2009), de Lecorvé et al. (2008) ou encore de Chen et al. (2003).

Approches intégrales

Dans certaines situations, le problème de la sélection des mots à ajouter au lexique initial ne se pose pas. C'est souvent le cas lorsque peu de données d'adaptation sont disponibles et qu'ajouter l'ensemble des mots inconnus observés dans ces données n'introduit que très peu de confusion acoustique. Par exemple Allauzen et Gauvain (2005b) proposent d'adapter le lexique d'un système de reconnaissance automatique de la parole conçu pour la transcription d'archives audiovisuelles. La quantité de données d'adaptation est très faible car il s'agit des résumés des documents à transcrire. En effet, il est difficile de trouver de gros volumes de données textuelles numérisées relatives à des documents audio anciens. Le nombre de mots inconnus présents dans ces résumés est de l'ordre de 1.6% et la taille des résumés est d'environ 350 mots, ce qui représente entre 5 et 6 mots inconnus à insérer dans le lexique par document.

Il en va de même lorsque la méthode de sélection des documents d'adaptation est particulièrement précise et retourne peu de documents très pertinents. C'est le cas par exemple des travaux de Bigi et al. (2004). Comme nous l'avons vu à la section 4.4.1, ils sélectionnent tous les mots inconnus des N documents d'adaptation les plus pertinents. Le contrôle de la taille du lexique adapté est fait en choisissant le nombre de documents d'adaptation utilisés pour le construire.

4.4.3 Phonétisation des nouveaux mots

Une autre difficulté rencontrée lors de l'ajout automatique de mots dans le lexique est leur phonétisation. En effet, dans les systèmes de reconnaissance automatique de la parole markoviens modernes, les mots du lexiques doivent être associées à leur représentation phonétique dont l'unité est habituellement le phonème.

La qualité de la phonétisation des mots est déterminante pour les performances des systèmes de reconnaissance automatique de la parole. Les mots mal phonétisés peuvent apparaître alors qu'ils ne devraient pas et peuvent perturber le décodage si on leur attribue une probabilité acoustique trop élevée. De plus, un mot est en général associé à plusieurs variantes de phonétisation, car il existe différentes prononciations suivant les accents régionaux, le contexte phonétique, etc. Le risque de commettre une erreur est alors accru car il faut posséder plus de connaissances linguistiques pour renseigner et vérifier les phonétisations.

Phonétisation manuelle

Pour obtenir les meilleures performances il est souvent nécessaire de recourir à la phonétisation manuelle par des spécialistes. Cependant, cette tâche est coûteuse et ne peut être réalisée sur demande lors de la phase d'adaptation, auquel cas il ne s'agirait plus d'adaptation automatique. Il existe plusieurs solutions pour contourner ce problème tout en conservant une bonne qualité de phonétisation.

Phonétisation automatique

Une première solution consiste à phonétiser automatiquement les mots à partir de leur graphie. Il existe aujourd'hui de nombreuses approches performantes pour effectuer cette tâche. Ceci est en partie dû au fait que cette tâche est souvent nécessaire à la synthèse vocale et que cette dernière a suscité beaucoup d'intérêt ces dernières années. En effet, la plupart des systèmes modernes transcrivent le texte à synthétiser en phonèmes qui sont ensuite fournis au moteur de synthèse vocale.

Cependant, le problème de la phonétisation n'est pas totalement résolu et il est encore un sujet de recherche actif. Plusieurs aspects du problème sont difficiles à résoudre, notamment le problème des homographes hétérophones, des noms propres, des emprunts à des langues étrangères, etc. Il existe deux approches majeures pour réaliser ce travail : celle basée sur des règles et celle basée sur l'apprentissage automatique.

Les premières approches étaient celles à base de règles car les approches automatiques nécessitent beaucoup de ressources computationnelles et humaines pour construire les exemples servant à l'entraînement des modèles. L'idée est d'appliquer un certains nombre de règles de transformation à la graphie du mot pour produire sa représentation phonétique. Ces règles peuvent être vue sous la forme d'un automate à états fini comme l'ont montré (Kaplan et Kay, 1994). Dans de tels systèmes, les exceptions aux règles sont en général traîtées explicitement à l'aide d'une liste de mots dont la phonétisation est renseignée explicitement. On peut par exemple citer les travaux de Allen (1976) ou encore ceux de Divay et Vitale (1997).

L'avantage de tels systèmes est qu'ils sont simples à mettre en oeuvre et rapides lors de l'exécution. Par contre, ils présentent plusieurs inconvénients. Le premier est

l'écriture des règles, qui est complexe. Cette tâche nécessite des compétences linguistiques poussées, bien que, comme le montre Torkkola (1993), il soit possible d'extraire automatiquement un certain nombre de règles à partir d'exemples. De plus, les langues naturelles admettent en général des exceptions qu'il faut gérer de manière plus ou moins élégante. L'interdépendence des règles est également problématique, notamment lorsque le nombre de règles est grand. Ce dernier point rend le développement et la maintenance des règles difficiles.

Plus tard, avec l'apparition de larges bases lexicales phonétisées, l'idée est apparue qu'il était peut-être possible de prédire la prononciation de mots inconnus par analogie avec un grand nombre d'exemples connus. L'avantage principal de cette approche est que l'intervention humaine se limite à fournir des exemples de phonétisations et non plus à écrire des règles générales et complexes. En effet, il est plus facile de produire la phonétisation d'un mot que de construire des règles générales de prononciation. Des algorithmes et des modèles statistiques ont donc commencé à être développés pour capturer automatiquement ces analogies.

Sejnowski et Rosenberg (1987) ont par exemple proposé une des premières approches à apprentissage automatique reposant sur l'utilisation de réseaux de neurones. Ce sont ces approches par apprentissage qui suscitent le plus d'intérêt actuellement (Taylor, 2005; Jiampojamarn et al., 2008; Rama et al., 2009).

Approches semi-automatiques

Un compromis entre la phonétisation manuelle et la phonétisation automatique est possible. Par exemple, il est possible de phonétiser manuellement *a priori* un lexique bien plus grand que celui utilisé dans le système de reconnaissance automatique de la parole. Lors de l'adaptation du lexique, les phonétisations du lexique constitué *a priori* sont utilisées pour tous les nouveaux mots introduits dans le lexique initial qui s'y trouvent. Pour les autres mots, un phonétiseur automatique peut être utilisé. En réduisant ainsi le nombre de mots phonétisés automatiquement, on réduit également le risque de commettre une erreur.

4.4.4 Score linguistique des nouveaux mots

En plus de la phonétisation des mots, les systèmes de reconnaissance automatique de la parole markoviens nécessitent des scores linguistiques associés aux mots. Le rôle du modèle de langage est de les fournir. La plupart des systèmes actuels utilisent des modèles n-grammes. Les scores linguistiques sont donc généralement des probabilités n-grammes qui sont estimées sur un gros volume de données.

Les mots qui sont ajoutés au lexique lors de la phase d'adaptation doivent être intégrés au modèle de langage n-gramme pour pouvoir être utilisés lors du décodage. Cette intégration consiste en l'affectation de probabilités non nulles aux mots tout en conservant l'intégrité du modèle initial. L'estimation de ces probabilités est délicate car elle

nécessite un corpus conséquent qui n'est pas nécessairement disponible pour les mots ajoutés. De plus, il faut intégrer les nouvelles probabilités estimées dans le modèle de langage initial en évitant d'y introduire du bruit.

On trouve dans la littérature plusieurs approches pour réaliser l'intégration de nouveaux mots dans le modèle de langage.

Probabilité basée sur la classe des mots inconnus

L'une des approches les plus simples à mettre en oeuvre pour intégrer de nouveaux mots dans un modèle de langage n-gramme est d'utiliser la classe des mots inconnus. La mise en place de cette classe passe généralement par l'ajout dans le lexique du modèle de langage d'une entrée spéciale qui représente n'importe quel mot non présent dans ce lexique. On utilise habituellement le mot "<UNK>" pour cet usage. Les probabilités n-grammes associées à ce mot seront celles modélisant l'apparition d'un mot non présent dans le lexique. On dit alors que ce mot modélise la classe des mots inconnus. Comme toutes les entrées du modèle de langage ne sont pas de classes, il s'agit d'un hybride entre un modèle de langage n-gramme de classes comme décrit à la section 2.4 et un modèle de langage n-gramme conventionnel décrit à la section 2.3.

L'estimation des probabilités de la classe des mots inconnus est très simple. Dans le cas où le lexique du modèle de langage possède la classe des mots inconnus, tous le mots n'appartenant pas au lexique trouvés dans le corpus d'entraînement sont remplacés par le mot représentant la classe des mots inconnus. L'algorithme d'apprentissage du modèle de langage le modélise alors comme les autres mots du lexique. Ce n'est que lors de l'utilisation du modèle de langage qu'il sera fait un usage particulier de cette entrée du lexique représentant une classe de mots.

La plupart des modèles de langages n-grammes traitent le problème des mots inconnus dans le corpus d'apprentissage de cette manière, ce qui est un gros avantage de cette méthode puisqu'il n'y a rien à développer.

La classe de mots fournit donc une probabilité de base qu'il va falloir utiliser de manière appropriée. Il existe plusieurs manières d'utiliser cette classe de mots pour insérer de nouveaux mots dans le lexique.

La première consiste à se servir de la probabilité de la classe pour dériver des probabilités unigrammes. Avec cette approche, les nouveaux mots sont ajoutés au lexique de manière classique. Une probabilité unigramme est ensuite dérivée de la probabilité de la classe de mots, par exemple de la manière présentée dans la formule 4.2 :

$$P(w) = P(U) \times P(w|U) \tag{4.2}$$

où $P(w)$ est la probabilité du nouveau mot w, $P(U)$ est la probabilité de la classe des mots inconnus et $P(w|U)$ est la probabilité unigramme du mot w au sein des mots inconnus.

La probabilité $P(w|U)$ peut par exemple être estimée en calculant la distribution unigramme dans les documents d'adaptation des mots ajoutés au lexique, comme le

proposent par exemple Federico et Bertoldi (2004) :

$$P(w|U) = \frac{C(w)}{\sum_{w' \in V'} C(w')} \tag{4.3}$$

où $C(w)$ est la fréquence du mot w dans le corpus d'adaptation et V' est l'ensemble des mots ajoutés au vocabulaire initial.

Cependant, lorsque les données d'adaptation ne sont pas suffisantes où pour des questions de performances, cette probabilité est simplifiée, voir remplacée par une valeur déterminée empiriquement. Ohtsuki et al. (2005) proposent par exemple d'estimer $P(w|U)$ de la manière suivante :

$$P(w|U) = \frac{1}{|V'|} \tag{4.4}$$

Une fois la probabilité unigramme du mot calculée pour chaque mot, il faut les insérer dans le modèle de langage en tant qu'unigrammes. Pour que l'intégrité du modèle soit préservée, il est nécessaire de retrancher de la probabilité de la classe des mots inconnus la masse de probabilité ainsi introduite. La probabilité de la classe des mots inconnus devient alors :

$$P'(U) = P(U) - \sum_{w \in V'} P(w|U) \tag{4.5}$$

On remarque que si la méthode de calcul des probabilités unigrammes des nouveaux mots est l'une de celles présentées ici, alors la somme des probabilités introduites est égale à la probabilité initiale de la classe des mots inconnus, cette dernière se verra donc attribuer une probabilité nulle. Par contre dans le cas où une constante remplace la probabilité $P(w|U)$, elle sera probablement non nulle.

Il est naturellement possible d'appliquer cette technique à tous les ordres du modèle de langage en modifiant les distributions de probabilités de manière analogue à la distribution unigramme :

$$P(w|h) = P(U|h) \times P(w|U) \tag{4.6}$$

où h est l'historique du mot w.

Cette méthode est intéressante car elle permet l'insertion des nouveaux mots sans la ré-estimation du modèle de langage initial. L'insertion des nouveaux n-grammes nécessite cependant dans la plupart des cas la recompilation du modèle de langage pour le système de reconnaissance automatique de la parole.

Un des inconvénients majeurs de cette approche est qu'elle n'est pas assez précise et ne modélise que grossièrement l'apparition des mots insérés.

Probabilité de la classe de mot

Cette approche est semblable à la précédente sauf qu'elle a pour but de modéliser plus précisément les nouveaux mots insérés dans le lexique. Avec cette méthode, la

classe des mots inconnus est divisée en plusieurs sous-classes qui modélisent le comportement des mots inconnus en fonction de leur nature.

La classe des mots inconnus vue précédemment modélise le comportement des mots inconnus sans distinction. Cependant, ce comportement est probablement différent pour chaque type de mot. Par exemple on sait que généralement les mots de différentes catégories morphosyntaxiques se comportent différemment. L'idée ici est donc de remplacer la classe des mots inconnus par les classes des mots inconnus de différentes catégories morphosyntaxique.

On peut noter que, comme pour le modèle n-gramme de classe décrit à la section 2.4, il est possible de choisir un autre critère de constitution des classes, comme par exemple des classes statistiques. Cependant la catégorie morphosyntaxique d'un nouveau mot peut être déterminée avec très peu de données, comme par exemple la phrase dans laquelle apparaît une unique occurrence du mot, et les ressources annotées existent pour entraîner ce type d'étiqueteurs.

Techniquement, il faut insérer dans le lexique du modèle de langage initial une entrée par catégorie de mot inconnu. Tous les mots inconnus du corpus d'entraînement du modèle de langage sont remplacés par leur étiquette morphosyntaxique. Le modèle de langage initial est ensuite estimé sur ces données de manière classique. De cette manière, la probabilité modélisée pour ces classes de mot sera la probabilité d'occurrence d'un mot inconnu appartenant à une classe donnée.

L'insertion de nouveaux mots dans le lexique initial se déroule comme avec la précédente technique. Chaque nouveau mot est ajouté comme une nouvelle entrée du lexique. Pour l'intégration de ces mots dans le modèle de langage il faut dériver des probabilités n-gramme à partir des probabilités de classe présentes dans le modèle. Cela peut se faire de la manière suivante :

$$P(w|h) = P(C_w|h) \times P(w|C_w) \tag{4.7}$$

avec C_w la classe du mot w, $P(C_w|h)$ la probabilité d'avoir un mot inconnu de classe C_w sachant l'historique h et $P(w|C_w)$ la probabilité du mot w au sein de la classe C_w.

L'estimation de la $P(w|C_w)$ peut se faire simplement en calculant la distribution unigramme des classes C_w, comme proposé par Allauzen et Gauvain (2005b) :

$$P(w|C_w) = \frac{C(w)}{\sum_{w' \in V' \cap C_w} C(w')} \tag{4.8}$$

où $C(w)$ est la fréquence du mot w dans le corpus d'adaptation et V' est l'ensemble des mots ajoutés au vocabulaire initial.

Cette approche modélise avec plus de précision le comportement des mots inconnus insérés dans le modèle de langage que la méthode présentée précédemment. Cependant cette approche s'appuye toujours sur des données imprécises que sont les probabilités des classes de mots.

Interpolation linéaire des modèles de langage

Afin de modéliser avec encore plus de précision le comportement des mots introduits dans le lexique, il a été proposé de se passer complètement des probabilités de classes calculées *a priori*. L'idée est que, lorsque les données d'adaptation sont importantes, il est possible d'estimer à partir de ces données des probabilités n-grammes fiables pour les nouveaux mots et que de ce fait l'utilisation des probabilités de classes ne feraient qu'introduire du bruit.

L'idée de cette approche est d'entraîner un ou plusieurs nouveaux modèles de langage sur les données sélectionnées pour l'adaptation. Le lexique de ces nouveaux modèles est composé des mots présents dans le lexique initial et des mots sélectionnés pour l'augmentation. De cette manière, les nouveaux modèles contiendront les distributions complètes des mots à introduire dans le lexique adapté. Il suffit alors de réaliser une interpolation linéaire comme décrite dans la section 2.9.1 pour obtenir un modèle final intégrant parfaitement les nouveaux mots :

$$\hat{P}(w|h) = \sum_{i=1}^{N} \alpha_i \times P_i(w|h) + (1 - (\sum_{i=1}^{N} \alpha_i)) \times P_{init}(w|h), \text{avec} 0 \leq \sum_{i=1}^{N} \alpha_i \leq 1 \qquad (4.9)$$

avec $P_i(w|h)$ la probabilité du mot w connaissant son historique h selon le modèle de langage i issu des documents d'adaptation, $P_{init}(w|h)$ la probabilité fournie par le modèle de langage initial et $(\alpha_i)_{i=1}^{N}$ l'ensemble des coefficients d'interpolation représentant le poids de chaque distribution dans le modèle final.

Lorsque les données d'adaptation sont suffisantes, cette solution est très souvent retenue dans le domaine de l'adaptation du lexique. On la retrouve par exemple dans les travaux de Meng et al. (2010), de Martins et al. (2007a), etc.

L'avantage majeur de cette approche est qu'elle permet une meilleure modélisation des mots introduits dans le lexique. Cependant elle nécessite de posséder suffisamment de données d'adaptation suffisamment représentatives de l'usage des mots introduits dans le lexique.

4.4.5 Autres approches

On trouve quelques approches qui n'entrent pas dans le schéma que l'on vient de présenter.

On peut notamment citer les approches à transcription phonème vers graphème, comme le proposent exemple Decadt et al. (2002). L'idée est d'utiliser la transcription phonétique des mots hors-vocabulaires détectés pour reconstruire leur graphie. Cette dernière est alors directement insérée dans l'hypothèse finale de transcription. Il existe plusieurs manières d'obtenir la transcription phonétique des mots hors-vocabulaires détectés. Il est possible de l'obtenir grâce au treillis de phonèmes, si le détecteur de

mots hors-vocabulaires localise précisément la position du mot. Il est également possible d'estimer un modèle de langage hybride contenant la phonétisation des mots hors-vocabulaires du corpus d'entraînement, comme le proposent Yazgan et Saraclar (2004).

Un avantage majeur de cette approche est qu'il n'y a pas besoin de chercher leur transcription phonétique ni leur score linguistique, puisque les mots ne sont pas insérés dans le processus de décodage. Un autre avantage qui découle du précédent est qu'elle ne nécessite pas de source d'information externe, le système est donc autonome.

L'inconvénient majeur est que les performances de cette approche sont fortement conditionnées par celles de la transcription phonèmes vers graphèmes. Elle souffre donc des mêmes faiblesses, notamment la forte dépendance aux caractéristiques de la langue utilisée. Par exemple plus cette langue admettra de lettres muettes, moins les performances seront bonnes. Dans beaucoup de langues la relation entre les phonétisations possibles et les graphies associés n'est pas une bijection, il existe des homophones hétérographes et des hétérophones homographes. De plus, une partie des mots hors-vocabulaires sont des mots empruntés aux autres langues ou des noms propres d'entités étrangères. Dans de telles situations, le moteur de transcription phonème vers graphème doit savoir s'adapter. Étant donné ces limites majeures, ce type d'approche est très peu utilisée en pratique.

4.5 Conclusion du chapitre

Nous avons vu en quoi consiste le problème des mots hors-vocabulaires en reconnaissance automatique de la parole et quelles en sont les causes et les conséquences. Nous avons présenté une vue synthétique des travaux déjà effectués dans le domaine en proposant un découpage de la tâche en quatre étapes qui peuvent être considérées comme indépendantes :

1. Collecter un corpus d'adaptation qui soit le plus proche possible du contenu des documents à transcrire
2. Sélectionner dans ce corpus les mots qui seront introduits dans le lexique à adapter
3. Associer une représentation phonétique à chacun des mots à introduire dans le lexique
4. Attribuer une probabilité linguistique à ces mots dans le modèle de langage.

Une approche typique de ce qui a été proposé jusqu'à présent dans la littérature pourrait consister à :

1. Collecter de documents relatifs au thème du document à transcrire par similarité avec une première passe de transcription, à l'aide de techniques de recherche d'information

2. Sélectionner dans ces dernier des N nouveaux mots les plus fréquents qui remplaceront les N mots les moins fréquents du lexique initial
3. Phonétiser automatique des nouveaux mots
4. Estimer un modèle de langage avec le nouveaux lexique sur les données collectées à l'étape 1 et l'interpoler avec le modèle de langage initial

Un avantage de cette approche est qu'elle est entièrement automatique et qu'elle est assez dynamique pour suivre l'évolution du contenu des documents.

Par contre, elle est fortement dépendante de la pertinence des données collectées et donc du contenu de la base de documents dans laquelle les recherches sont effectuées. De plus, afin que l'estimation des probabilités des nouveaux mots soit pertinente, il faut que les documents collectés soient de bonne qualité et peu bruités, ce qui nécessite un travail de nettoyage important.

Chapitre 5

Adaptation locale et dynamique du lexique

Sommaire

Comme nous l'avons vu au chapitre 4, beaucoup de travaux ont porté sur le problème de la couverture lexicale en reconnaissance automatique de la parole. La tâche de transcription de journaux radiodiffusés a suscité beaucoup d'intérêt, probablement parce que les mots hors-vocabulaires posent un réel problème pour indexer ce type d'émissions.

Nous allons présenter ici les travaux que nous avons menés dans le cadre de l'augmentation du lexique pour la reconnaissance automatique de la parole. Les résultats sont présentés sur deux tâches. La première est une tâche classique de transcription de journaux radiodiffusés en français et la seconde consiste à transcrire des commentaires de films d'opérations chirurgicales académiques en anglais.

5.1 Introduction

La plupart des systèmes de transcription automatique de journaux radio ou télédiffusés s'appuient sur des modèles de langage n-grammes estimés à partir de journaux d'information disponibles sous forme écrite, comme les journaux papiers ou des transcriptions manuelles de journaux télévisés ou radiodiffusés. Cependant, les journaux d'information traitent de sujets récents et variés qui sont difficiles à anticiper. De ce fait, il y a toujours un décalage entre le lexique des modèles de langage et les journaux d'information à transcrire.

La manière la plus évidente pour résoudre le problème de couverture lexicale est d'augmenter la taille du lexique constitué *a priori* à partir du corpus d'entraînement du modèle de langage. Hormis le fait qu'à partir d'une certaine taille de lexique, les performances du système de reconnaissance automatique de la parole chutent (Rosenfeld, 1995), cette approche comporte deux inconvénients. Le premier est qu'elle ne permet pas de compenser le décalage lexical qu'il pourrait exister entre le corpus d'entraînement et le corpus de test. Le second inconvénient est que le lexique ainsi obtenu est figé dans le temps, ce qui est peu pertinent lorsque les données traitées sont hautement dynamiques comme les journaux d'information.

Une autre approche serait de toujours avoir un lexique construit *a priori* mais d'utiliser une source de nouveaux mots autre que celle que constitue le corpus d'entraînement du modèle de langage. Nous avons vu à la section 4.4.1 un grand nombre d'approches exploitant ce principe.

La quantité de données textuelles accessibles depuis les moteurs de recherche du Web connaît une croissance exponentielle depuis de nombreuses années, comme nous l'avons vu au chapitre 3. A ce titre, on peut considérer le Web comme une source de données textuelle quasi-infinie et qui est mise à jour constamment.

Dans le cadre de la modélisation linguistique, ces deux caractéristiques font du Web une ressource de choix qu'il est intéressant d'exploiter. Pour la tâche d'enrichissement du lexique de reconnaissance automatique de la parole, le fait que le Web soit de taille considérable augmente la probabilité qu'il contienne les mots manquant au lexique.

Son caractère dynamique nous assure qu'il sera mis à jour au gré des événements se produisant dans le monde, comme l'arrivée sur la scène médiatique de personnalités inconnues jusqu'alors.

Cette idée a déjà été développée dans le contexte de la reconnaissance automatique de la parole, comme nous l'avons vu au chapitre 4. Pour réaliser l'adaptation du modèle de langage et du lexique, les auteurs proposent généralement de télécharger de grande quantités de données qui soient assez proche du contexte linguistique et sémantique des données à transcrire (Federico et Bertoldi, 2004; Allauzen et Gauvain, 2005a; Martins et al., 2010). Ces approches permettent d'améliorer considérablement la couverture lexicale mais présentent des inconvénients. Tout d'abord, le Web est un corpus plutôt bruité et les modèles de langage entraînés sur ces données sont généralement moins performants que les modèles appris sur des corpus propres et bien ciblés, mais souvent plus petits, comme le montrent Lapata et Keller (2005). Un autre inconvénient est que les mots hors-vocabulaires sont peu fréquents et les statistiques n-grammes qui leurs sont associées sont souvent mal estimées, ce qui peut induire une mauvaise utilisation des nouveaux mots introduits dans le lexique. Finalement, pour obtenir une bonne couverture lexicale pour un domaine en particulier, il faut en général augmenter considérablement la taille du lexique car il est difficile de bien cibler les mots nécessaires, ce qui induit une augmentation de la complexité du décodage et peut conduire à une diminution des performances du système de reconnaissance automatique de la parole, comme le montre Rosenfeld (1995).

Au lieu de procéder à une augmentation lexicale globale basée sur des descripteurs sémantiques comme cela a été fait jusqu'à maintenant, nous proposons d'utiliser la structure syntaxique des mots du contexte d'apparition des mots hors-vocabulaires. Une première passe de transcription automatique fournit une hypothèse de transcription et des descripteurs sémantiques et syntaxiques sont extraits du contexte des mots hors-vocabulaires détectés dans celle-ci. Des requêtes Web sont alors formulées à partir des descripteurs et permettent de collecter des candidats. Pour chaque mot hors-vocabulaire, un nouveau lexique est construit en ajoutant au lexique initial les mots récupérés sur le Web. Finalement, les segments contenant des mots hors-vocabulaires sont à nouveau décodés avec ces nouveaux lexiques.

5.2 Principes de l'augmentation locale et dynamique du lexique

L'approche que nous proposons s'appuie sur l'hypothèse que le Web est une source d'information linguistique quasi-exhaustive. A ce titre, on peut le voir comme un corpus complet, c'est à dire qui contient toutes les suites de mots possibles pour une langue donnée. De ce fait, il devrait être possible de récupérer les mots hors-vocabulaires d'une transcription en cherchant ceux qui apparaissent dans le même contexte.

On peut voir sur la figure 5.1 un schéma en flux représentant de manière synthétique l'architecture de recherche automatique de mots hors-vocabulaires que nous pro-

posons. Comme on peut le voir, chaque segment de parole subit les traitements suivants :

1. Une première passe de reconnaissance automatique de la parole avec un modèle de langage initial est effectuée à partir du signal acoustique, ce qui produit une première hypothèse de transcription.
2. Les mots hors-vocabulaires sont ensuite automatiquement détectés et étiquetés dans cette transcription.
3. Pour chaque mot hors-vocabulaire, un patron de mots est extrait de son contexte.
4. Des documents contenant des séquences de mots correspondant à ce patron sont récupérés sur le Web, à l'aide d'un moteur de recherche.
5. Les mots de ces documents qui se trouvent à l'emplacement du mot hors-vocabulaire dans le patron sont extraits et ajoutés au lexique du modèle de langage initial.
6. Enfin, une dernière passe de transcription est appliquée au signal de parole, en utilisant le modèle de langage enrichi des mots extraits des documents Web.

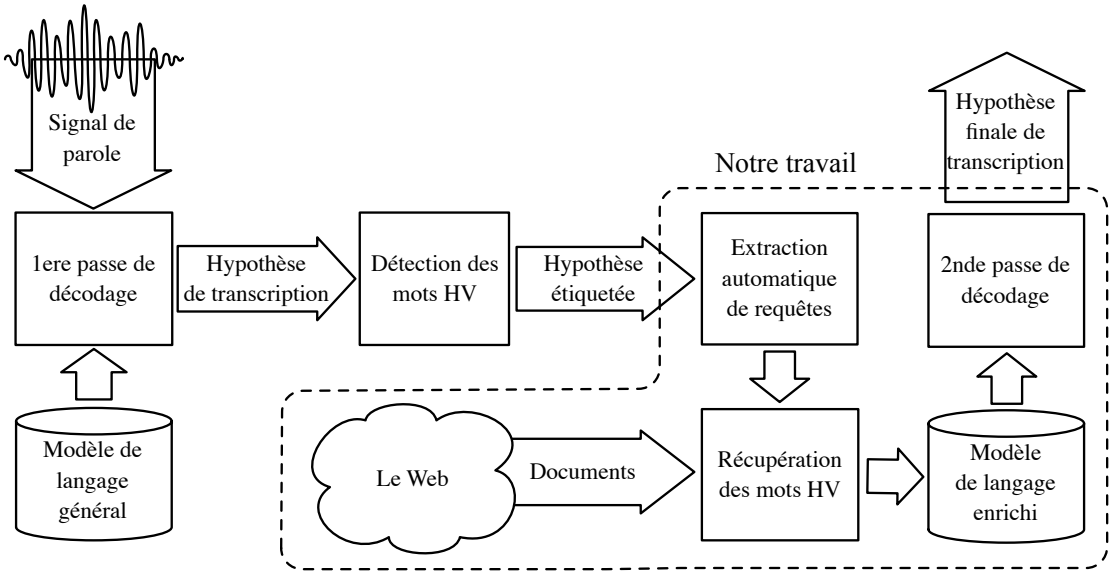

FIGURE 5.1 – *Schéma global de la procédure d'augmentation locale et dynamique du lexique reposant sur l'extraction de patrons syntaxiques.*

Dans ces travaux, la tâche de détection des mots hors-vocabulaires est supposée être effectuée par ailleurs. On trouve beaucoup de travaux à ce sujet dans la littérature, comme on peut le voir à la section 4.3. On remarque notamment les travaux de Lecouteux et al. (2009), membre de l'équipe dans laquelle j'ai effectué les travaux présentés ici.

Notre approche est novatrice car, d'une part, elle repose une augmentation locale du lexique et que, d'autre part, elle s'appuie sur des patrons syntaxiques et sémantiques locaux.

5.2.1 Augmentation locale

L'augmentation du lexique est locale car elle intervient pour chaque segment de parole contenant au moins un mot hors-vocabulaire. Le système utilise donc le lexique initial pour transcrire l'intégralité des documents hormis les passages contenant des mots hors-vocabulaires. Un lexique adapté est construit pour chaque mot hors-vocabulaire rencontré, ce qui a deux avantages majeurs. D'une part, le nombre de mots à ajouter au lexique est réduit car l'augmentation n'est valable que pour un mot hors-vocabulaire. La confusion acoustique introduite par l'ajout de mots dans le lexique s'en trouve donc mieux maîtrisée. D'autre part, il est possible d'optimiser l'augmentation pour un mot en particulier.

Avec une approche globale, si un document à transcrire comporte N_{hv} mots hors-vocabulaires et que le lexique initial est de taille $|V|$, pour proposer en moyenne C candidats par mot hors-vocabulaire, il faudra augmenter le lexique de $C \times N_{hv}$ mots. La taille moyenne du lexique pour décoder le document sera alors de $|V| + C \times N_{hv}$. Par exemple, pour un lexique initial de 65k mots, si l'on voulait introduire 500 candidats en moyenne par mot hors-vocabulaire, il faudrait doubler la taille du lexique pour transcrire une heure de parole avec 1% de mots HV [1]. Comme le montre Rosenfeld (1995), une telle augmentation de la taille du lexique conduirait très probablement à une dégradation de la qualité des transcriptions.

Avec l'approche locale, le document est découpé en $|D|$ passages. Pour chaque passage contenant des mots hors-vocabulaires, un lexique est construit en ajoutant au lexique initial les C candidats de chaque mot hors-vocabulaire qu'il contient. Les passages qui ne contiennent pas de mots hors-vocabulaires sont décodés avec le lexique initial. Il y a en moyenne $\frac{N_{hv}}{|D|}$ mots hors-vocabulaires par passage, la taille moyenne du lexique sur l'ensemble du document est donc de $|V| + C \times \frac{N_{hv}}{|D|}$. Plus le nombre de passages est grand, plus cette approche est intéressante par rapport à l'approche globale.

La différence fondamentale entre l'approche globale et l'approche locale est la périodicité de l'adaptation du lexique. Avec l'approche globale, le lexique est considéré comme une ressource figée qu'il est possible d'adapter périodiquement (chaque document, chaque jour, etc.). Avec l'approche locale, le lexique est dynamique. Il est mis à jour chaque fois qu'il est nécessaire de le faire. Avec l'approche globale, le lexique est fermé et adapté périodiquement et avec l'approche locale il est ouvert et dynamique.

Pour des raisons pratiques, on considère par la suite qu'un passage est un segment. Comme nous l'avons présenté au chapitre 1, le segment est une sous-partie homogène

1. Une heure de parole du corpus de test de la campagne d'évaluation ESTER 2005 contient environ 30000 mots.

du document à transcrire. Le système de reconnaissance automatique de la parole contient généralement un étage de segmentation qui fournit ce découpage afin de contrôler la consommation de ressources computationnelles du décodeur. Un segment ne couvre généralement pas plus de 30 secondes de parole et la moyenne est autour de 15 secondes [2].

5.2.2 Augmentation contextuelle

La recherche de mots candidats que nous proposons s'appuie sur le contexte de chaque mots hors-vocabulaires. Il est ainsi possible de modéliser plus précisément celui-ci et donc de collecter une liste de candidats plus précise qu'avec, par exemple, une méthode reposant sur la thématique globale.

5.3 Extraction de requêtes caractéristiques des mots Hors-vocabulaires

Toute la difficulté de l'approche proposée réside dans la formulation de requêtes Web qui permettent de cibler des documents dans lesquels les mots hors-vocabulaires recherchés apparaissent. Il est nécessaire de bien connaître les possibilités des moteurs de recherche pour pouvoir développer des requêtes efficaces.

5.3.1 Les moteurs de recherche Web

La plupart des moteurs de recherche Web permettent deux types de requêtes :

- rechercher par séquences de mots : la requête contient une suite de mots et le moteur de recherche va retourner les documents contenant cette suite de mots exacte. Il y a la possibilité d'utiliser un ou plusieurs jokers, qui autorisent l'insertion de quelques mots (en général de 1 a 5) dans la séquence spécifiée.
- rechercher par sac de mots : la requête contient une liste de mots et le moteur de recherche va retourner les documents qui contiennent le plus grand nombre de ces termes, quelque soient leurs positions relatives dans les documents.

Il est aussi possible de mélanger les deux types de requêtes pour rechercher des documents qui contiennent une séquence de mots donnée et qui contiennent également d'autres mots, indifféremment de leur position par rapport à la séquence de mots. Il est aussi possible de rechercher des documents qui contiennent plusieurs séquences de mots quelque soit leur positions relatives.

A partir de ces possibilités de recherche dans les documents Web, plusieurs stratégies pour construire les requêtes sont possibles.

2. Par exemple avec la segmentation automatique du système SPEERAL du LIA sur le corpus de test de la campagne d'évaluation ESTER 2005

5.3.2 Stratégie *n*-gramme

La manière la plus intuitive de récupérer les mots hors-vocabulaires sur le Web consiste à récupérer les mots qui apparaissent dans exactement le même contexte que celui de la transcription. Pour récupérer de tels mots candidats sur le Web on peut construire des requêtes comme suit :

1. On construit une requête avec la séquence de mots (*n*-grammes) contenant le mot hors-vocabulaire détecté
2. On remplace le mot hors-vocabulaire par un joker pour qu'il puisse être substitué à n'importe quel mot dans les documents retournés par le moteur de recherche.

Par exemple, pour la transcription "[...] Les otages Christian *chaîne aux* et Georges [...]", avec *chaîne aux* les mots qui se sont substitués au mot hors-vocabulaire *Chesnot*, la requête de longueur 3 construite suivant cette stratégie serait "otages Christian (*)".

La liste des mots candidats pour le mot hors-vocabulaire en question sont les mots absents du lexique du modèle de langage initial qui seront substitués au joker, "(*)" dans l'exemple précédent, dans les documents retournés par le moteur de recherche auquel la requête aura été soumise.

5.3.3 Stratégie patrons

Les requêtes construites avec la stratégie *n*-gramme décrite précédemment sont très contraintes par le contexte d'apparition des mots hors-vocabulaires. Cependant, dans le cadre de transcriptions issues de reconnaissance automatique de la parole, ce contexte peut être erroné, ce qui a pour effet de diminuer l'efficacité de ces requêtes. Il est également possible que le mot hors-vocabulaire cherché n'apparaisse sur le Web que dans un contexte légèrement différent de celui de la transcription et ne pourrait donc pas non plus être retrouvé avec cette technique.

Pour pallier ces inconvénients, nous proposons de relâcher les contraintes sur le contexte capturé par les requêtes en autorisant des insertions et des substitutions sur les mots moins importants : les mots outils de la langue. Par la suite, nous appelons ces contextes souples des patrons.

Nous proposons de construire de tels patrons de la manière suivante :

1. Les mots-outils de la langue sont supprimés de l'hypothèse de transcription
2. Un jocker est inséré entre chaque mot
3. Les requêtes sont ensuite extraites de cette transcription modifiée de la même manière que pour la stratégie *n*-gramme

On peut noter que les jokers ne comptent pas dans le calcul de la taille de la requête car ils peuvent ne correspondre à aucun mot dans les documents retournés par le moteur de recherche.

La suppression des mots outils de la langue est effectuée à l'aide d'une anti-liste qui contient tous ces mots. De telles listes se trouvent facilement sur internet [3]. Elles sont généralement construites sur un critère statistique : les n mots les plus fréquents d'un corpus très grand. Le moteur de recherche auquel la requête est soumise fera correspondre chaque joker introduit à un maximum de cinq mots.

La méthode que nous venons de présenter génère des requêtes génériques qui nous permettent de récupérer sur le Web des séquences de mots qui ne sont pas exactement celles observées dans le contexte d'apparition des mots outils, mais qui conservent néanmoins l'ordre des mots importants.

Par exemple, pour la transcription "[...] Les otages Christian *chaîne aux* et Georges [...]", avec *chaîne aux* les mots qui se sont substitués au mot hors-vocabulaire *Chesnot*, la requête de longueur 3 construite suivant cette stratégie serait "otages * Christian (*)".

La liste des mots candidats pour le mot hors-vocabulaire en question sont les mots absents du lexique du modèle de langage initial qui seront substitués au joker collecteur, "(*)" dans l'exemple précédent, dans les documents retournés par le moteur de recherche auquel la requête aura été soumise.

5.3.4 Stratégie basée sur la sémantique à court terme

L'idée de cette stratégie est de construire des requêtes en relâchant encore plus les contraintes syntaxiques sur le contexte d'apparition du mot hors-vocabulaire. Les documents que l'on souhaite récupérer sur le Web doivent contenir le plus de mots possible en commun avec ceux du contexte du mot hors-vocabulaire en question, sans aucune restriction quant à leur position relatives dans ces documents.

Pour réaliser cela, des mots pertinents sont extraits du contexte des mots hors-vocabulaires et une recherche de type "sac de mots" est effectuée pour éliminer les contraintes syntaxiques. Afin d'aider le moteur de recherche à cibler les documents pertinents, seul les mots les plus discriminants du contexte servent à construire les requêtes.

La procédure de construction d'une requête est la suivante :

1. Les mots-outils de la langue sont supprimés de l'hypothèse de transcription
2. Les mots qui se trouvent dans une fenêtre de quelques mots autour du mot hors-vocabulaire sont sélectionnés
3. Ces mots sont triés par ordre décroissant de fréquence dans un gros corpus textuel
4. Les n premiers mots de la liste sont sélectionnés pour construire la requête "sac de mots' de taille n

3. Le site du stemmer Snowball (http://snowball.tartarus.org) contient des anti-listes de mots outils pour plus de dix langues.

Par exemple, pour la transcription "[...] Les otages Christian *chaîne aux* et Georges [...]", avec *chaîne aux* les mots qui se sont substitués au mot hors-vocabulaire *Chesnot*, la requête de longueur 3 contiendrait les mots clés suivante : "otages", "Christian" et "Georges".

La liste des mots candidats pour le mot hors-vocabulaire en question sont tous les mots absents du lexique du modèle de langage initial qui se trouvent dans les documents retournés par le moteur de recherche auquel la requête aura été soumise. Cette approche est donc censée récupérer une liste de candidats bien plus importante que les autres stratégies.

5.3.5 Stratégie *n*-gramme et patrons guidée par la sémantique

La plupart des requêtes *n*-gramme ou patron retournent plusieurs centaines voir plusieurs milliers de résultats. Par exemple, les requêtes 3-gramme du corpus de test de la campagne d'évaluation ESTER permettent de récupérer en moyenne 16 000 documents chacune.

Pour des raisons techniques, il n'est pas possible d'analyser tous les documents retournés par le moteur de recherche pour une requête donnée. Par conséquent il faut faire en sorte que les documents ayant le plus de chances de contenir les mots hors-vocabulaires cherchés se trouvent en tête de la liste de résultats.

En ce qui concerne les requêtes basées sur des séquences de mots ou des patrons, une solution peut consister à mieux cibler le contexte thématique local d'apparition des mots hors-vocabulaires. Pour cela, on peut ajouter à ces requêtes des mots qui sont représentatifs de la thématique, mais sans contrainte de positionnement. On obtient ainsi, en tête de la liste des documents résultats, ceux qui contiennent des séquences de mots correspondant aux requêtes initiales, mais qui, en plus, contiennent les mots porteurs de la thématique qui y ont été ajoutés. Nous appellerons mot pilotes les mots qui servent à orienter les résultats de la recherche vers des documents porteur de la thématique locale des mots hors-vocabulaires.

Nous proposons de sélectionner les mots pilotes de la même manière que nous sélectionnons les mots porteurs de sens pour la méthode basée sur la sémantique à court terme décrite à la section 5.3.4. Il faut noter que les mots qui composent les requêtes à guider, qu'elles reposent sur des *n*-grammes ou sur des patrons, sont retirés de la liste des mots-pilote.

Par exemple, pour l'hypothèse de transcription "[...] Les otages Christian *chaîne aux* et Georges [...]", avec *chaîne aux* les mots qui se sont substitués au mot hors-vocabulaire *Chesnot*, la requête *n*-gramme de longueur 3 accompagnée d'un mot pilote serait : "otages Christian (*)" + Georges. Pour la stratégie reposant sur les patrons, cette requête serait : "otages * Christian (*)" + Georges.

Pour ces deux stratégies de formulation de requête, les mots candidats sont ceux qui se substituent au joker collecteur dans les documents résultat, "(*)" dans les exemples précédents, et qui sont absents du lexique du modèle de langage initial.

5.4 Injection des nouveaux mots dans le processus de transcription

Les stratégies de construction de requêtes présentées à la section 5.3 permettent d'obtenir une liste de mots candidats pour chaque mot hors-vocabulaire détecté dans l'hypothèse de transcription. Plusieurs solutions sont envisageables pour insérer ces mots dans le processus de transcription.

5.4.1 Substitution dans la transcription

Si la position précise du mot hors-vocabulaire est connue, il est possible de sélectionner le bon mot dans la liste de candidats et de l'insérer dans la transcription à la place des mots qui se sont substitués au mot HV. La sélection du mot peut se faire selon la distance d'édition entre sa phonétisation et la transcription phonétique de la portion de signal correspondant au mot HV. La phonétisation des candidats peut être obtenue à l'aide d'un phonétiseur automatique.

Cette solution présente l'avantage d'être simple et rapide. Elle comporte par contre un inconvénient majeur : il est supposé que la position exacte du mot hors-vocabulaire dans le signal est connue, ce qui est techniquement difficile. Les détecteurs de mots hors-vocabulaires n'obtiennent des performances raisonnables qu'à partir d'une imprécision sur les frontières des mots hors-vocabulaires d'une demi seconde (Lecouteux et al., 2009), ce qui suffit à prononcer entre un et deux mots en français [4]. Dans une telle situation, la distance phonétique permettant de choisir un candidat ne peut fonctionner de manière pertinente.

5.4.2 Insertion des nouveaux mots dans le lexique

Une autre solution consiste à adapter le système de reconnaissance automatique de la parole plutôt que de corriger la transcription. Le principe est d'ajouter au lexique et au modèle de langage initial du système de reconnaissance automatique de la parole les mots candidats. Pour chaque segment donné du document audio dans lesquels des mots hors-vocabulaires ont été détectés, les candidats leurs correspondant sont inséré dans le lexique du système initial et une seconde passe de décodage est effectuée avec celui-ci. Le décodeur se chargera alors de trouver le mot candidat correspondant le mieux d'un point de vue acoustique et linguistique au contexte d'apparition du mot hors-vocabulaire.

L'avantage de cette solution est qu'elle ne nécessite pas de connaître avec précision les frontières des mots hors-vocabulaires. De plus, le choix du candidat est alors effectué en tenant compte de la correspondance linguistique en plus de l'acoustique.

4. Le débit de mots observé sur le corpus de test de la campagne d'évaluation ESTER est en moyenne de 3,3 mots par seconde.

Comme nous l'avons vu au chapitre 1, pour qu'un mot puisse être utilisé lors du processus de décodage, il est nécessaire de l'insérer dans le lexique avec sa prononciation et de lui donner une probabilité non nulle dans le modèle de langage.

La prononciation des candidats est obtenue ici à l'aide d'un phonétiseur automatique.

Comme nous l'avons vu à la section 4.4.4, il existe trois méthodes principales pour attribuer une probabilité à de nouveaux mots dans le modèle de langage : l'interpolation linéaire, la probabilité du mot inconnu et la probabilité de la classe de mots.

L'interpolation linéaire nécessite un corpus conséquent où les nouveaux mots sont bien représentés et qui est peu bruité. Dans notre cas nous ne contrôlons pas la nature des documents retournés par le moteur de recherche Web, il est donc difficile de s'assurer que le contenu linguistique de ces documents corresponde bien à celui des documents à transcrire. De plus, étant donné la nature hétérogène des documents, leur nettoyage est ardu. Ces difficultés rendent la collecte d'un corpus adéquat difficile, c'est pour cette raison que nous n'utiliserons pas cette approche.

Les deux autres approches sont, en fait, très similaire puisque qu'elles reposent sur un modèle de classes de mots. Dans l'approche par classe, chaque mot appartient à une ou plusieurs classe. La probabilité d'apparition du mot est alors déterminée en fonction des probabilités des classes auxquelles il appartient. L'approche par classe de mots inconnus est un cas particulier de l'approche par classe dans lequel il n'y à qu'une classe : celle des mots inconnus. Dans ce contexte, l'intégration d'un mot au modèle nécessite simplement de connaître les classes auxquelles il appartient. On s'affranchit donc du de la nécessité d'avoir un corpus d'adaptation conséquent comme avec la première approche.

Probabilité du mot inconnu

Avec cette stratégie d'intégration des nouveaux mots, la probabilité de chaque mot inséré dans le lexique est dérivée de la probabilité du mot inconnu.

Étant donné que la probabilité du mot inconnu est une probabilité de classe, il est nécessaire de la remettre à l'échelle du mot. La probabilité d'un nouveau mot est alors la probabilité de classe divisée par le nombre de mots qui appartiennent à cette classe dans le corpus d'entraînement du modèle de langage, ce qui revient à considérer que le nouveau mot est apparu une fois à l'échelle du corpus d'entraînement. Cette approximation est justifiée pour les mots inconnus car le fait qu'ils ne soient pas présents dans le lexique indique qu'ils appartiennent à la queue de la loi de Zipf.

La probabilité d'un nouveau mot w appartenant à l'ensemble N des candidats ajoutés au lexique est alors :

$$P(w|h) = \frac{1}{|U|} \times P(U|h) \tag{5.1}$$

avec $P(U|h)$ la probabilité d'apparition d'un mot de la classe U étant donné l'historique h et $|U|$ le nombre de mots appartenant à la classe U dans le corpus d'entraînement.

Si le nombre de mots insérés est différent du nombre de mots appartenant à la classe des mots inconnus dans le corpus d'entraînement du modèle de langage par classe, alors la probabilité des mots insérés ne sommera pas à 1. Si le nombre de mot est plus petit, la somme sera inférieure à 1 et sinon elle sera supérieure à 1.

Dans le premier cas, la masse de probabilité manquante sera fournie par la classe du mot inconnu. Elle modélisera alors l'apparition d'un mot inconnu autre que ceux ajoutés au lexique. La probabilité de la classe devient alors :

$$P'(U|h) = 1 - (\sum_{w \in N} P(w|h)) \tag{5.2}$$

avec N l'ensemble des mots ajoutés au lexique.

Si par contre on insère plus de mots dans le lexique qu'il n'y avait de mots inconnus dans le corpus d'entraînement du modèle de langage, il est possible de modifier la formule donnant la probabilité d'un nouveau mot pour que la distribution soit normale. On peut par exemple opter pour la formule suivante :

$$P(w|h) = \frac{1}{|N|} \times P(U|h) \tag{5.3}$$

Probabilité de la classe de mots

Cette stratégie consiste à attribuer une probabilité aux nouveaux mots à partir de la probabilité de la classe à laquelle ils appartiennent. Comme précédemment, on considère que chaque nouveau mot serait apparu une fois dans un corpus comparable à celui qui a servi à l'estimation du modèle de langage par classes. Contrairement à l'approche précédente, un mot peut appartenir ici à plusieurs classes. C'est pour cela que la probabilité accordée à un nouveau mot est pondérée par la probabilité que ce mot appartienne à la classe.

La probabilité d'un nouveau mot w dont la probabilité d'appartenir à la classe C est $P(w|C)$ s'écrit alors :

$$P(w|h) = \sum_{C} P(w|C) \times \frac{1}{|C|} P(C|h) \tag{5.4}$$

avec $P(C|h)$ la probabilité d'apparition d'un mot de la classe C étant donné l'historique h et $|C|$ le nombre de mots appartenant à la classe C dans le corpus d'entraînement.

Comme avec la méthode précédente, si le nombre de mots insérés appartenant à une classe donnée, pondéré par la probabilité d'appartenance des mots à la classe, est différent du nombre de mots appartenant à cette classe dans le corpus d'entraînement

du modèle de langage par classe, alors la probabilité des mots insérés de cette classe ne sommera pas à 1. Si le nombre de mot est plus petit, la somme sera inférieure à 1 et sinon elle sera supérieure à 1.

Ces deux cas de figure peuvent se traiter de la même manière que pour la méthode précédente. Si le nombre de mot est plus petit, le manque de probabilité peut être compensé par la classe du mot considéré. Elle modélisera alors l'apparition d'un mot inconnu appartenant à cette classe et qui n'est pas un de ceux ajoutés au lexique. La probabilité de la classe devient alors :

$$P'(C|h) = 1 - (\sum_{w \in N} P(w|C) \times \frac{1}{|C|} P(C|h)) \tag{5.5}$$

avec N l'ensemble des mots ajoutés au lexique.

Si par contre le nombre de mots insérés est plus grand que le nombre de mots inconnus de la classe C dans le corpus d'entraînement du modèle de langage, il est possible de modifier la formule donnant la probabilité d'un nouveau mot pour que la distribution soit normale. La formule devient alors :

$$P(w|h) = \sum_{C} P(w|C) \times \frac{1}{\sum_{w \in N} P(w|C)} P(C|h) \tag{5.6}$$

Nous avons vu à la section 2.4 que l'ensemble de classes donnant les meilleures performances est déterminé automatiquement de manière non supervisée. Déterminer à quelle classe appartiennent les nouveaux mots nécessite un corpus dans lequel ils apparaissent suffisamment. Dans notre cas, un tel corpus n'est pas disponible.

Les classes morphosyntaxiques, bien que légèrement moins performantes, ont l'avantage de pouvoir être déterminées avec peu de données. Dans notre cas, les phrases auxquelles les nouveaux mots appartiennent dans les documents Web suffisent. C'est donc cette solution qui sera retenue.

5.5 Dispositif expérimental

L'approche que nous proposons pour l'enrichissement local et dynamique du lexique est évaluée sur deux tâches très différentes, mais pour lesquelles les mots hors-vocabulaires posent un réel problème. La première est une tâche de transcription automatique de journaux radiodiffusés en français dans un but d'indexation. La seconde tâche consiste à transcrire automatiquement des commentaires audio de films d'opérations chirurgicales. Là encore il s'agit d'utiliser la transcription automatique pour indexer les documents afin qu'il soit rapidement accessibles.

Dans les deux cas, la présence de mots hors-vocabulaires entraîne un défaut systématique d'indexation sur ces termes. Étant donné que, comme nous l'avons vu à la

section 4.1.2, plus de 95% des mots hors-vocabulaires sont des noms propres ou des termes techniques, il est évident que ces lacunes auront des répercussions sur les performances du processus d'indexation et de recherche.

5.5.1 Les corpus d'évaluation

Pour simuler les deux tâches, nous avons utilisé différents corpus. Pour la tâche de transcription de journaux radiodiffusés, nous avons utilisé le corpus de la campagne d'évaluation ESTER. La tâche de transcription de films chirurgicaux a été simulée par le corpus AVISON.

Le corpus ESTER

Le corpus utilisé pour la tâche de transcription automatique de journaux radiodiffusés est celui de la campagne d'évaluation ESTER (Gravier et al., 2004; Galliano et al., 2005). Cette campagne a pour but dévaluer les performances des systèmes de transcription riches sur des journaux d'information radiodiffusés francophones. La campagne d'évaluation était divisée en trois sous-évaluations : la transcription orthographique, la détection d'événements (parole, musique, suivi des locuteurs, etc.) et finalement l'extraction d'informations (détection des entités nommées, etc.).

Les données audio fournies pour la tâche étaient divisées en 4 corpus :

- Le corpus d'entraînement non supervisé : 1600 heures d'émissions de radio francophones antérieures à octobre 2003 non transcrites.
- Le corpus d'entraînement : 82 heures d'émissions de radio francophones antérieures à octobre 2003 transcrites manuellement.
- Le corpus de développement : 8 heures d'émissions de radio francophones antérieures à octobre 2003 transcrites manuellement.
- Le corpus d'évaluation : 10 heures d'émissions de radio francophones enregistrées entre octobre et décembre 2003 et transcrites manuellement.

Ces données ont été enregistrées sur les stations de radio suivantes : France Inter, France Info, Radio France International (RFI), Radio Télévision Marocaine (RTM), France Culture et Radio Classique.

Pour l'entraînement des modèles de langage, un corpus d'articles de journaux était fourni. Il s'agissait des articles du journal français Le Monde sur la période de 1987 à 2003, ce qui représente environ 400M mots.

Le corpus AVISON

Le corpus AVISON a été créé afin d'évaluer les systèmes proposés dans le cadre du projet ANR AVISON [5]. Le but du projet est de développer des outils de navigation dans

5. Projet numéro ANR-07-RIAM-0903

le fond documentaire de l'Institut de Recherche contre les Cancer de l'Appareil Digestif et de l'European Institute of TeleSurgery (IRCAD-EITS), qui est composé de films d'opérations, d'avis d'experts et de cours dans le domaine de la chirurgie robotisée.

Les données collectés pour ce projet sont divisées en plusieurs corpus :

1. Un corpus d'entraînement composé d'environ 20 heures de vidéos pour lesquelles une transcription approchée est disponible
2. Un corpus de test composé de 4 heures de vidéos pour lesquelles une transcription manuelle est disponible

Ce corpus comporte différents styles de parole (non étiquetés) : descriptions du déroulement d'opérations enregistrées en studio, descriptions en parole spontanées faites par les praticiens dans l'action ou encore dialogue entre chirurgiens et étudiants.

Pour l'entraînement des modèles de langage, les transcriptions approchées des 20 heures du corpus audio représentent 600k mots. Un corpus supplémentaire de 350k mots composé de rapports techniques est également disponible.

5.5.2 Le système de reconnaissance automatique de la parole

Le système de reconnaissance automatique de la parole continue grand vocabulaire SPEERAL a été utilisé pour mettre en pratique les différentes approches proposées ici sur les deux tâches décrites précédemment. Ce système de reconnaissance automatique de la parole statistique repose sur un algorithme A* pour le décodage, une modélisation n-gramme du langage et sur des modèles de Markov cachés contextuels à état partagés pour la modélisation acoustique. Les variantes de prononciation d'un mot du lexique ne sont pas probabilisées. Il est développé au Laboratoire Informatique d'Avignon et est décrit plus en détails par Nocera et al. (2002a), Nocera et al. (2002b) et Nocéra et al. (2004).

Le système français journalistique

Le système pour le français utilise les ressources linguistiques de la campagne d'évaluation ESTER décrites précédemment. Le modèle de langage est un 3-gramme obtenu par combinaison linéaire de deux modèles : un 3-gramme estimé sur les 400M de mots du journal *Le Monde* et un autre estimé sur environ 1M de mots du corpus d'entraînement acoustique de la campagne d'évaluation ESTER constitué de transcriptions manuelles de journaux radiodiffusés. Le poids de chaque modèle dans la combinaison est choisi pour maximiser la perplexité du modèle résultant sur le corpus de développement de la campagne.

Les mots du lexique sont phonétisés à l'aide du phonétiseur automatique à base de règles LIA_PHON [6] (Bechet, 2001) et sont vérifiées à la main.

6. http://pageperso.lif.univ-mrs.fr/~frederic.bechet

Le modèle de langage par classes morphosyntaxiques de mots inconnus, comme décrit à la section 5.4.2, est construit en étiquetant les mots inconnus du corpus d'entraînement du modèle de langage avec l'outil LIA_TAGG [6]. Cet outil fournit un jeu d'étiquettes morphosyntaxiques très détaillé.

Les modèles acoustiques sont des modèles de Markov cachés *gauche-droite* à trois états comme décrit a la section 1.3. Les unités acoustiques sont des triphones et les états des modèles sont partagés pour réduire la complexité de l'apprentissage. Les 82 heures d'enregistrement démissions annotées du corpus ESTER ont été utilisées pour l'estimation de ces modèles.

Dans les conditions de test de la campagne d'évaluation ESTER, ce système obtient un taux d'erreur mot d'environ 20%.

Le système anglais chirurgical

La modélisation acoustique du système anglais est effectuée de manière similaire à celle du système français. Le jeu de phonèmes est celui utilisé par le *CMU Pronouncing Dictionary*, composé de 39 phonèmes [7]. Pour l'estimation des modèles, nous ne disposions pas de corpus représentatifs des conditions acoustiques des documents à transcrire suffisant pour estimer les modèles acoustiques. Ils ont donc été estimés à partir du corpus le plus proche dont nous disposions. Il s'agit du corpus d'entraînement de la campagne d'évaluation HUB4 (Stern, 1997), qui contient des enregistrements d'émissions d'information américaines transcrits manuellement.

Ces modèles sont ensuite adaptés au corpus AVISON grâce à quelques itérations de l'algorithme *Maximum A Posteriori* décrit dans le chapitre 1. Les données d'adaptation sont issues de l'alignement par décodage guidé des transcriptions imparfaites des 20h du corpus d'entraînement d'AVISON, ce qui a permis d'obtenir 12h de corpus aligné (Lecouteux et al., 2011).

Le lexique est constitué de tous les mots présents plus d'une fois dans le corpus d'entraînement d'AVISON (15k mots) et des 30k mots les plus fréquents du corpus d'entraînement de HUB4. La prononciation de ces mots est celle du *CMU Pronouncing Dictionary* si elle existe et sinon elle est obtenue à l'aide du phonétiseur automatique anglais Festival (Taylor et al., 1998).

Le modèle de langage est un 3-gramme obtenu par la combinaison de plusieurs modèles estimés sur des corpus de nature différentes. Le poids des modèles est déterminé pour maximiser la perplexité du modèle final sur un corpus de développement. Quatre sources de données sont utilisées pour estimer les modèles atomiques. La première est constituée des transcriptions approchées disponibles dans le corpus AVISON. La seconde consiste en l'ensemble des documents techniques disponibles dans le corpus AVISON. Le troisième est composée des transcriptions d'émissions américaines disponibles dans le corpus HUB4. La dernière regroupe l'ensemble des articles de jour-

7. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

naux d'information américains disponibles dans le corpus North American News Text Corpus (Graff, 1995), fourni pour la campagne d'évaluation HUB4.

Sur les données de test du corpus AVISON, ce système permet d'obtenir un taux d'erreur mot de 48,7%. Cette valeur élevée peut s'expliquer par trois éléments.

Le premier est que le corpus de test possède un grand nombre de mots hors-vocabulaires pour le lexique décrit précédemment. 6% des mots du corpus de test sont hors-vocabulaires, or on sait que chaque mot hors-vocabulaire produit environ deux erreurs de transcription (Gauvain et al., 1995), ce qui représente environ 12% d'erreurs uniquement à cause des mots hors-vocabulaires.

Le second point faible du système est le modèle de langage. L'ensemble des données d'entraînement qui correspondent à la nature des données de test ne représentent qu'environ 600k mots. Il est évident que dans de telles conditions il n'est pas possible d'obtenir des performances semblables à celles obtenues sur le corpus ESTER où environ 5M mots de transcription sont disponibles et où 400M mots de textes écrits de même nature (le journal *Le Monde*) sont disponibles.

Le dernier point faible du système est le modèle acoustique. Ils ont été estimés à partir d'un corpus dont les conditions acoustiques sont assez éloignées de celles du corpus AVISON et les données alignées automatiquement utilisées pour l'adaptation ne représentent qu'une dizaine d'heures et contiennent quelques erreurs de transcription.

Il est intéressant de noter que ce dispositif reproduit bien le cas réel où l'on doit indexer des documents audio relatifs à un domaine pointu. Dans une telle situation on dispose de peu de données d'entraînement et donc d'un système de reconnaissance automatique de la parole mal entraîné. L'approche proposée dans ce chapitre est alors utilisée pour récupérer automatiquement les mots hors-vocabulaires sur le Web afin d'améliorer l'indexation malgré de nombreuses erreurs de reconnaissance automatique de la parole.

5.5.3 Détection des mots hors-vocabulaire

Le processus de détection des mots hors-vocabulaires est ici effectué manuellement. Dans les transcriptions produites pour les deux corpus, les mots hors-vocabulaires sont identifiés manuellement afin de simuler un détecteur de mot hors-vocabulaire parfait. On notera que l'état de l'art dans le domaine de la détection des mots hors-vocabulaires permet d'obtenir un taux d'égale erreur (*Equal Error Rate*) supérieur à 90% (Lecouteux et al., 2009).

5.5.4 L'augmentation lexicale

Le processus d'augmentation lexicale utilise, pour les expériences présentées ici, le moteur de recherche *Google* [8].

8. http://www.google.com

Recherche des candidats

Pour chaque mot hors-vocabulaire rencontré dans la transcription, une requête est générée et soumise au moteur de recherche. Les 100 premier résultats sont alors utilisés pour construire la liste des candidats. La manière dont les candidats sont sélectionnés dépend de la méthode utilisée pour construire la requête.

Injection des candidats

Chaque candidat de la liste est phonétisé automatiquement à l'aide des outils LIA_PHON (Bechet, 2001) pour le français et de Festival (Taylor et al., 1998) pour l'anglais.

Pour insérer les mots candidats dans le modèle de langage, il est nécessaire de leur attribuer une probabilité. Nous avons vu qu'il est possible d'utiliser un modèle de langage de classes morphosyntaxiques. Il est donc nécessaire d'attribuer de telles classes aux mots candidats.

La liste de mots candidats est en fait une liste de graphies. Une même graphie peut être associée plusieurs mots et chacun peut avoir plusieurs fonctions dans une phrase. Une graphie donnée sera donc associée à plusieurs classes morphosyntaxiques en fonction du contexte dans lequel elle apparaît. Dans l'approche proposée ici, les graphies de la liste de candidats sont en fait associées à un seul mot, celui qui a été détecté comme hors-vocabulaire, et à un seul contexte, celui d'apparition du mot hors-vocabulaire. Il s'agit donc ici de trouver la classe morphosyntaxique qui correspond au mot hors-vocabulaire dans son contexte parmi toutes les classes morphosyntaxiques qui peuvent être associées aux différentes graphies de la liste des candidats.

L'approche que nous proposons pour résoudre ce problème est la suivante. Pour chaque mot de la liste des candidats, l'ensemble des phrases des documents Web correspondant au patron de transcription ayant servi à construire la requête sont extraites. De cette manière, on s'assure que seul les contextes correspondants à celui de la transcription sont sélectionnés. Toutes les phrases extraites sont étiquetées à l'aide du logiciel LIA_TAGG. Il est alors possible qu'il soit associé à un mot candidat plusieurs classes morphosyntaxiques qui sont tout à fait légitimes car le contexte de la transcription a été préservé lors de l'extraction des phrases.

On calcule, pour chaque candidat, la probabilité d'appartenance à chaque classe morphosyntaxique de la manière suivante :

$$P(w|C) = \frac{N(w, w \in C)}{N(w)} \tag{5.7}$$

avec $N(w, w \in C)$ le nombre d'occurrences du mot w associé à la classe C et $N(w)$ le nombre total d'occurrences de w.

Cette probabilité est utilisée dans la formule de calcul de la probabilité linguistique du mot avec le modèle de langage par classe décrit à la section 5.4.2.

5.6 Experimentations

Les techniques d'enrichissement dynamique du lexique été évaluées sur les deux corpus présentés, ESTER et AVISON, dans les conditions décrites précédemment. Dans un premier temps nous allons étudier le potentiel du Web dans ce contexte précis. Ensuite nous analyserons en détails les performances des différentes techniques de formulation de requête et d'extraction de mots candidats. Finalement nous présenterons les résultats obtenus après intégration de ces nouveaux mots dans le processus de reconnaissance automatique de la parole.

5.6.1 L'importance du moteur de recherche

L'approche proposée pour récupérer les mots hors-vocabulaires absent de la transcription s'appuie sur un moteur de recherche pour accéder aux données du Web. Il est évident que la taille de l'index du moteur de recherche, la manière dont il est constitué ainsi que la nature des documents qu'il référence influence les performances de la méthode.

Afin d'étudier les performances des moteurs de recherche pour la tâche de récupération des mots hors-vocabulaires, nous allons effectuer un ensemble de requêtes afin de vérifier que ces mots sont bien présents dans leur index. On vérifiera également qu'ils apparaissent dans le même contexte que celui de la transcription.

Les requêtes sont construites à partir des transcriptions de référence afin de s'affranchir des éventuelles erreurs de reconnaissance automatique de la parole. Les requêtes sont construites suivant la stratégie n-gramme et le mot hors-vocabulaire recherché remplace le joker collecteur de mots candidats. Cette technique revient à soumettre aux moteurs de recherche l'ensemble des n-grammes contenant des mots hors-vocabulaires dans la transcription de référence.

Le tableau 5.1 contient les résultats obtenus pour deux des moteurs de recherches Web les plus connus, Google [9] et Yahoo [10], avec des requêtes n-grammes d'ordre compris entre 1 et 5. Une requête d'ordre 1 est en fait constituée uniquement du mot hors-vocabulaire et permet de s'assurer de sa présence sur le Web quelque soit le contexte. Le tableau contient les taux de requêtes ayant renvoyé au moins un résultat. Les corpus ESTER et AVISON ont été testés afin d'étudier le comportement des moteurs de recherche lorsqu'ils sont confrontés à des données de nature différentes (des informations pour le corpus ESTER et des données spécialisées pour le corpus AVISON).

9. http://www.google.com
10. http://www.yahoo.com

	AVISON			ESTER		
n	**Yahoo**	**Google**	**Δ**	**Yahoo**	**Google**	**Δ**
1	99.6	100	0.4	100	100	0.0
2	95.8	98.8	3.0	91.8	88.2	3.6
3	77.8	87.1	9.3	63.5	50.5	13.0
4	45.2	56.6	11.4	37.8	27.3	10.5
5	19.2	32.0	12.8	22.8	16.1	6.7

TABLE *5.1* – *Performances des moteurs de recherche Yahoo et Google pour la récupération des mots hors-vocabulaire*

La première chose que l'on remarque est que toutes les requêtes d'ordre 1 ramènent des documents, ce qui indique que tous les mots hors-vocabulaires recherchés sont présents sur le Web. La seule exception est pour le corpus AVISON avec le moteur de recherche Yahoo, où quelques mots hors-vocabulaires ne sont pas présents. Ces résultats indiquent que les mots hors-vocabulaires sont globalement tous présents sur le Web quelque soit le moteur de recherche. Il suffit donc de formuler correctement les requêtes pour parvenir à les récupérer.

Lorsque l'ordre des requêtes augmente, leur pouvoir de filtrage contextuel augmente également et on constate une baisse du nombre de mots hors-vocabulaires qu'elles permettent de récupérer. On en déduit que malgré le fait que tous les mots hors-vocabulaires soient présents sur le Web, ils n'apparaissent pas toujours dans le même contexte que celui de la transcription.

Lorsque l'ordre des requêtes augmente, les performances obtenues avec le corpus de spécialité AVISON sont meilleures que celles obtenues avec le corpus d'informations ESTER. Ce résultat est contre intuitif car il est légitime de supposer que le Web contient moins de documents relatifs à la chirurgie robotisée que de documents d'informations. Cependant, étant donné que tous les mots hors-vocabulaires, ceux d'AVISON comme ceux d'ESTER, sont présents sur le Web, la différence de performance lorsque l'ordre des requêtes augmente indique simplement que pour AVISON, les mots hors-vocabulaires apparaissent plus souvent dans le même contexte que celui de la transcription que pour le corpus ESTER. Cela est probablement dû au fait que les données d'AVISON contiennent beaucoup de mots de vocabulaire techniques relatifs à la chirurgie. En effet, une grande partie des mots hors-vocabulaires de ce corpus sont des termes techniques qui apparaissent dans des expressions répondant aux normes linguistiques du domaine. A l'inverse, dans le corpus ESTER la majorité des mots hors-vocabulaires sont des noms propres qui peuvent être utilisés dans des contextes variés.

En comparant les résultats obtenus en utilisant les deux moteurs de recherche, on constate que Google est plus performant pour les données de spécialité du corpus AVISON, alors que Yahoo donne de meilleurs résultats pour les données d'information provenant du corpus ESTER. Cette différence de performances est probablement due aux décisions prises concernant la sélection des documents à indexer. Comme nous l'avons vu au chapitre 3, les moteurs de recherche Yahoo et Google ne partagent qu'environ 56% de leur index, ce qui indique qu'ils ont des stratégies de collecte de docu-

ments différentes. Il est alors possible que le moteur de recherche Yahoo ait optimisé l'indexation des documents d'information. Un autre phénomène qui entre en jeux est la langue, puisque les corpus ESTER et AVISON sont respectivement en français et en anglais. Il est donc possible que Yahoo ait privilégié les documents francophones.

Ces expériences montrent que le choix du moteur de recherche est important pour obtenir de bonnes performances. Il ne semble donc pas exister de moteur de recherche idéal, l'important est qu'il soit adapté à la nature des documents à traiter. Malheureusement, les expériences montrent également qu'il est très difficile de déterminer *a priori* le meilleur moteur car leur fonctionnement est obscur et leurs stratégies d'indexation sont souvent secrètes. Enfin, pour des contextes suffisamment discriminant ($n > 2$), les expériences montrent que le choix du moteur de recherche n'a d'impact que sur environ 10% des mots hors-vocabulaires.

5.6.2 Performances des requêtes

Les différentes stratégies de construction de requêtes Web à partir des emplacements des mots hors-vocabulaires dans la transcription, présentées à la section 5.3, ont pour but de récupérer des documents contenant les mots hors-vocabulaires absent de la transcription.

Par exemple, pour la transcription "[...] Les otages Christian *chaîne aux* et Georges [...]", avec *chaîne aux* les mots qui se sont substitués au mot hors-vocabulaire *Chesnot*, la requête construite avec la stratégie n-gramme d'ordre 3 est "otages Christian (*)". Le but d'une telle requête est de récupérer des documents dans lesquels ce patron apparaît et dont le joker *(*)* est remplacé par le mot *Chesnot* : "otages Christian Chesnot".

La partie "otages Christian" de la requête précédente sert à contextualiser la recherche dans le but de filtrer les mauvais contextes pour cibler le mot manquant. Il faut que ce filtrage soit très discriminant afin qu'il n'y ait pas beaucoup de candidats pour une requête donnée. Il faut également faire attention à ce qu'il ne le soit pas trop car le mot cible pourrait être filtré s'il apparaît uniquement dans un contexte proche de celui de la transcription, mais pas identique.

Evaluation du rappel maximum

Afin de mesurer le potentiel de chacune des mesures, il est possible de sonder le Web de la même manière que précédemment : en générant des requêtes sur la transcription de référence avec les différentes stratégies et en remplaçant le joker collecteur par le mot hors-vocabulaire recherché. Si les requêtes renvoient au moins un document, c'est qu'elles permettent potentiellement de récupérer le mot hors-vocabulaire.

Les résultats de cette expérience sont présentés dans le tableau 5.2. Le moteur de recherche Yahoo a été utilisé pour effectuer les requêtes. Le tableau contient les taux de requêtes ayant renvoyé au moins un résultat.

	AVISON		ESTER	
n	***n*-grammes**	**patrons**	***n*-grammes**	**patrons**
1	99.6	99.6	100	100
2	95.8	86.0	91.8	81.1
3	77.8	26.9	63.5	33.2
4	45.2	2.5	37.8	13.2
5	19.2	0.3	22.8	4.9
6	7.3	0.2	13.1	2.0

TABLE 5.2 – *Performances des méthodes d'extraction de contexte, en utilisant le moteur de recherche Yahoo*

On constate que pour les deux stratégies étudiées, plus la taille des requêtes augmente, plus la recherche contextuelle est discriminante et moins elles permettent de récupérer de mots hors-vocabulaires. On remarque également que la stratégie reposant sur des patrons syntaxiques est plus discriminante que la stratégie basée sur les *n*-grammes. Ceci est dû au fait que dans un patron, tous les mots outils sont éliminés. Pour une taille de requête donnée, une requête basée sur les patrons contiendra alors plus de mots porteurs de sens qu'une requête *n*-gramme. Les requêtes *n*-grammes vont filtrer le contexte sur un critère plutôt syntaxique alors que les requêtes basées sur les patrons vont plutôt filtrer le contenu sémantique du contexte. On notera qu'une partie du filtrage syntaxique est conservé dans le patron puisque l'ordre des mots doit être respecté dans les documents résultat.

Potentiel des stratégies de base

L'expérience précédente montre quels sont les taux de rappels maximum que l'on peut obtenir avec les différentes stratégies de requêtes. Afin d'apprécier plus précisément les caractéristiques des différentes stratégies, il est intéressant de mesurer la précision et le rappel sur la transcription de référence.

Le rappel est ici le nombre de mots hors-vocabulaires que chaque stratégie permet de récupérer parmi les candidats en provenance du Web. La précision est la taille moyenne des listes de candidats.

Pour effectuer cette mesure, les différentes stratégies proposées sont appliquées aux mots hors-vocabulaires de la transcription de référence. Pour chaque stratégie, la procédure de construction de requêtes et de récolte des candidats décrite à la section 5.3 est respectée. Les listes de candidats produites sont ensuite analysées afin de déterminer le rappel et la précision de chaque méthode.

Le tableau 5.3 contient le rappel obtenu avec les différentes stratégies de collecte de mots candidats proposées sur les corpus AVISON et ESTER. Le tableau 5.4 contient le nombre moyen de mots dans les listes de candidats produites par les stratégies sur ces corpus.

	ESTER			AVISON		
n	***n*-grammes**	**patrons**	**sémantique**	***n*-grammes**	**patrons**	**sémantique**
2	14.0	20.0	32.6	18.1	21.1	68.5
3	18.1	20.3	39.7	20.5	18.7	61.2
4	16.4	17.5	45.9	21.4	6.5	55.6
5	13.8	12.3	50.2	14.1	1.5	23.5

TABLE 5.3 – *Taux de rappel des requêtes de récupération de mots hors-vocabulaires mesurés sur les 100 premiers documents retournés par Google à partir des transcriptions de référence*

	ESTER			AVISON		
n	***n*-grammes**	**patrons**	**sémantique**	***n*-grammes**	**patrons**	**sémantique**
2	145	411	16.0k	526	356	15.9k
3	49	139	19.0k	147	88	30.2k
4	13	34	37.9k	62	18	33.6k
5	4	15	44.9k	25	8	20.3k

TABLE 5.4 – *Nombre moyen de candidats par requête de récupération de mots hors-vocabulaires mesurés sur les 100 premiers documents retournés par Google, à partir de la transcription de référence*

D'après les résultats présentés dans les tableaux 5.3, on constate que le rappel de la méthode *n*-gramme et patrons est meilleur sur le corpus AVISON, relatif à un domaine de spécialité, que sur le corpus ESTER, contenant des données généralistes. On peut attribuer ce comportement au contenu linguistique du corpus AVISION, qui est plutôt technique et dont les structures linguistiques sont très codifiées et souvent récurrentes. Dans la plupart des phrases, les termes techniques ("forceps", "ciseaux", etc.) peuvent être interchangés. Il est donc plus probable que les mots hors-vocabulaires recherchés apparaissent sur le Web dans un contexte identique à celui de la transcription.

A l'inverse, le corpus ESTER est constitué de données moins spécifiques mais plus variées dans les thématiques et dans le style (parole spontanée ou non), ce qui implique une plus grande diversité dans les constructions linguistiques. Il est donc moins probable qu'un mot hors-vocabulaire apparaisse sur le Web dans un contexte identique à celui de la transcription.

On constate également que la stratégie reposant sur les patrons donne de meilleurs résultats que la stratégie *n*-gramme sur le corpus ESTER, alors qu'elle est légèrement moins performante que cette dernière sur le corpus AVISON. Ces résultats indiquent que le fait de relâcher les contraintes syntaxiques sur le contexte utilisé pour les requêtes, comme le fait la stratégie reposant sur des patrons, n'est pas nécessaire sur un corpus comme AVISON. Cette observations confirme l'hypothèse selon laquelle les structures linguistiques présentes dans le corpus AVISON sont récurrentes. Relâcher les contraintes syntaxiques diminue alors l'information contenue dans les requêtes.

Concernant la stratégie sémantique, on constate que le rappel obtenu sur le cor-

pus AVISON diminue lorsque la taille des requêtes augmente, alors que sur le corpus ESTER on observe une tendance inverse. Cela peut s'expliquer par la nature très spécifique du contenu du corpus AVISON. Les requêtes sémantiques sont constituées d'un ensemble de termes pertinents extraits du contexte des mots hors-vocabulaires. Les documents retournés par le moteur de recherche doivent contenir tous les termes présents dans les requêtes. Plus elles sont longues, plus les contraintes sur les document sont grandes, particulièrement sur le corpus AVISON où les termes pertinents sont très spécifiques et donc plutôt rares en général. Quand les requêtes sont trop longues ou contient trop de mots peu fréquents, le nombre de documents contenant tous les termes diminue fortement et donc un plus grand nombre de mots hors-vocabulaires sont manqués.

A l'inverse, les termes pertinents extraits du corpus ESTER sont moins spécifiques et plus fréquents sur le Web. Lorsque la taille des requêtes augmente, elles deviennent plus précises et permettent de mieux cibler les documents pertinents.

Bien que le rappel obtenu sur le corpus AVISON chute lorsque la taille des requêtes sémantiques augmente, on constate que le nombre de mots candidats collectés augmente. Ce comportement peut s'expliquer par le fait que les documents retournés par les requêtes sémantiques doivent contenir tous les termes qu'elles contiennent. Plus la taille des requêtes est grande ou plus les mots qu'elles contiennent sont rares, plus il est probable que seul des documents volumineux y correspondent.

Globalement, les meilleurs résultats, en terme de rappel, sont obtenus avec la stratégie sémantique, alors que les meilleurs résultats en terme de précision le sont avec les méthodes n-gramme et patrons. Il semble donc intéressant d'évaluer les stratégies hybride n-gramme et patrons guidées par la sémantique, qui permettrait de tirer le meilleur de chaque approche.

Potentiel des stratégies hybrides

Le tableau 5.5 contient le rappel et le nombre moyen de mots par liste de candidats pour la méthode hybride reposant sur des n-grammes guidés par des mots pilotes et le tableau 5.6 contient les résultats obtenus par la méthode hybride reposant sur des patrons guidés par des mots pilotes. Comme décrit dans la section 5.3, ces deux stratégies consistent à enrichir les requêtes de type n-gramme ou patron avec des mots pilotes. Il s'agit de mots porteurs de sens extraits du contexte des mots hors-vocabulaires qui sont ajoutés aux requêtes avec pour seul contrainte d'apparaître dans les documents retournés par le moteur de recherche.

Comme on pouvait s'y attendre, les taux de rappel obtenus avec les méthodes hybrides sont meilleurs que ceux obtenus avec les stratégies n-gramme et patron initiales, sur les deux corpus.

Cependant, on remarque que parfois les gains obtenus sont plus importants lorsque le nombre de mots pilotes ajoutés aux requêtes est faible. Les mots pilotes ajoutent

	ESTER		AVISON	
n/m	**rappel**	**candidats**	**rappel**	**candidats**
2/1	24.0	268	41.8	1.3k
2/2	26.1	789	57.8	3.9k
2/3	27.0	1.3k	60.5	5.8k
3/1	19.1	16	31.1	215
3/2	15.0	15	33.1	379
3/3	13.3	19	32.7	500

TABLE 5.5 – *Rappel et taille des liste de candidats pour la stratégie de requêtes n-gramme guidé par la sémantique, avec m guides sémantiques sur les 100 premiers documents retournés par Google, à partir de la transcription de référence*

	ESTER		AVISON	
n/m	**rappel**	**candidats**	**rappel**	**candidats**
2/1	23.6	299	31.6	360
2/2	24.8	328	38.1	539
2/3	22.3	397	38.9	625
3/1	18.6	93	17.0	89
3/2	14.9	76	15.4	79
3/3	13.0	66	12.7	88

TABLE 5.6 – *Rappel et taille des liste de candidats pour la stratégie de requêtes patron guidé par la sémantique, avec m guides sémantiques sur les 100 premiers documents retournés par Google, à partir de la transcription de référence*

des contraintes aux requêtes, ce qui réduit le nombre de documents qui leurs correspondent. Au delà d'un certain nombre de mots ajoutés les requêtes deviennent trop discriminantes et des documents pertinents sont manqués. On remarque d'ailleurs que le seuil de dégradation est plus bas pour les requêtes de base de taille 3 que pour des requêtes de base de taille 2 car ces requêtes contribuent également à contraindre la recherche.

Sur les deux corpus, la stratégie hybride reposant sur les patrons fournit de moins bons résultats que celle reposant sur les *n*-grammes. Ce résultat est probablement la conséquence de la forte densité de mots discriminants que contiennent les requêtes provenant de la stratégie patrons. Dans une telle situation, l'allongement des requêtes augmente fortement leur pouvoir discriminant.

Une autre conséquence de l'ajout de contraintes aux requête est l'augmentation des listes de candidats. En effet, plus une requête possède de contraintes, plus la taille moyenne des documents lui correspondant est grande.

On constate que pour le corpus AVISON, les gains en terme de rappel sont plus importants que ceux obtenus sur le corpus ESTER. Cela s'explique par la récurrence des structures syntaxiques employées dans le corpus AVISON. En effet, étant donné que les termes techniques sont souvent interchangeables, augmenter la taille du contexte des

requêtes *n*-gramme ne permet pas de mieux cibler les documents pertinents. Par contre, ajouter des contraintes sémantiques locales à ces requêtes par l'intermédiaire des mots pilotes permet un filtrage en faveur des documents pertinents.

Le rapport entre le rappel et le nombre moyen de candidats par mot hors-vocabulaire est bien meilleur avec les stratégies hybrides qu'avec les stratégies de base. On obtient des taux de rappel comparables à la stratégie sémantique, mais avec un nombre moyen de candidats par liste proche de celui des stratégies patrons et *n*-grammes. Les meilleurs performances sur les deux corpus sont obtenues avec la stratégie *n*-gramme guidée par la sémantique. Ces stratégies hybrides permettent donc de cumuler les avantages des stratégies reposant sur des descripteurs syntaxiques et sémantiques locaux.

5.6.3 Robustesse des requêtes

Les expériences précédentes ont permis d'évaluer le potentiel des différentes stratégies. Elles ont été conduites en utilisant la transcription de référence afin de s'affranchir des erreurs de transcription. Nous allons maintenant étudier la robustesse de ces approches aux erreurs de reconnaissance automatique de la parole en effectuant les mêmes expériences mais en utilisant la transcription automatique à la place de la transcription manuelle.

Performances des stratégies de base

Les tableaux 5.7 et 5.8 contiennent respectivement les taux de rappel et la taille moyenne des listes de candidats obtenus par les différentes stratégies présentées à la section 5.3 sur les transcriptions automatiques des corpus ESTER et AVISON.

	ESTER			AVISON		
n	***n*-grammes**	**patrons**	**sémantique**	***n*-grammes**	**patrons**	**sémantique**
2	4.7	7.3	18.5	15.1	15.3	55.4
3	5.1	5.0	27.8	11.6	9.9	44.4
4	2.3	2.0	35.2	9.2	3.8	35.1
5	1.9	1.2	40.9	4.1	1.0	24.1

TABLE 5.7 – *Taux de rappel des requêtes de récupération de mots hors-vocabulaires mesurés sur les 100 premiers documents retournés par Google, à partir de la transcription automatique*

On constate que les performances des différentes stratégies sont moins bonnes que celles obtenues à partir des transcriptions de référence. Cette dégradation systématique s'explique par le fait que toutes les stratégies proposées reposent sur l'extraction de descripteurs locaux, or la probabilité d'erreur de transcription autour des mots hors-vocabulaires est plus importante que dans le reste de la transcription car le processus de reconnaissance automatique de la parole est perturbé.

	ESTER			AVISON		
n	***n*-grammes**	**patrons**	**sémantique**	***n*-grammes**	**patrons**	**sémantique**
2	322	475	13.7k	439	309	28.7k
3	207	166	38.1k	108	70	23.3k
4	34	21	42.6k	38	16	44.0k
5	9	8	45.0k	13	9	26.2k

TABLE 5.8 – *Nombre moyen de candidats par requête de récupération de mots hors-vocabulaires mesurés sur les 100 premiers documents retournés par Google, à partir de la transcription automatique*

On remarque cependant que la dégradation provoquée par les erreurs de transcription est globalement moins importante pour le corpus AVISON que pour le corpus ESTER, alors que la transcription automatique du corpus AVISON contient plus d'erreurs que celle du corpus ESTER. Le rappel moyen obtenu par les stratégies *n*-gramme et patron est environ trois fois moins élevé que sur la transcription de référence pour le corpus ESTER alors qu'il n'est que 0.75 fois moins élevé sur le corpus AVISON.

Ce résultat peut s'expliquer par le fait que les structures linguistiques du corpus AVISON sont régulières et donc plutôt redondantes. Comme le modèle de langage capture facilement les structures redondantes, ces dernières ont tendance à soit être totalement erronées lorsque le processus de reconnaissance automatique de la parole s'est orienté vers la mauvaise structure, soit être totalement juste. Comme le corpus ESTER possède des structures linguistiques moins régulières, les erreurs sont mieux réparties et la probabilité qu'une erreur se glisse dans une requête est donc plus importante.

Un autre élément qui favorise ce phénomène est l'important déséquilibre entre le volume de données Web relatives au contenu du corpus ESTER et AVISON. En effet, si une erreur se glisse dans une requête, comme par exemple l'insertion d'un mot porteur de sens n'ayant pas de rapport avec la transcription, et que des documents correspondant fortement à cette requête existent, ils vont prendre la place des documents pertinents attendus. Si par contre aucun document ne correspond fortement à la requête erronée, mais que des documents pertinents répondent tout de même aux contraintes, ils seront retournés. Comme le Web recèle bien plus de documents relatifs au corpus ESTER que de documents relatifs au corpus AVISON, il est plus probable qu'une erreur dans une requête du corpus ESTER fasse dévier complètement le résultat de la recherche qu'une erreur dans une requête du corpus AVISON.

Les résultats obtenus sur la transcription automatique sont globalement moins bons que ceux obtenus sur la transcription de référence. On observe un rappel environ 3 fois moins important sur le corpus ESTER et environ 0.75 fois moins important sur le corpus AVISON. Cependant, le comportement des différentes stratégies sur la transcription automatique est semblable à celui observé sur la référence.

Performances des stratégies hybride

La robustesse des stratégies hybrides a été étudiée de la même manière que pour les stratégies de base. Le tableau 5.9 contient le rappel et le nombre moyen de mots par liste de candidats pour la méthode hybride reposant sur des n-grammes guidés par des mots pilotes et le tableau 5.10 contient les résultats obtenus par la méthode hybride reposant sur des patrons guidés par des mots pilotes.

	ESTER		AVISON	
n/m	**rappel**	**candidats**	**rappel**	**candidats**
2/1	8.7	292	29.1	1.2k
2/2	8.1	306	37.4	4.1k
2/3	6.5	295	38.1	6.2k
3/1	4.0	87	13.8	129
3/2	3.9	79	13.9	209
3/3	3.1	98	13.8	255

TABLE *5.9* – *Rappel et taille des liste de candidats pour la stratégie de requêtes n-gramme guidé par la sémantique, avec m guides sémantiques sur les 100 premiers documents retournés par Google, à partir de la transcription automatique*

	ESTER		AVISON	
n/m	**rappel**	**candidats**	**rappel**	**candidats**
2/1	8.2	266	20.1	317
2/2	7.9	325	22.0	397
2/3	5.7	354	23.2	440
3/1	3.7	158	8.4	63
3/2	3.6	133	6.8	52
3/3	2.9	97	5.3	48

TABLE *5.10* – *Rappel et taille des liste de candidats pour la stratégie de requêtes patron guidé par la sémantique, avec m guides sémantiques sur les 100 premiers documents retournés par Google, à partir de la transcription automatique*

On constate là aussi que les résultats obtenus sur la transcription automatique sont moins bons que ceux obtenus sur la transcription de référence. Les requêtes du corpus AVISON semblent plus robustes aux erreurs de reconnaissance automatique de la parole puisque le rappel moyen obtenu sur le corpus ESTER diminue d'environ 3 fois alors qu'il ne diminue que de 0.75 fois sur le corpus AVISON. On notera qu'on observe la même dégradation sur les deux corpus qu'avec les stratégies de base.

La dégradation des performances est uniforme sur les méthodes et on observe toujours que la stratégie hybride reposant sur des requêtes n-grammes est plus performante que la stratégie reposant sur des requêtes patrons. Il est ainsi possible de récupérer environ 30% des mots hors-vocabulaires du corpus AVISON avec une augmentation moyenne du lexique d'environ 1000 mots.

Ces expériences nous ont permis de mesurer le potentiel de chaque stratégie dans

les deux situations considérées, représentées par les corpus AVISON et ESTER. Nous avons montré que la stratégie offrant les meilleures performances sur ces deux corpus est la stratégie hybride reposant sur des requêtes syntaxiques n-grammes combinées à un descripteur sémantique local.

5.6.4 Performances de l'injection des mots

Afin d'évaluer les performances de la stratégie d'injection des mots dans le processus de reconnaissance automatique de la parole, les listes de candidats produites par la méthode hybride n-grammes et sémantique vont être insérées dans le processus de reconnaissance automatique de la parole en utilisant les deux stratégies présentées à la section 5.4.2. Les paramètres de la stratégie hybride sont $n = 2$ et $m = 1$, c'est à dire des requêtes bigrammes avec un mot-pilote. Nous avons vu que 29.1% des mots hors-vocabulaires se trouvent dans les listes de candidats du corpus AVISON et que 8.7% se trouvent dans celles du corpus ESTER.

Le tableau 5.11 contient les résultats des deux approches d'injection de mots présentées à la section 5.4.2 sur le corpus ESTER, en terme de précision et de rappel.

Le rappel est ici définit comme le rapport entre le nombre de mots hors-vocabulaires présents dans la transcription finale et le nombre de mots hors-vocabulaires présents dans les listes de candidats. Le rappel absolu est défini comme le rapport entre le nombre de mots hors-vocabulaires présents dans la transcription finale et le nombre de mots hors-vocabulaires présents dans la transcription de référence.

La précision est définie comme le rapport entre le nombre de mots hors-vocabulaires présents dans la transcription finale et le nombre total de nouveaux mots insérés dans cette dernière.

	Rappel	**Précision**	**Rappel absolu**	**WER**	**WER initial**
Mot inconnu	57.5	22.0	5.0	24.6	24.5
Morphosyntaxique	70.1	55.1	6.1	24.3	24.5

TABLE 5.11 – *Performances des stratégies d'injection des mots hors-vocabulaires dans le processus de reconnaissance automatique de la parole sur le corpus ESTER*

Ces résultats montrent sans équivoque que la meilleure approche pour injecter les mots candidats dans le processus de reconnaissance automatique de la parole est d'utiliser la technique reposant sur les classes morphosyntaxiques. Avec cette approche, la modélisation du comportement linguistique des mots insérés est plus fine, ce qui augmente la précision d'insertion. Avec l'approche reposant sur la classe des mots inconnus, tous les mots candidats sont interchangeables du point de vue linguistique et donc la correspondance acoustique est le seul facteur déterminant l'apparition des mots, ce qui favorise leur insertion inappropriée.

Le rappel bénéficie lui aussi de l'approche morphosyntaxique. En effet, dans certaines situations où le signal acoustique observé ne correspond pas complètement à la

phonétisation des mots hors-vocabulaires, la probabilité linguistique conditionnée par la classe morphosyntaxique apporte l'information nécessaire pour aboutir à la la bonne solution.

Le tableau 5.11 contient les résultats obtenus avec la méthode d'injection des mots par classe morphosyntaxique pour les corpus ESTER et AVISON, en terme de précision et de rappel.

	Rappel	**Précision**	**Rappel absolu**	**WER**	**WER initial**
ESTER	70.1	55.1	6.1	24.3	24.5
AVISON	98.6	49.9	28.7	47.7	48.7

TABLE 5.12 – *Performances globales de l'approche d'enrichissement dynamique du lexique*

Sur le corpus ESTER, environ 70% des mots hors-vocabulaires récupérés par les techniques de requêtes sont correctement insérés dans la transcription finale, ce qui représente environ 6% du nombre total de mots hors-vocabulaires de la transcription. La précision est d'environ 50%, ce qui veut dire que pour chaque mot hors-vocabulaire correctement inséré, une erreur est commise. Le bilan en terme d'erreurs de transcription devrait être nul, or on observe une diminution du taux d'erreur mot global de 0.2% absolu. Comme il y a environ 1% de mots hors-vocabulaires dans le corpus ESTER et que 6% sont récupérés, le gain de 0.2% indique que la correction de chaque mot hors-vocabulaire a permis de corriger en moyenne 3 erreurs de transcription. Ce résultat doit cependant être modéré par le fait qu'une diminution du taux d'erreur mot de 0.2% n'est pas très significative étant donné la taille du corpus de test d'ESTER.

Les performances obtenues sur le corpus AVISON sont bien meilleures que celle obtenues sur le corpus ESTER. Environ 98% des mots hors-vocabulaires présents dans les listes de candidats sont correctement transcrits dans la transcription finale, ce qui permet de récupérer environ 29% du nombre total de mots hors-vocabulaires de la transcription. La précision est là encore d'environ 50%, ce qui fait que le bilan global des erreurs de transcription doit être nul. Cependant, on observe là encore une diminution du taux d'erreur mot global de la transcription de 1% absolu. Étant donné qu'il y a environ 5% de mots hors-vocabulaires dans le corpus AVISON, on en déduit que chaque mot hors-vocabulaire corrigé a permis de réparer en moyenne 0.7 autres erreurs.

5.7 Conclusion du chapitre

Nous avons proposé une approche permettant d'augmenter localement et dynamiquement le lexique du moteur de reconnaissance automatique de la parole. Elle possède plusieurs caractéristiques intéressantes et inédites.

Les mots servant à l'augmentation sont dynamiquement récupérés sur le Web, ce qui permet de déléguer la fastidieuse tâche de collecte de documents pertinents pour l'adaptation au moteur de recherche. De plus, comme le Web est un corpus qui évolue

constamment, il est possible de traiter les documents relatifs à des sujets récents de manière totalement automatique.

Les mots hors-vocabulaires sont caractérisés par deux descripteurs locaux : un descripteur syntaxique et un descripteur sémantique. Cette manière de modéliser l'emploi des mots hors-vocabulaires permet d'effectuer une augmentation lexicale très précise. Nous avons montré expérimentalement qu'il est possible de récupérer 30% des mots hors-vocabulaires en effectuant une augmentation lexicale de seulement 1000 mots sur le corpus AVISON

Comme l'augmentation du lexique est locale, il est possible de l'effectuer uniquement lorsque c'est nécessaire, c'est à dire à chaque fois qu'un mot hors-vocabulaire apparaît dans les documents à transcrire. Cela évite de polluer inutilement l'ensemble du document avec un lexique inadapté.

Les expériences ont été effectuées dans deux contexte caractéristiques :

1. la transcription de données généralistes issues du corpus de journaux radiodiffusés ESTER
2. la transcription de données spécifiques à un domaine issues du corpus de vidéos d'opérations chirurgicales AVISON

L'approche permit dans les deux situations de récupérer de manière totalement automatique des mots hors-vocabulaires absents du lexique initial. Le taux d'erreur global des transcriptions finales a ainsi été réduit. Un des bénéfices de cette approche, outre le fait de réduire le taux d'erreur mots, est qu'un certain nombre de mots hors-vocabulaires porteurs de sens peuvent être récupérés. Si les transcriptions ont pour but d'alimenter l'index d'un moteur de recherche, ces mots sont très importants et permettent de trouver des documents qui normalement n'auraient pas pu l'être.

L'utilisation du Web comme un corpus ouvert dont l'accès se fait par l'intermédiaire des moteurs de recherche ouvre de nouvelles perspectives en terme d'adaptation automatique des modèles de langage et du lexique.

Quatrième partie

Adaptation des scores linguistiques à partir du Web

Chapitre 6

Etat de l'art : Adaptation des scores linguistiques à partir du Web

Sommaire

Ce chapitre présente un état de l'art de l'utilisation de données Web dans le cadre de la modélisation linguistique et plus particulièrement pour l'adaptation des probabilités n-grammes.

6.1 Introduction

La littérature se rapportant au domaine de l'adaptation des modèles de langage est prolifique. Nous allons présenter ici à une catégorie particulière de travaux : ceux dont le but est d'exploiter les données disponibles sur le Web. La motivation généralement rapportée est la quantité colossale de données textuelles qui se trouvent sur le Web.

Banko et Brill (2001) ont montré que dans certaines situations il était plus judicieux d'augmenter la taille du corpus d'entraînement plutôt que d'optimiser les paramètres des algorithmes d'apprentissage ou de mieux nettoyer les données. Sur une tâche de désambiguïsation de mots, les performances du système augmentent avec la quantité de données d'entraînement suivant une loi log-linéaire, même en utilisant des algorithmes d'apprentissage très simples.

Comme nous l'avons montré au chapitre 3, le Web constitue une source de données textuelle dont la taille est sans commune mesure avec les corpus utilisés traditionnellement dans le domaine du traitement automatique de la langue naturelle. De plus, les données qu'il contient sont en perpétuelle évolution, ce qui garantit la mise a jour automatique et continue. Le Web semble constituer le plus gros corpus textuel et permettrait donc d'augmenter considérablement la taille des corpus d'entraînement utilisés dans les tâches de traitement automatique de la langue naturelle.

On trouve deux grandes familles d'approches. La première consiste à utiliser une technique pour cibler un ensemble de documents pertinents pour la tâche de reconnaissance automatique de la parole considérée, à les nettoyer et à les utiliser comme un corpus classique. La seconde consiste à utiliser les moteurs de recherche Web pour obtenir des statistiques nécessaires à la construction de modèles de langage classiques. Nous allons présenter quelques travaux représentatifs de ces approches.

6.2 Collecte de documents

Les premières approches visant à utiliser le Web dans le cadre de la reconnaissance automatique de la parole ont surtout consisté à rechercher les documents pertinents pour la tâche considérée. Cette recherche peut être effectuée soit de manière non supervisée si l'on n'a aucune information sur les documents à transcrire, soit de manière supervisée si l'on connaît *a priori* le contenu de ces documents.

Beaucoup des travaux présentés ici proposent, en plus de l'adaptation du modèle de langage, une adaptation du lexique. Ils ont donc pour la plupart été décrits en détails a la section 4.4.1 du chapitre 4, relative à la collecte de documents d'adaptation.

6.2.1 Collecte dynamique

L'approche consistant à collecter des documents de manière non supervisée a de loin été la plus largement utilisée dans le cadre de la reconnaissance automatique de la parole. Généralement les auteurs utilisent un moteur de recherche Web pour sélectionner les documents pertinents. Toute la difficulté de ce type d'approche est de formuler des requêtes représentatives du contenu des documents à transcrire alors qu'on dispose de peu ou pas d'informations sur ceux-ci.

Lorsque sont associées aux documents à transcrire des informations les concernant, il est possible de les utiliser pour formuler les requêtes Web. Par exemple, Munteanu et al. (2007) et Rogina et Schaaf (2002) utilisent les mots-clés présents dans les diaporamas de présentation associés à des enregistrement de conférences et de cours pour trouver sur le Web des documents relatifs aux sujets des exposés.

Lorsqu'aucune information n'est disponible, l'extraction du contenu linguistique des documents à transcrire est généralement effectué par une première passe de transcription automatique. Des mots-clés sont alors extraits de cette transcription et servent

à construire des requêtes. C'est ce qu'on proposé par exemple Berger et Miller (1998) ou Vaufreydaz et al. (1999).

Un grand nombre de travaux se sont intéressés à l'optimisation des requêtes. Il a par exemple été proposé différentes techniques pour regrouper les mots-clés et former des requêtes moins bruitées. Les travaux de Ito et al. (2008) ou Suzuki et al. (2006) en sont un bon exemple. Il a également été proposé de modifier la nature des termes recherchés. Par exemple Sarikaya et al. (2005b) proposent d'utiliser des n-grammes comme requête afin de cibler des documents linguistiquement comparables aux données à transcrire. Meng et al. (2010), Tsiartas et al. (2010) ou Wan et Hain (2006) proposent d'aller plus loin en utilisant des phrases entières.

Cette technique a fait ses preuves lorsqu'aucune information n'est disponible *a priori* sur le contenu des documents à transcrire. Elle a également été utilisée avec succès par Bulyko et al. (2003) pour transcrire des documents dans une langue peu dotée.

6.2.2 Collecte *a priori*

Lorsque la tâche de reconnaissance automatique de la parole est bien définie et que l'on connaît à l'avance la nature et le contenu des documents à transcrire, il est possible d'utiliser le Web pour collecter de manière systématique des documents comparables.

La collecte a généralement lieu sur un ensemble de sites Web déterminés *a priori*. A intervalles réguliers, un robot vient télécharger les nouveaux document. Il s'agit d'une forme de réplication en temps réel du contenu des sites Web. Les documents ainsi extraits sont nettoyés et constituent un corpus qui peut être utilisé de différentes manières.

Par exemple Federico et Bertoldi (2004) proposent d'adapter quotidiennement un système de transcription de journaux télédiffusés. Leur approche consiste à récolter les articles publiés quotidiennement sur des sites d'information (60k mots/jour) et à les utiliser pour adapter le modèle de langage et le lexique initial du système de reconnaissance automatique de la parole. Leurs expériences mettent en évidence une réduction importante du taux d'erreur mot et du taux de mots hors-vocabulaires.

Allauzen et Gauvain (2003) proposent une approche semblable mais pondèrent les documents collectés par leur ancienneté par rapport au document à transcrire. Ils obtiennent ainsi un modèle de langage plus représentatif des actualités au moment où a été enregistré le document à transcrire.

Kemp et Waibel (1998) collectent également des articles d'actualité pour transcrire des enregistrements d'information. Ils proposent par contre une étape de filtrage qui consiste à ne conserver que les document les plus proches d'une première passe de transcription automatique.

6.2.3 Collecte hybride

Nous avons vu qu'il existe deux approches pour collecter des documents servant à construire un corpus d'entraînement ou d'adaptation en reconnaissance automatique de la parole. Il est possible de combiner ces deux approches pour cumuler leurs avantages.

C'est par exemple ce que proposent Kemp et Waibel (1998) pour une tâche de transcription de journaux radiodiffusés. D'une manière similaire à ce que proposent Federico et Bertoldi (2004), ils effectuent une collecte quotidienne de documents relatifs à l'actualité sur des sites Web d'information sélectionnés *a priori*. Ces documents sont ensuite nettoyés et indexés par un moteur de recherche local. Une première passe de transcription est effectuée sur les données à transcrire et les mots-clés qu'elle contient sont soumis au moteur de recherche pour trouver des documents similaires. Un modèle de langage est alors estimé à partir de ces données et est utilisé pour produire la transcription finale.

Martins et al. (2010) ont récemment proposé une technique d'adaptation hybride plus aboutie que la précédente. La tâche sur laquelle ils expérimentent leur approche est la transcription de journaux d'information télévisées.

Ils effectuent une collecte quotidienne de documents relatifs à l'actualité sur des sites Web d'information sélectionnés *a priori*. Ces documents sont ensuite nettoyés et servent à construire des corpus positionnés dans le temps. Les documents collectés sur internet correspondants aux 7 jours précédents le jour d forme le corpus $O_7(d)$.

Pour chaque document à transcrire datant du jour d, ils estiment un modèle de langage à partir du corpus d'entraînement généraliste et du corpus $O_7(d)$. Ils effectuent une première passe de transcription avec ce modèle et l'utilisent pour effectuer une segmentation thématique du document à transcrire.

Pour chaque thématique, des mots-clés sont extraits et servent à construire des requêtes qui sont soumises à un moteur de recherche local. Ils récupèrent ainsi un ensemble de documents relatifs à cette thématique dans l'ensemble des corpus locaux (les corpus collectés sur le Web et les corpus d'entraînement du modèle de langage initial).

Ces documents servent à estimer un second modèle de langage qui est interpolé avec le précédent et sert à produire la transcription finale du document.

Cette approche est très intéressante car elle tire parti des avantages des deux techniques. D'une part, la base de documents locale est enrichie avec des documents dont on sait *a priori* qu'ils seront pertinents pour la tâche considérée. On constitue ainsi progressivement un corpus ciblé faiblement bruité. D'autre part, l'aspect dynamique de la sélection de documents servant à estimer le modèle de langage final assure une forte adéquation entre le modèle de langage et les données à transcrire.

6.2.4 Traitement des documents

Les documents collectés sur le Web peuvent subir différents traitements et sélections avant de servir de corpus d'entraînement du modèle de langage.

Filtrage temporel

Allauzen et Gauvain (2003) et Martins et al. (2010) ont pour objectif de transcrire des documents relatifs à l'information. En connaissant la date de production des documents à transcrire, ils proposent de pondérer l'importance des différentes sources d'informations en fonction de leur date de production. Les modèles de langage ainsi estimés sont plus adaptés aux données du document à transcrire.

Filtrage thématique

Il arrivé également que les données collectées de manière systématiques sur le Web soient ensuite sélectionnées pour constituer des corpus thématiques adaptés aux documents à transcrire. Nous avons vu que c'est la solution que proposent Martins et al. (2010) lors de la construction du modèle de langage utilisé pour la seconde passe de transcription. Kemp et Waibel (1998) ont également proposé une approche similaire en indexant les documents collectés et en utilisant la première passe de transcription pour rechercher ceux qui se rapprochent le plus du contenu linguistique des documents à transcrire.

6.3 Exploitation des moteurs de recherche

Les approches qui consistent à télécharger des documents sur le Web et à utiliser des techniques classiques d'adaptation ou d'estimation des modèles de langage se heurtent à plusieurs problèmes. Le premier consiste à trouver un moyen de sélectionner un sous-ensemble de documents du Web pertinents pour la tâche considérée. Le second problème est le nettoyage de ces documents afin que les outils classiques de modélisation du langage puissent les traiter.

Il se trouvent que les moteurs de recherche Web mettent a disposition des statistiques sur l'ensemble des documents qu'ils indexent. On a par exemple vu au chapitre 3 qu'il était facile d'obtenir le nombre de document contenant une suite de mot quelconque. S'il était possible d'obtenir le nombre d'occurrences d'une suite de mots quelconque sur le Web, il serait alors possible de construire un modèle de langage n-gramme avec l'ensemble des documents du Web, uniquement en se basant sur les statistiques des moteurs de recherche.

Malheureusement, la plupart des moteurs de recherche ne fournissent que le nombre de documents qui correspondent à une requête. Si cette requête est un n-gramme,

alors on obtient le nombre de documents qui contiennent ce n-gramme.

On trouve dans la littérature plusieurs travaux exploitant ces statistiques à des fins de modélisation du langage. Par exemple, Zhu et Rosenfeld (2001) proposent une formule permettant d'estimer le nombre d'occurrence d'une séquence de mots à partir du nombre de documents dans lesquels elle apparaît sur le Web :

$$f^{Web}(w_{i-n+1}, w_{i-n}, ..., w_i) \approx \alpha \times df^{Web}(w_{i-n+1}, w_{i-n}, ..., w_i)^{\beta} \tag{6.1}$$

avec $f^{Web}(W)$ la fréquence de la séquence de mots W sur le Web et $df^{Web}(W)$ le nombre de documents contenant la séquence de mots W. α et β sont des constantes pour un ordre de n-gramme n donné.

Zhu et Rosenfeld (2001) ont estimé les valeurs de α et de β pour des n-grammes d'ordre 1 à 3. Leurs résultats en fonction de la valeur de n sont reportés dans le tableau 6.1.

n	α	β
1	2.427	1.019
2	1.209	1.014
3	1.174	1.025

TABLE 6.1 – *Constantes de proportionnalité entre fréquence de documents et fréquences d'occurrences de n-grammes*

Ils ont ainsi pu utiliser le Web pour estimer les probabilités des n-grammes non observés dans le corpus d'entraînement du modèle de langage. Ils ont montré que ces probabilités permettaient d'obtenir de meilleurs performances que celles obtenues avec une méthode de repli état de l'art.

Keller et Lapata (2003) ont également utilisé la fréquence de documents pour diverses applications. Ils ont d'abord montré que le nombre de documents Web contenant un bigramme extrait d'un corpus était proportionnel à la fréquence de ce bigramme dans le corpus. Ils ont alors estimé la probabilité de bigrammes non observés dans leur corpus à partir du nombre de documents les contenant sur le Web et ont observé une forte corrélation entre cette probabilité et l'estimation de celle-ci reposant sur un modèle à classes morphosyntaxiques. Ils ont finalement estimé un modèle bigramme à partir du web et ont montré qu'il fournissait de bonnes performances sur une tâche de désambiguïsation de mots. La principale force de ce modèle est la taille colossale du corpus avec lequel il est estimé (le Web).

Lapata et Keller (2005) ont poursuivi leurs travaux en testant leur modèle Web dans différents contextes de traitement automatique de la langue naturelle. Ils ont montré que ces modèles non supervisés fonctionnement mieux que les modèles simples estimés sur corpus. Par contre, ils fonctionnent moins bien que les modèles état de l'art supervisés faisant intervenir des sources de connaissances à forte valeur ajoutée comme les taxinomies.

Nakov et Hearst (2005b) a proposé une amélioration du modèle Web proposé par Lapata et Keller (2005) reposant sur la combinaison des mesures Web avec d'autres informations et ont obtenu des performances semblables à celle de l'état de l'art dans les tâches de désambiguïsation de mots.

On notera cependant que la mesure du nombre de documents contenant une requête est parfois irrégulière pour plusieurs raisons. Tout d'abord, le nombre de documents correspondant à une requête est en général destiné à informer le visiteur et il est souvent le fruit d'une estimation arrondie. De plus, étant donné que l'index des moteurs de recherche est en perpétuelle évolution, le nombre de documents retournés par une requête varie avec le temps. Finalement, pour des raisons techniques, les index des moteurs de recherche sont souvent stockés sur plusieurs serveurs et un mécanisme de répartition de charge redirige les requêtes clients sur l'un d'eux. Lorsque l'index est mis à jour, le temps que la modification se propage à tous les serveurs, un client peut effectuer deux fois la même requête, être redirigé sur deux serveurs différents et ainsi obtenir deux résultats différents. Ces éléments ont, en pratique, un effet assez limité sur les performances des modèles de langage estimés à partir de ces statistiques comme le montrent Nakov et Hearst (2005b) et Lapata et Keller (2005).

6.4 Conclusion

Beaucoup de travaux portant sur l'utilisation des données du Web pour l'estimation et l'adaptation des modèles de langage dans le cadre de la reconnaissance automatique de la parole consistent à sélectionner un sous-ensemble du Web et à l'exploiter comme un corpus classique. Cette voie a été très largement explorée, notamment en ce qui concerne l'aspect le plus délicat : optimiser la formulation des requêtes Web en fonction des informations dont on dispose sur les données à transcrire.

Une autre piste a commencé à être explorée et consiste à utiliser le Web comme un corpus ouvert. Les statistiques des moteurs de recherche Web permettent de dériver une mesure comparable à celles d'un modèle de langage conventionnel. Les quelques travaux mettant cette idée en pratique montrent de bons résultats et laissent penser qu'il s'agit d'une approche intéressante à développer.

Dans le chapitre suivant, nous développer cette dernière idée en proposant plusieurs modèles de langage reposant sur une utilisation dynamique du Web. Nous étudierons différentes stratégies permettant d'intégrer ces modèles au processus de reconnaissance automatique de la parole et aux modèles de langages classiques.

Chapitre 7

Modèles de langage probabilistes et possibilistes Web

Sommaire

Ce chapitre présente une nouvelle approche pour tirer parti de l'information disponible sur le Web. Nous proposons un modèle de langage qui utilise le Web comme un corpus ouvert et dont l'estimation est effectuée dynamiquement à l'aide des statistiques des moteurs de recherches. Ce modèle repose sur la théorie des possibilités. Plusieurs stratégies sont proposées pour l'intégrer dans le processus de reconnaissance automatique de la parole.

Ces propositions sont évaluées dans deux situations fréquentes en reconnaissance automatique de la parole : la transcription de données généralistes pour lesquelles on dispose d'un gros volume de données d'entraînement et la transcription de données relatives à un domaine de spécialité pour lesquelles on dispose de peu de données d'entraînement.

7.1 Introduction

Les modèles de langage sont utilisés dans de nombreux domaines, comme la recherche d'information, la traduction automatique ou la reconnaissance automatique de la parole. La qualité de ces modèles est déterminante pour les performance du système les utilisant.

Beaucoup de travaux ont porté sur l'amélioration des modèles de langage et quelques exemples sont présentés au chapitre 2. Il s'est avéré qu'il était très difficile d'améliorer significativement le modèle de langage n-gramme proposés par Jelinek (1976). Les principales avancées ont été faites dans des situations particulières, comme la transcription de langues peu dotées. Une piste d'amélioration intéressante est l'intégration de nouvelles données dans des modèles de langage classiques déjà estimés. Cette adaptation peut se faire de différentes manières et dépend de la nature des informations à intégrer dans le modèle.

Le Web peut être vu comme un corpus textuel de taille colossale et qui est en constante évolution. Ces deux caractéristiques, présentées plus en détail au chapitre 3, en font une source de données textuelle très intéressante du point de vue de la modélisation du langage. D'une part sa taille permet d'estimer des modèles fiables et d'autre part son caractère évolutif assure une mise a jour constante de son contenu.

Les travaux présentés au chapitre 6, qui ont porté sur l'utilisation de ressources Web pour améliorer la modélisation du langage, consistent globalement à récupérer des données textuelles sur le Web et à utiliser les techniques de modélisation ou d'adaptation classiques pour en tirer parti.

Nous l'avons vu, la principale force du Web est le volume de documents textuels qu'il contient. Ces documents sont très variés et leur nombre augmente sans cesse. L'estimation de probabilités à partir de ceux-ci présente quelques difficultés comme la sélection d'un sous-ensemble de documents homogènes correspondant à la tâche pour laquelle le modèle de langage est estimé, la récupération de ces documents ou encore

leur nettoyage pour les débarrasser de toutes les informations de mise en forme qu'ils contiennent.

Les caractéristiques du Web nous laissent penser qu'il serait plus intéressant de développer des approches spécifiques pour exploiter un tel corpus. Au lieu d'utiliser une sous-partie du Web comme un corpus fermé, par exemple en récupérant les documents qu'il contient pour les traiter comme un corpus classique, nous le considérons comme un corpus ouvert. Nous proposons alors d'utiliser les statistiques des moteurs de recherche comme moyen de manipuler les informations qu'il contient.

Nous avons vu que les travaux de Zhu et Rosenfeld (2001) et de Lapata et Keller (2005) ont exploité le Web d'une manière similaire. Les premiers n'ont pas poussé l'approche très loin puisqu'ils n'estiment que des probabilités de repli. Les seconds n'ont estimés que des modèles probabilistes bigrammes qui n'ont pas été évalués dans le contexte de la RAP.

Dans ces deux études, le Web a été utilisé principalement pour estimer des probabilités n-grammes à partir des documents qu'il contient. Cette mesure ne tient compte que des événements observés sur le Web. Cependant, de par le caractère colossal du Web, les événements non observés peuvent apporter une information. En effet, si une séquence de mot n'existe pas sur le Web, il est tout à fait légitime de remettre en cause la possibilité de son existence. C'est par exemple ce que montrent Keller et Lapata (2003) en observant une corrélation entre le nombre de pages Web contenant une séquence de mots et la plausibilité que lui accorde un humain.

Cette caractéristique du Web permet en outre de solutionner un problème majeur rencontré par Langlois et al. (2003) lorsqu'ils essayent d'intégrer dans le processus de reconnaissance automatique de la parole la notion de n-grammes impossibles : la génération *a priori* des n-grammes impossibles. Ici, il n'y a pas besoin de les connaître *a priori* puisque le Web peut être utilisé pour estimer la possibilité d'un n-gramme en temps réel.

De même que la Théorie des Probabilités est utilisée pour tirer parti des événements observés sur le Web, nous proposons d'utiliser la Théorie des Possibilités (de Cooman, 1997; Dubois, 2006) pour tirer parti des événements non observés et permettre la quantification du caractère possible des n-grammes.

Dans ce chapitre, nous allons proposer une mesure permettant d'estimer la possibilité d'une séquence de mots à partir du Web. Cette mesure sera étendue à l'estimation d'une possibilité à partir de n'importe quel corpus textuel. Une amélioration et une généralisation de la mesure de probabilité Web proposée par Zhu et Rosenfeld (2001) sera également proposée.

Nous proposons ainsi de mesures probabilistes et possibilistes pouvant être appliquées à des corpus conventionnels ainsi qu'au Web. Nous allons alors proposer différentes stratégies pour combiner ces mesures et ainsi produire une mesure linguistique intégrant toutes ces informations.

7.2 Probabilités

Les modèles de langage probabilistes sont très largement répandus dans le domaine de la modélisation du langage. Ils sont traditionnellement estimés à partir de documents textuels nettoyés appartenant à un corpus.

7.2.1 Probabilités estimées sur corpus

L'essentiel des modèles de langage probabilistes estimés sur corpus sont des modèles de langage n-grammes. Quelques variantes ont été proposées dans la littérature, comme les modèles *cache* et *trigger*, mais nous ne considérerons dans ce chapitre uniquement les modèles n-grammes classique. Des informations plus détaillées sur ce type de modèles se trouvent au chapitre 2.

7.2.2 Probabilités estimées sur le Web

La plupart des approches présentées au chapitre 6 permettant d'estimer un modèle de langage à partir des données du Web consistent à collecter un ensemble de documents pertinent sur le Web, par exemple à l'aide d'un moteur de recherche, à les nettoyer et à estimer un modèle de langage n-gramme en utilisant les techniques classiques. Un sous-ensemble du Web est alors utilisé comme un corpus fermé classique.

Cette approche pose plusieurs problèmes. Tout d'abord elle est très limitée quant au nombre de documents utilisés pour la modélisation du langage. En effet, il faut télécharger tous les documents, ce qui prend du temps et est coûteux en terme d'espace de stockage. De plus, sélectionner *a priori* un ensemble de documents nécessite une expertise permettant de déterminer lesquels seront pertinents. Finalement, le travail de nettoyage est complexe car les documents Web sont en général très pollués par les informations de mise en forme.

Considérer le Web comme un corpus ouvert dans le cadre de l'estimation d'un modèle de langage nécessite de pouvoir mesurer facilement la fréquence de séquences de mots sur l'ensemble du Web. Ces statistiques sont dynamiques puisque le Web est considéré comme ouvert. Il faut donc une mesure qui s'adapte dynamiquement aux variations du contenu du Web.

Les moteurs de recherche Web existent depuis de nombreuses années et certains arrivent à indexer une très grande partie des documents. De plus, pour réaliser leur index, ces moteurs ont dû nettoyer les documents. Si nous avions accès directement à l'index dynamique d'un moteur de recherche, notre tâche serait facilitée. Il suffirait d'adapter les outils d'exploitation de l'index à nos besoins pour obtenir des comptes de séquences de mots dynamiquement. Malheureusement l'accès direct à l'index est en général impossible et on doit utiliser une couche logicielle d'exploitation, fournie par le moteur de recherche. Cette couche nous permet, en général, de n'effectuer que les

opérations qui sont possible à un utilisateur du portail Web du moteur de recherche, c'est à dire rechercher des documents.

Cependant, sur certains moteurs, le nombre de documents qui contiennent les termes recherchés est accessible. En fait il s'agit d'une estimation car cette statistique est trop lourde à calculer en temps réel, même pour les plus gros moteurs de recherche. Si nous fournissons une requête sous la forme d'une séquence de mots, le moteur de recherche nous retourne une estimation du nombre de documents dans son index qui contiennent la séquence de mots. Pour obtenir la fréquence de la séquence de mots, il faudrait normalement récupérer tous les documents et effectuer le compte de la séquence dans chacun d'eux. Cette tâche est elle très lourde et ne correspond pas à l'utilisation du Web comme corpus ouvert.

Comme nous l'avons vu au chapitre 6, Zhu et Rosenfeld (2001) ont étudiés la relation qui existe entre le nombre de documents Web retournés par un moteur de recherche contenant une séquence de mots et le nombre d'apparitions de la séquence dans ces documents. Les auteurs montrent qu'il est possible d'estimer la fréquence d'un n-gramme à partir du nombre de documents qui le contiennent à partir de la formule 7.1 :

$$f^{Web}(w_{i-n+1}, w_{i-n}, ..., w_i) \approx \alpha \times df^{Web}(w_{i-n+1}, w_{i-n}, ..., w_i)^{\beta} \tag{7.1}$$

avec $f^{Web}(W)$ la fréquence de la séquence de mots W sur le Web et $df^{Web}(W)$ le nombre de documents contenant la séquence de mots W. α et β sont des constantes pour un ordre de n-gramme n donné.

Zhu et Rosenfeld (2001) ont estimé les valeurs de α et de β pour des n-grammes d'ordre 1 à 3. Leurs résultats en fonction de la valeur de n sont reportés dans le tableau 7.1. On constate que la valeur de β qu'ils ont trouvé est très proche de 1, ce qui indique que la relation qui lie ces deux mesures est une relation de proportionnalité. Comme on pourrait s'y attendre, plus l'ordre du n-gramme est élevé, plus la valeur de α est basse. En effet, plus une séquence de mots est grande, moins il est probable qu'on la trouve plusieurs fois dans un même document. On constate également que la valeur de α est très rapidement proche de 1, comme par exemple pour $n = 2$ on obtient une valeur d'environ 1.2.

n	α	β
1	2.427	1.019
2	1.209	1.014
3	1.174	1.025

TABLE 7.1 – *Constantes de proportionnalité entre fréquence de documents et fréquences d'occurrences de n-grammes*

Étant donné ces considérations, il nous faut estimer au moins la valeur de proportionnalité α si nous souhaitons obtenir des fréquences de n-grammes du Web. Cependant, le but des mesures de fréquences n-gramme Web est d'estimer des probabilités

n-grammes. Le calcul d'une telle probabilité est donné par la formule 7.2 :

$$P(w_i|h_i^n) \approx \frac{f(h_i^n, w_i)}{f(h_i^n)} \tag{7.2}$$

avec $f(W)$ la fréquence de la séquence de mots W dans le corpus qui sert à l'estimation et h_i^n l'historique de taille n du mot w_i.

En utilisant la formule de Zhu et Rosenfeld (2001) pour obtenir les fréquences de séquences de mots à partir du Web, on obtient la formule 7.3 :

$$P_{Web}(w_i|h_i^n) \approx \frac{\alpha^n \times df^{Web}(h_i^n, w_i)}{\alpha^{n-1} \times df^{Web}(h_i^n)} \tag{7.3}$$

avec $df^{Web}(W)$ le nombre de documents contenant la séquence de mots W et α^n le coefficient de proportionnalité de Zhu et Rosenfeld (2001) pour l'ordre n.

On constate que l'information nécessaire à l'estimation d'une probabilité d'ordre n est le rapport des coefficients de proportionnalités des ordres n et $n-1$. Le tableau 7.2 contient les rapports de coefficients de proportionnalité des ordres 2 et 3, les seuls que l'on puisse calculer à l'aide des valeurs du tableau 7.1. Étant donné que la différence entre les coefficients d'ordre 1 et 2 est importante, le rapport entre ces coefficients, nécessaire pour estimer des probabilités 2-grammes, est d'environ 0.5. Par contre, comme nous l'avons vu, les coefficients de proportionnalité tendent rapidement vers 1 avec l'augmentation de l'ordre n, ce qui fait que dès le rapport suivant, entre les ordres 2 et 3, on obtient une valeur très proche de 1. On peut en déduite que la valeur du rapport des coefficients de proportionnalité d'ordres n et $n-1$ peut être considéré comme valant 1 à partir de l'ordre 2.

n	α^n/α^{n-1}
2	0.498
3	0.971

TABLE 7.2 – *Rapport de coefficients de proportionnalité entre fréquence de documents et fréquences d'occurrences de n-grammes*

On peut ainsi estimer une probabilité n-gramme Web directement à partir des fréquences de document Web avec une bonne précision pour les ordres strictement supérieurs à 2 par la formule 7.4 :

$$
\begin{aligned}
P_{Web}(w_i|h_i^n) &\approx \frac{f^{Web}(h_i^n, w_i)}{f^{Web}(h_i^n)} \\
&\approx \frac{\alpha^n \times df^{Web}(h_i^n, w_i)}{\alpha^{n-1} \times df^{Web}(h_i^n)} \\
&\approx \frac{df^{Web}(h_i^n, w_i)}{df^{Web}(h_i^n)}
\end{aligned}
\tag{7.4}
$$

avec $f^{Web}(W)$ la fréquence de la séquence de mots W sur le Web et $df^{Web}(W)$ le nombre de documents contenant la séquence de mots W. α^n est la constante de proportionnalité pour l'ordre n.

Comme c'est le cas pour l'estimation classique des modèles de langage n-grammes, la formule 7.4 affecte une probabilité nulle aux séquences de mots qui n'ont pas été observées. Ce problème est généralement résolu en redistribuant une partie des probabilités des événements observés aux événements non observés.

Étant donné que les statistiques nécessaires à l'estimation d'un repli état de l'art, comme la technique de Kneser-Ney modifiée par Goodman (2006), ne sont pas disponibles en exploitant les données Web de cette manière, une autre technique de lissage sera utilisée : l'interpolation avec la distribution d'ordre inférieure. Bien que cette technique de lissage soit simple, Goodman (2006) montre qu'elle fournit de bonnes performances.

La probabilité n-gramme est alors calculée de la manière suivante :

$$
\begin{aligned}
P^*_{\text{web}}(w_i|h_i^n) = \alpha_n \cdot P_{\text{web}}(w_i|h_i^n) + \\
\alpha_{n-1} \cdot P_{\text{web}}(w_i|h_i^{n-1}) + \cdots + \alpha_1 \cdot P_{\text{web}}(w_i)
\end{aligned}
\tag{7.5}
$$

avec le vecteur $(\alpha_i)_{i=1}^n$ l'ensemble des coefficients d'interpolation tel que $\sum_{i=1}^{n} \alpha_i = 1$.

On notera que cette formulation présente une difficulté : l'estimation des fréquences unigrammes à partir du Web. Cette probabilité est définie comme le rapport entre le nombre de documents contenant un terme donné et le nombre total de documents se trouvant sur le Web. L'estimation du nombre total de documents que contient le Web pour une langue donnée est effectué en mesurant la fréquence de documents d'un des mots les plus fréquents de la langue considérée qui présente le moins de conflits avec d'autres langues, par exemple le mot "the" pour l'anglais.

A partir de la manière classique d'estimer des probabilités à partir de corpus textuels fermés, nous avons proposé une manière d'estimer des probabilités n-grammes à partir de l'ensemble du Web. Il est alors utilisé comme un corpus ouvert dont les moteur de recherche permettent la manipulation. La mesure est calculée à la volée à partir

des statistiques produites par les moteurs de recherche Web. Cette approche est intéressante puisqu'elle ne dépend d'aucune statistique *a priori* et permet donc à la mesure d'évoluer en même temps que le corpus.

7.3 Possibilités

Nous proposons ici une mesure de possibilité pouvant être estimée aussi bien sur des corpus textuels classiques que sur l'ensemble des données que contient le Web.

7.3.1 Introduction

La théorie des possibilités a été proposée par Zadeh (1978). Il s'agit d'un cadre mathématique permettant de manipuler l'incertain résultant d'une connaissance incomplète (Dubois, 2006). En ce sens, elle est un complément à la théorie des probabilités. Bien que conçue à l'origine pour formaliser la notion d'incertitude linguistique, un cadre plus formel lui a récemment été donné par de Cooman (1997), en s'appuyant sur des concepts de la théorie de la mesure. Ces travaux offrent ainsi un cadre mathématique pour manipuler des connaissances incomplètes.

Cette théorie est née du constat que la théorie des probabilités, très largement utilisée, est insuffisante dans certaines situations car elle utilise une seule mesure pour représenter deux concepts qui sont fondamentalement différents : l'incertitude et l'imprécision. Par exemple, si l'affirmation "cet homme a 42 ans" a une probabilité de 80%, cela peut signifier deux choses. La première relève de l'imprécision : l'homme a environ 42 ans, avec une marge d'erreur de 10%. La seconde relève de l'incertitude : il y a 80% de chances que l'homme ait exactement 42 ans.

Afin de séparer ces informations, la théorie des possibilités propose deux mesures : la *nécessité* et la *possibilité*. La première peut être vue comme une mesure de l'imprécision et la seconde comme une mesure de l'incertitude. Plus simplement, la possibilité indique à quel point les informations dont nous disposons nous permettent d'affirmer qu'un événement donné est possible.

Une première notion fondamentale est la notion de *distribution possibiliste*. Il s'agit d'une fonction π qui associe à chaque élément e d'un ensemble d'événements E, une valeur de l'intervalle unité $[0;1]$. Cette fonction représente l'information distinguant ce qui est possible de ce qui l'est moins. Par convention, pour un événement e, on a :

- si $\pi(e) = 0$, l'événement e est impossible
- si $\pi(e) = 1$ l'événement e est totalement possible.

Pour un univers fini et dénombrable Ω dont tous les sous-ensembles sont mesurables, on a les axiomes suivants :

$$\begin{aligned} &\pi(\varnothing) = 0 \\ &\pi(\Omega) = 1 \\ &\pi(A \cup B) = \max(\pi(A), \pi(B)) \quad \forall A, B \subset \Omega \end{aligned} \tag{7.6}$$

De plus, on a généralement :

$$\pi(A \cap B) \leq \min(\pi(A), \pi(B)) \quad \forall A, B \subset \Omega \tag{7.7}$$

Comme pour la théorie des probabilités, une mesure possibiliste peut être construite à partir d'une distribution possibiliste, si l'ensemble des événements est fini (de Cooman, 1997). Une mesure de possibilité Π peut être définie sur un ensemble d'événements E tel que :

$$\Pi(E) = \max_{e \in E} \pi(e) \tag{7.8}$$

avec π la distribution de possibilités de E.

$\Pi(E)$ évalue alors dans quelle mesure l'ensemble E est cohérent avec la connaissance π.

Nous allons proposer une méthode d'estimation de la mesure de possibilité pour des séquences de mots à partir des données provenant d'un corpus fermé ou à partir de l'ensemble des données du Web.

7.3.2 Possibilités estimées sur le Web

Dans cette section nous allons montrer comment obtenir une mesure possibiliste pour des séquences de mots, en utilisant des statistiques issues du Web.

La mesure possibiliste doit représenter la possibilité qu'une séquence de mots existe. Pour cela, nous nous appuyons sur l'existence ou non de cette séquence et des sous-séquences la composant sur le Web. Nous entendons ici par *existence* sur le Web le fait qu'il existe au moins un document web contenant la séquence de mots en question. Ainsi, plus il y a de sous-séquences de la suite de mots qui existent sur le Web, plus la suite de mots est possible. Cependant, il est nécessaire de borner la recherche de sous-séquences pour obtenir une mesure fiable. En effet, plus le corpus considéré pour calculer la mesure possibiliste est petit, moins la non-existence de séquences longues sera significative. C'est l'ordre du modèle qui sert de borne.

Tout d'abord, pour l'ordre désiré n du modèle de langage, nous construisons récursivement un ensemble distinct de distributions de possibilités π_n à π_1, selon l'équation 7.9 :

$$\pi_n(W) = \frac{|W_n \cap \text{Web}_n| + \alpha \cdot |W_n \backslash \text{Web}_n| \cdot \pi_{n-1}(W)}{|W_n|} \tag{7.9}$$

avec W une séquence de n mots ou plus, W_n l'ensemble des séquences de mots de taille n composant W, Web_n est l'ensemble des séquences de mots de taille n sur le Web, $\backslash$ est l'opérateur de différence ensembliste, et $0 \leq \alpha \leq 1$ est un coefficient de repli. La condition terminale de la récursion est $\pi_0(W) = 0$.

Pour une séquence de mots W, la valeur résultante de cette formule est le nombre de sous-séquences de taille n de W présentes sur le web, normalisée par le nombre total de sous-séquences de taille n de W. Pour augmenter la résolution de la mesure, un repli est effectué en interpolant la distribution possibiliste avec celle d'ordre inférieur pour les événements non observés. Le coefficient α permet de calibrer l'importance de ce repli dans le score final.

Les distributions possibilistes que l'on vient de définir nous permettent de construire un ensemble de mesures possibilistes correspondantes Π_n, à l'aide de la formule 7.10 :

$$\Pi_n(\Theta) = \max_{W \in \Theta}(\pi_n(W)) \tag{7.10}$$

avec Θ un ensemble de séquences de n mots ou plus. Si Θ a un seul élément W, alors $\Pi_n(\{W\}) = \pi_n(W)$.

On peut noter que contrairement à la modélisation n-gramme classique, cette mesure de possibilité évalue une séquence de mots dans sa globalité et ne nécessite pas de décomposer la possibilité globale en possibilités conditionnelles.

Dans cette modélisation, l'existence des séquences de mots sur le Web est vérifiée en mesurant une fréquence de document positive auprès d'un moteur de recherche Web. De manière analogue au modèle probabiliste Web que nous avons proposé précédemment, nous utilisons ici le Web comme un corpus ouvert.

7.3.3 Possibilités estimées sur corpus

Le modèle possibiliste Web présenté précédemment utilise le Web comme un corpus ouvert. Il est possible de généraliser la formule d'estimation des distributions de possibilités pour permettre leur estimation sur n'importe quel corpus, qu'il soit ouvert ou fermé.

La mesure de possibilité Web que nous avons proposé repose sur l'idée que, pour une séquence de mots W, plus il existe de longues sous-séquences de W sur le Web, plus W peut être considéré comme possible. Il est facile d'étendre cette affirmation à n'importe quel corpus C.

Comme précédemment, pour chaque ordre n du modèle de langage souhaité, on construit récursivement un ensemble de distributions de possibilités π_i^c avec $i \in \{1, \ldots, n\}$:

$$\pi_n^c(W) = \frac{|W_n \cap C_n| + \alpha \cdot |W_n \backslash C_n| \cdot \pi_{n-1}(W)}{|W_n|} \tag{7.11}$$

avec W une séquence de n mots ou plus, W_n l'ensemble des séquences de mots de taille n composant W, C_n est l'ensemble des séquences de mots de taille n dans le corpus, $\backslash$ est l'opérateur de différence ensembliste, et $0 \leq \alpha \leq 1$ est un coefficient de repli. La condition terminale de la récursion est $\pi_0^c(W) = 0$.

Pour une séquence de mots W, la valeur résultante de cette formule est le nombre de sous-séquences de taille n de W présentes dans le corpus C, normalisée par le nombre total de sous-séquences de taille n de W. Pour augmenter la résolution de la mesure, un repli est effectué en interpolant la distribution possibiliste avec celle d'ordre inférieur pour les événements non observés. Le coefficient α permet de calibrer l'importance de ce repli dans le score final.

Il est alors possible de définir une mesure de possibilité Π_n^c à partir de la distribution de possibilité π_n^c définie précédemment :

$$\Pi_n^c(A) = \max_{W \in A}(\pi_n^c(W)) \tag{7.12}$$

avec A un ensemble de séquences de n mots ou plus. Si A a un seul élément W, alors $\Pi_n^c(\{W\}) = \pi_n^c(W)$.

Nous avons proposé une formulation permettant d'estimer une mesure de possibilité à partir du Web ou de corpus conventionnels. Nous allons maintenant étudier la manière dont ces mesures peuvent être intégrées au sein du processus de reconnaissance automatique de la parole.

7.4 Intégration dans le processus de reconnaissance automatique de la parole

Afin d'évaluer les différentes hypothèses, le système de reconnaissance automatique de la parole utilise normalement la formule 1.4. Deux mesures sont utilisées : la probabilité acoustique et la probabilité linguistique. Cependant, pour des raisons pratiques, des libertés sont prises sur cette formulation théorique. Tout d'abord, la probabilité acoustique est généralement replacée par la vraisemblance acoustique, afin d'accélérer son calcul. De plus, on ajoute souvent une pénalité d'insertion de mots dans l'hypothèse pour contrôler plus finement le découpage en mots du flux de phonèmes. Étant donné ces considérations, une hypothèse n'est plus associée à une probabilité, mais à un score. Nous parlerons donc ici de scores acoustiques et linguistiques. La formule utilisée pour calculer le score d'une hypothèse H est alors :

$$S(H) = S_a(X|H) \times S_l(H)^{fl} \times fm^{|H|} \tag{7.13}$$

avec $S_a(X|H)$ la vraisemblance acoustique de H par rapport à l'observation X, $S_l(H)$ le score linguistique de H, fl le facteur linguistique permettant d'équilibrer les deux scores et fm le facteur qui pénalise l'augmentation du nombre de mots dans l'hypothèse.

Il est alors possible d'intégrer les modèles présentés précédemment de différentes manières.

7.4.1 Score linguistique à part entière

La méthode la plus simple consiste à utiliser les scores produits par les différents modèles directement comme score linguistique dans l'équation 7.13. On combine ainsi les scores fournis par le modèle en question avec les scores acoustiques pour obtenir le score de l'hypothèse.

Possibilités

La formule 7.11 définit la mesure possibiliste d'une séquence de mots quelconque. Il est alors possible d'obtenir la possibilité d'une séquence de mots de taille variable telle que les hypothèses produites par un système de reconnaissance automatique de la parole. Le score linguistique d'une hypothèse H est donc :

$$S_l(H) = \Pi(H) \tag{7.14}$$

Étant donné que la nature du score linguistique change, il est nécessaire d'adapter le facteur linguistique fl de la formule 7.13. On notera qu'étant donné que la possibilité n'est qu'une des deux informations modélisées par la probabilité, cette formulation conduit à une perte d'information.

7.4.2 Probabilités

L'intégration des modèles probabilistes n-grammes, qu'ils soient issus du Web ou de corpus, est triviale, dans la mesure où la plupart des systèmes de reconnaissance automatique de la parole utilisent cette famille de modèles. Il suffit donc de remplacer le score linguistique de l'équation 7.13 par la probabilité fournie par le modèle. Par exemple pour un modèle n-gramme Web :

$$S_l(H) = P_{web}(H) \tag{7.15}$$

7.5 Combinaison de probabilités et de possibilités

Nous disposons de quatre modèles de langage :

- Un modèle probabiliste n-gramme avec repli estimé sur corpus
- Un modèle probabiliste n-gramme avec lissage estimé sur le Web
- Un modèle possibiliste estimé sur corpus
- Un modèle possibiliste estimé sur le Web

Il existe plusieurs manières de combiner ces modèles pour obtenir un score linguistique unique qui pourra être utilisé dans l'équation 7.13.

7.5.1 Possibilités comme borne supérieur des probabilités

Plusieurs définitions de la relation entre possibilités et probabilités ont été proposées. Dans la définition de Dubois et Prade (1988), s'il existe une mesure de probabilité P et une mesure de possibilité Π définies sur un univers Ω, alors les deux mesures doivent respecter l'équation suivante :

$$\forall A \subseteq \Omega, P(A) \leq \Pi(A) \tag{7.16}$$

On peut utiliser cette propriété pour améliorer un modèle de langage probabiliste à partir d'informations de nature possibilistes. En effet, il arrive que le modèle de langage probabiliste assigne une probabilité plus importante qu'il ne devrait à certains événements. C'est systématiquement le cas des modèles de langage n-gramme probabilistes qui assignent une probabilité non nulle à tous les événements, même ceux qui sont impossibles.

Si une mesure de possibilité fiable est disponible, il est alors possible de l'utiliser comme borne supérieure de la probabilité. La masse de probabilité perdue peut ensuite être redistribuée aux autres événements. L'équation 7.17 formalise cette idée pour un modèle de langage initial P :

$$\hat{P}(w_i|h_i^n) = \begin{cases} \Pi_n(h_i^n, w_i)^\alpha, & \text{si } \Pi_n(h_i^n, w_i)^\alpha < P(w_i|h_i^n) \\ \beta \cdot P(w_i|h_i^n), & \text{sinon} \end{cases} \tag{7.17}$$

avec Π_n la mesure de possibilité, α un facteur d'échelle qui permet de contrôler l'impact de la distribution de possibilités sur la distribution de probabilité (si $\alpha = 0$, aucune probabilité n'est modifiée). β est le facteur de normalisation qui permet de redistribuer la masse de probabilité perdue et est définit ainsi :

$$\beta = \frac{1 - \sum_{u \in U_{h_i^n}} \hat{P}(u|h_i^n)}{1 - \sum_{u \in U_{h_i^n}} P_{\text{LM}}(u|h_i^n)} \tag{7.18}$$

avec $U_{h_i^n}$ l'ensemble des mots w_i d'historique h_i^n de taille $n-1$, pour lesquels la probabilité initiale était plus élevée que la possibilité.

Il est possible de combiner ainsi les deux modèles de langage probabilistes avec les deux modèles de langage possibilistes. Cependant, la combinaison la plus intéressante semble être le modèle de langage probabiliste estimé sur corpus modifié par le modèle possibiliste estimé sur le Web. En effet, le modèle de langage probabiliste estimé sur corpus devrait estimer correctement une grande partie des événements qui peuvent être rencontrés dans les documents à transcrire, si le corpus est suffisamment représentatif. Par contre, pour les événements peu fréquents et ceux qui n'ont pas été observés, son estimation est peu fiable et dans ces situations la mesure de possibilité permettra d'interdire les événements impossible.

7.5.2 Probabilités Web comme modèle de repli

Le mécanisme de repli dans un modèle de langage n-gramme sert à corriger les lacunes du corpus à partir duquel il est estimé. Les événements qui ne sont pas observés dans ce corpus sont considérés comme néanmoins possibles et une probabilité non nulle leur est attribuée grâce à une technique de repli. Cette probabilité est distribuée sur les événements non observés à l'aide d'un modèle n-gramme d'ordre inférieur.

Les probabilités de repli sont généralement de mauvaise qualité car elles sont estimées à l'aide d'heuristiques qui s'appuient sur des statistiques trop généralistes, comme les fréquences de fréquences de mots. Le tableau 7.3 contient les taux d'erreur mots obtenus sur les corpus ESTER et AVISON, présentés au chapitre 5.5.1, en fonction de la présence des n-grammes d'ordres 3 à 6 dans les modèles de langage ayant servi à la transcription. Les résultats sont présentés pour la meilleure hypothèse (colonne *1-best*) et pour les 100 meilleures hypothèses (colonne *100-best*). On voit très clairement que la majorité des erreurs sont rencontrées lorsque le modèle de langage effectue un repli linguistique.

	ESTER		**AVISON**	
	1-best	**100-best**	**1-best**	**100-best**
$\exists$ **3-gramme**	18.9	26.9	27.9	31.4
$\nexists$ **3-gramme**	42.6	53.4	57.6	59.5
$\exists$ **4-gramme**	17.7	22.1	21.8	24.4
$\nexists$ **4-gramme**	31.8	41.5	50.9	52.5
$\exists$ **5-gramme**	11.1	17.2	13.7	15.4
$\nexists$ **5-gramme**	26.5	35.8	46.0	47.9
$\exists$ **6-gramme**	8.0	13.4	7.0	7.9
$\nexists$ **6-gramme**	23.4	32.6	42.9	44.7

TABLE 7.3 – *Répartition des erreurs en fonction de la présence des n-grammes dans le modèle de langage sur les corpus AVISON et ESTER.*

Il est raisonnable de considérer que le modèle de langage probabiliste Web est de meilleure qualité que les probabilités résultantes d'une technique de repli linguistique. Il est alors possible de remplacer les probabilités obtenues avec la technique de repli du modèle de langage initial par les probabilités fournies par le modèle de langage Web.

L'intérêt de cette approche est que la partie du modèle de langage initial qui est considérée comme fiable est inchangée.

Une approche semblable a été proposée initialement par Zhu et Rosenfeld (2001), bien que la nature du modèle de langage Web qu'ils utilisent soit différente et qu'ils ne modifient pas toutes les situations de repli.

En considérant $U_{h_i^n}$ comme l'ensemble des mots w_i d'historique h_i^n de taille $n-1$ pour lesquels le modèle de langage P se replie, cette approche peut être formalisée ainsi :

$$\hat{P}(w_i|h_i^n) = \begin{cases} \alpha \cdot P(w_i|h_i^n) + (1-\alpha) \cdot P^*_{\text{web}}(w_i|h_i^n), & \text{si } w_i \in U_{h_i^n} \\ \beta \cdot P(w_i|h_i^n), & \text{sinon} \end{cases} \tag{7.19}$$

où α est un facteur d'échelle déterminé empiriquement et β est le facteur de normalisation qui permet de redistribuer la masse de probabilité perdue et est définit à l'équation 7.18.

On notera que l'utilisation de cette approche implique d'accorder une plus grande confiance au modèle de langage estimé sur corpus qu'au modèle de langage Web. Cette formulation est donc compatible avec la conclusion des travaux de Lapata et Keller (2005) qui montrent que dans certaines situations, un modèle de langage estimé sur un corpus bien ciblé est meilleur qu'un modèle de langage estimé sur le Web dans sa globalité.

7.5.3 Possibilités comme facteur de repli linguistique

La mesure de possibilité nous renseigne sur la confiance que l'on peut avoir sur la possible existence d'une séquence de mots. Si l'on accorde au corpus d'entraînement d'un modèle de langage une confiance plus élevée qu'au corpus servant à estimer la mesure de possibilité, alors tous les n-grammes vus dans ce corpus sont totalement possibles ($\pi_n(h_i^n, w_i) = 1$). Par contre, les n-grammes composés grâce aux stratégies de repli sont sujets à controverse. Nous proposons donc de pondérer la probabilité qu'accorde le modèle de langage aux n-grammes non vus dans le corpus d'entraînement par la possibilité de ces n-grammes. Cette approche est formalisée dans l'équation 7.20 :

$$\hat{P}(w_i|h_i^n) = \begin{cases} \Pi_n(\{h_i^n, w_i\}) \cdot \alpha(h_i^n) \cdot P(w_i|h_i^{n-1}), & \text{si } w_i \in U_{h_i^n} \\ \beta \cdot P_{\text{LM}}(w_i|h_i^n), & \text{sinon} \end{cases} \tag{7.20}$$

avec $\alpha(h_i^n)$ le coefficient de repli du modèle de langage original.

La masse de probabilité attribuée à tort à des événements impossibles du point de vue de la connaissance Π sera redistribuée aux événements vus dans le corpus d'apprentissage, par l'intermédiaire du coefficient β définit dans l'équation 7.18.

7.5.4 Combinaison log-linéaire

Au lieu de délimiter les domaines de compétence des modèles de langage probabilistes et possibiliste proposés comme nous l'avons fait avec les stratégies de combinaison précédente, on peut considérer ces modèles comme complémentaires.

Comme nous l'avons vu, lors de l'implémentation d'un système de reconnaissance automatique de la parole, des libertés sont prises sur la formulation théorique. Chaque hypothèse H est évaluée à l'aide d'un score qui résulte d'une combinaison linéaire entre son score acoustique ($S_a(X|H)$), son score linguistique ($S_l(H)$) et une pénalité d'insertion de mots. Nous proposons d'aller plus loin en estimant le score linguistique d'une hypothèse par une combinaison linéaire des différents modèles, quelque soit leur nature.

Par exemple, il est possible de cumuler les informations fournies par le modèle de langage initial P avec les informations fournies par le modèle possibiliste Π en modifiant la formule d'évaluation du score linguistique de la manière suivante :

$$S_l(H) = P(H)^{\alpha} \times \Pi(H)^{\beta} \tag{7.21}$$

avec α et β les facteurs de la combinaison linéaire, déterminés empiriquement.

Cette approche permet de combiner les quatre modèles de langage proposés de toutes les manières possibles afin d'évaluer la pertinence d'une hypothèse.

7.6 Dispositif expérimental

Les approches proposées dans cet article pour utiliser l'information présente sur le Web pour construire des modèles de langage sont évaluées sur deux tâches de reconnaissance automatique de la parole :

1. la transcription de journaux radiodiffusés en langue anglaise, avec un modèle de langage initial appris sur un corpus approprié
2. la transcription de commentaires audio relatifs à un domaine spécialisé, en anglais, avec un modèle de langage initial appris avec peu de données.

L'intérêt de ces deux tâches est qu'elles représentent bien les deux situations les plus courantes en reconnaissance automatique de la parole. La première consiste à transcrire des données généralistes pour lesquels on dispose de grande quantités de données d'entraînement et de toute l'expertise résultant de plus de dix ans de travaux de

recherche évalués sur ces données. La seconde est typique lorsque l'on souhaite faire de la reconnaissance automatique de la parole de documents pour lesquels rien n'a encore été fait et ou on ne dispose que de peu de données d'entraînement.

Nous allons donc étudier le comportement des approches proposées face à ces deux situations.

7.6.1 Les corpus

Pour représenter les deux situations évoquées précédemment, les corpus AVISON et HUB4 ont été utilisés. Le corpus HUB4 représente la tâche pour laquelle beaucoup de données d'entraînement sont disponibles et le corpus AVISON représente l'autre tâche.

Le corpus AVISON

Le corpus AVISON utilisé ici est exactement le même que celui décrits à la section 5.5.1 du chapitre 5.

Le corpus HUB4

Le corpus utilisé pour la tâche de transcription automatique de journaux radiodiffusés est celui de la campagne d'évaluation HUB4 (Stern, 1997). Cette campagne a pour but dévaluer les performances des systèmes de transcription riches sur des journaux d'information radiodiffusés en langue anglaise.

Les données audio fournies pour la tâche étaient divisées en 3 corpus :

- Le corpus d'entraînement : 50 heures d'émissions d'information américaines transcrites manuellement.
- Le corpus de développement : 3 heures d'émissions d'information américaines transcrites manuellement.
- Le corpus d'évaluation : 2.5 heures d'émissions d'information américaines transcrites manuellement.

Ces données sont des enregistrements d'émissions produites par ABC, CNN, CSPAN, NPR et PRI.

7.6.2 Les systèmes de reconnaissance automatique de la parole

Le système de reconnaissance automatique de la parole continue grand vocabulaire SPEERAL, développé au LIA (Nocéra et al., 2004), est utilisé pour la transcription des deux corpus.

Intégrer directement les estimateurs Web présentés dans l'algorithme de recherche du moteur de reconnaissance automatique de la parole nécessiterait un nombre considérable de requêtes Web. Pour contourner ce problème, nous effectuons un décodage des 100 meilleures hypothèses (100-best) avec le modèle de langage initial appris sur corpus, et nous utilisons les techniques Web proposées ici pour réordonner ces hypothèses. Le moteur de recherche Google est utilisé pour effectuer les requêtes Web.

Le système anglais chirurgical

La manière dont le système SPEERAL est adapté au corpus AVISON est décrite en détails à la section 5.5.2 du chapitre 5.

Cependant, des améliorations ont été apportées au système entre les précédentes expériences et celles présentées dans ce chapitre. Tout d'abord, la phonétisation des mots les plus fréquents a été vérifiée manuellement afin de corriger les erreurs de phonétisation automatiques. Ensuite, les scripts de nettoyage des corpus textuels servant à la modélisation du langage ont été améliorés. Finalement, un nouvel arrivage de transcriptions de vidéos chirurgicales a permis d'augmenter la taille du corpus d'entraînement du modèle de langage.

Ces améliorations ont eu une répercussion sur les performances du système puisque le taux d'erreur mot obtenu sur le corpus de test est descendu à 27.9%.

Le système anglais journalistique

La modélisation acoustique du système est effectuée de manière similaire à celle du système anglais chirurgical. Le jeu de phonèmes est aussi celui utilisé par le *CMU Pronouncing Dictionary*, composé de 39 phonèmes [1]. Les modèles acoustiques sont estimés à partir des enregistrements d'émissions d'information américaines transcrits manuellement et fournis comme corpus d'entraînement de la campagne d'évaluation HUB4.

Le lexique est constitué des 45k mots les plus fréquents dans le corpus d'entraînement des modèles acoustique de la campagne d'évaluation HUB4. La prononciation de ces mots est celle du *CMU Pronouncing Dictionary* si elle existe et sinon elle est obtenue à l'aide du phonétiseur automatique anglais Festival (Taylor et al., 1998).

Le modèle de langage est un 3-gramme obtenu par la combinaison de plusieurs modèles estimés sur des corpus de nature différentes. Le poids des modèles est déterminé pour maximiser la perplexité du modèle final sur le corpus de développement fourni pour la campagne. Trois sources de données sont utilisées pour estimer les modèles atomiques. La première est constituée des transcription du corpus d'entraînement des modèles acoustiques de la campagne. La seconde regroupe l'ensemble des articles de journaux d'information américains disponibles dans le corpus North Ameri-

1. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

can News Text Corpus (Graff, 1995), fourni pour la campagne d'évaluation HUB4. La dernière est constituée des 2.7G mots du corpus Gigaword (Graff et al., 2003), qui contient un recueil de dépêches d'agences de presse en langue anglaise.

Sur les données de test du corpus HUB4, ce système permet d'obtenir un taux d'erreur mot de 27.0%.

7.6.3 Optimisation des paramètres

L'optimisation des différents facteurs de la combinaison linéaire et des formules de repli pose problème car il faut posséder une grande quantité de données pour les estimer correctement. Dans le cadre du corpus AVISON on ne dispose pas de telles données. Pour contourner cette difficulté, l'optimisation des coefficients est effectués à l'aide d'une validation croisée *k-fold* avec $k = 10$. Le fonctionnement du processus peut être résumé ainsi :

1. Le corpus de test est d'abord partitionné en dix sous-corpus
2. Les paramètres sont ensuite optimisés sur neuf sous-corpus et testés sur le dixième
3. L'étape précédente est répétée dix fois, pour que chaque sous-corpus ait été utilisé comme test une et une seule fois.

Le nombre d'erreurs global obtenu sur le corpus de test est alors la somme des erreurs commises sur chacun des sous-corpus.

7.7 Expérimentations

Les différents modèles ont été évalués selon les différentes stratégies de combinaison sur les deux corpus ESTER et AVISON. Les résultats, en terme de taux d'erreur mot, sont présentés dans le tableau 7.4.

Étant donné que les résultats proposés sont le fruit d'un réordonnancement des 100 meilleurs hypothèses produites par le système initial, l'amélioration maximale que peuvent produite les modèles proposés correspond à la meilleure des 100 hypothèses. Le tableau 7.5 contient la mesure oracle effectuée sur les 100 meilleures hypothèses des corpus HUB4 et AVISON. On y trouve le taux d'erreur mot qu'on obtiendrait si l'on choisissait systématiquement la meilleure hypothèse, ainsi que la position moyenne des meilleures hypothèses.

On constate qu'il est possible de réduire le taux d'erreur mot de 5.9% absolus pour le corpus AVISON (de 27.9% à 22.0%) et de 1.1% absolus pour le corpus HUB4 (de 27.0% à 25.9%).

	AVISON		HUB4	
	$n=3$	$n=6$	$n=3$	$n=6$
P_c	27.9	28.0	**27.0**	26.9
P_w	**27.2**	25.5	**27.0**	**26.0**
Π_w	28.5	**25.2**	27.7	26.9
Π_c	28.0	28.1	27.9	27.9
P_w BO P_c	27.6	**26.1**	**26.7**	**26.6**
Π_w BO P_c	**27.5**	26.5	26.8	26.7
Π_c BO P_c	27.6	27.4	26.8	26.9
$P_w \leq \Pi_c$	27.8	25.6	27.2	26.1
$P_c \leq \Pi_w$	28.1	**26.0**	27.0	**26.7**
$P_w \leq \Pi_w$	**27.6**	**25.3**	**27.0**	**26.1**
$P_w + P_c$	**27.1**	25.4	**26.6**	**25.9**
$P_w + \Pi_w$	27.1	**24.7**	27.0	26.1
$P_c + \Pi_w$	27.7	24.8	26.8	26.6
$\Pi_w + \Pi_c + P_w + P_c$	**26.8**	**24.5**	**26.4**	**25.9**

TABLE 7.4 – *Taux d'erreur mots obtenus par les différents modèles de langages proposés et leurs combinaisons sur les corpus AVISON et HUB4*

	AVISON	HUB4
TEM	22.0	25.9
Position	16	22

TABLE 7.5 – *Taux d'erreur mots (TEM) oracle et position moyenne de l'oracle sur les 100 meilleurs hypothèses pour les corpus AVISON et HUB4*

7.7.1 Mesures de probabilité et de possibilité

Les quatre premières lignes du tableau 7.4 contiennent les taux d'erreur mot obtenus avec les modèles probabilistes estimés sur corpus et sur le web (respectivement P_c et P_w) et les modèles possibilistes estimés sur corpus et sur le web (respectivement Π_c et Π_w).

Les performances des modèles de langage peuvent varier fortement avec l'ordre choisi. La nature du modèle ainsi que le volume de données d'apprentissage doivent être en adéquation avec l'ordre pour obtenir des résultats optimaux. Afin d'étudier cet axe de variabilité, des modèles de langage d'ordre 6 ont été estimés et utilisés pour le réordonnancement des 100 meilleurs hypothèses.

Probabilités Corpus

La ligne intitulée P_c du tableau 7.4 contient les taux d'erreur mot obtenus avec les modèles probabilistes estimés sur corpus. Il s'agit de l'évaluation de référence qui correspond à un système de reconnaissance automatique de la parole classique en 2010.

On y retrouve les résultats des systèmes décrits précédemment pour les modèles

d'ordre 3. On constate qu'augmenter l'ordre des modèles de langage probabilistes estimés sur corpus à 6 ne modifie pas les performances des systèmes. Cela indique que les corpus utilisés pour l'estimation des modèles ne sont pas suffisamment volumineux pour permettre une estimation correcte des probabilités n-gramme avec un tel historique.

On se serait attendu à une baisse de performances sur le corpus AVISON étant donné la faible taille du corpus d'entraînement réellement représentatif des données de test. La stabilité des résultats est probablement due à la nature redondante des structures linguistiques que contient le corpus AVISON. Il faut aussi noter que le corpus journalistique qui est intégré dans le modèle de langage final du corpus AVISON est de taille suffisante pour bénéficier de l'augmentation de la portée des n-grammes. L'amélioration obtenue sur ce sous-modèle a pu compenser la perte éventuelle du sous-modèle chirurgical. Finalement, étant donné que les modèles d'ordre 6 ne sont pas utilisés pour réordonner les 100 meilleurs hypothèses produites avec le modèle de langage 3-gramme initial, la marge d'erreur reste limitée.

Probabilités Web

La ligne intitulée P_w du tableau 7.4 contient les taux d'erreur mot obtenus avec les modèles probabilistes estimés sur le Web.

On remarque que le modèle de langage Web donne de meilleurs résultats lorsque l'ordre du modèle augmente. La taille du Web étant sans commune mesure avec celle des corpus habituellement utilisés pour estimer les modèles de langage, on s'attend donc à ce qu'il soit en mesure de fournir des statistiques pertinentes pour l'estimation de probabilités n-grammes de longue portée. On peut également s'interroger sur la pertinence d'une dépendance linguistique aussi longue dans le contexte d'ESTER et d'AVISON.

Le modèle Web d'ordre 3 fournit d'aussi bons résultats que le modèle estimé sur corpus de même ordre pour les deux corpus. On constate même une réduction significative du taux d'erreur mots sur le corpus AVISON. Les modèles Web d'ordre 6 sont par contre systématiquement plus performants que les modèles corpus. Ils permettent de réduire le taux d'erreur mot de 0.9% absolu sur le corpus HUB4 et de 2.5% absolus sur le corpus AVISON. Le modèle semble en apparence plus performant sur le corpus AVISON. Cependant, si l'on compare ces résultats avec ceux obtenus par l'oracle présentés dans le tableau 7.5, on constate que ce modèle permet de s'approcher fortement de la solution optimale pour le corpus HUB4 alors que pour le corpus AVISON, il ne permet qu'environ la moitié de l'amélioration potentielle.

On peut voir sur les figures 7.1 et 7.3 la relation qui existe entre les probabilités 3-grammes estimées sur corpus et celles estimées avec la méthode *ad-hoc* Web présentée à la section 7.2.2. Les n-grammes présents et absents du corpus d'entraînement ont été mis en évidence. Les figures 7.2 et 7.4 contiennent les répartitions, sous forme d'histogramme, des n-grammes en fonction de l'écart entre les probabilités corpus et Web.

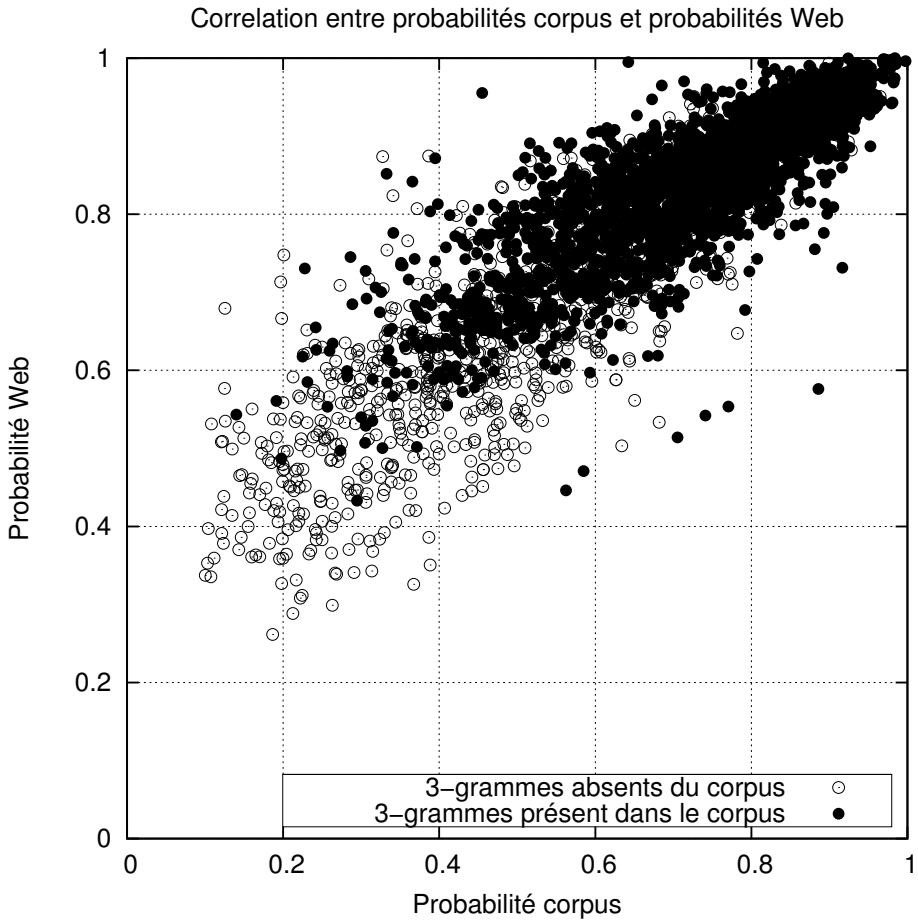

FIGURE 7.1 – *Corrélation entre les probabilités* 3*-grammes corpus et Web observées sur les données de test du corpus HUB4.*

On constate visuellement qu'il existe une forte corrélation entre la mesure Web et la mesure corpus. Le coefficient de corrélation est de 80.3% pour les 3-grammes, ce qui confirme cette première impression. On constate que les nuages de points des n-grammes présents et absents sont bien plus éclatés pour les faibles valeurs de probabilités, ce qui montre que la corrélation est bien plus forte pour les probabilités corpus élevées que pour les probabilités faibles, qui sont estimées avec peu de données. De plus, les histogrammes montrent que les écarts entre probabilités corpus et Web sont bien plus importants pour les n-grammes absents du corpus, qui sont de moins bien estimés que les n-grammes présents. Comme les performances des modèles Web sont meilleures que celle des modèles corpus, cela laisse supposer que l'apport principal du Web consiste en une meilleure modélisation des événements peu fréquents dans le corpus. On voit ici tout l'intérêt d'utiliser un corpus de la taille du Web.

En comparant les nuages de corrélation des 3-grammes et des 6-grammes, on constate que le nuage 6-grammes est plus éclaté. Étant donné que le modèle Web fournit les meilleurs performances, on peut supposer que ses probabilités sont plus justes que celle estimées sur le corpus. L'éclatement du nuage 6-gramme met alors en évidence le

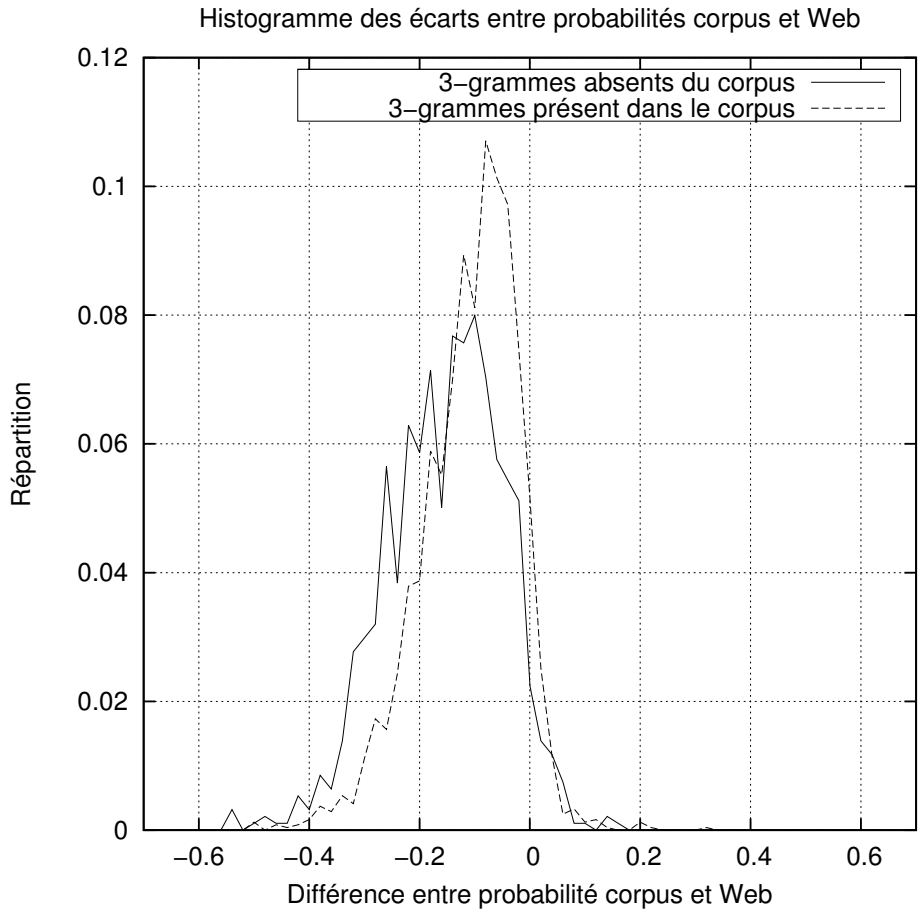

FIGURE 7.2 – *Répartition des écarts entre probabilités* 3*-grammes corpus et Web, observées sur les données de test du corpus HUB4.*

fait que les probabilités estimées sur le corpus sont moins fiables.

On peut déduire de ces expériences que les modèles probabilistes Web sont plus performants lorsque l'ordre est élevé et qu'ils sont dans tous les cas de meilleure qualité que les modèles estimés sur corpus dans les deux situations proposées.

Possibilités

Les lignes intitulées Π_w et Π_c du tableau 7.4 contiennent les taux d'erreur mot obtenus avec les modèles possibilistes estimés respectivement à partir du Web et de corpus.

Comme on pouvait s'y attendre, la mesure de possibilité estimée à partir de corpus est bien moins performante que les trois autres modèles. Ce résultat est tout a fait normal puisqu'un modèle possibiliste contient moins d'informations qu'un modèle probabiliste estimé sur les mêmes données. Cependant, la dégradation des résultats n'est

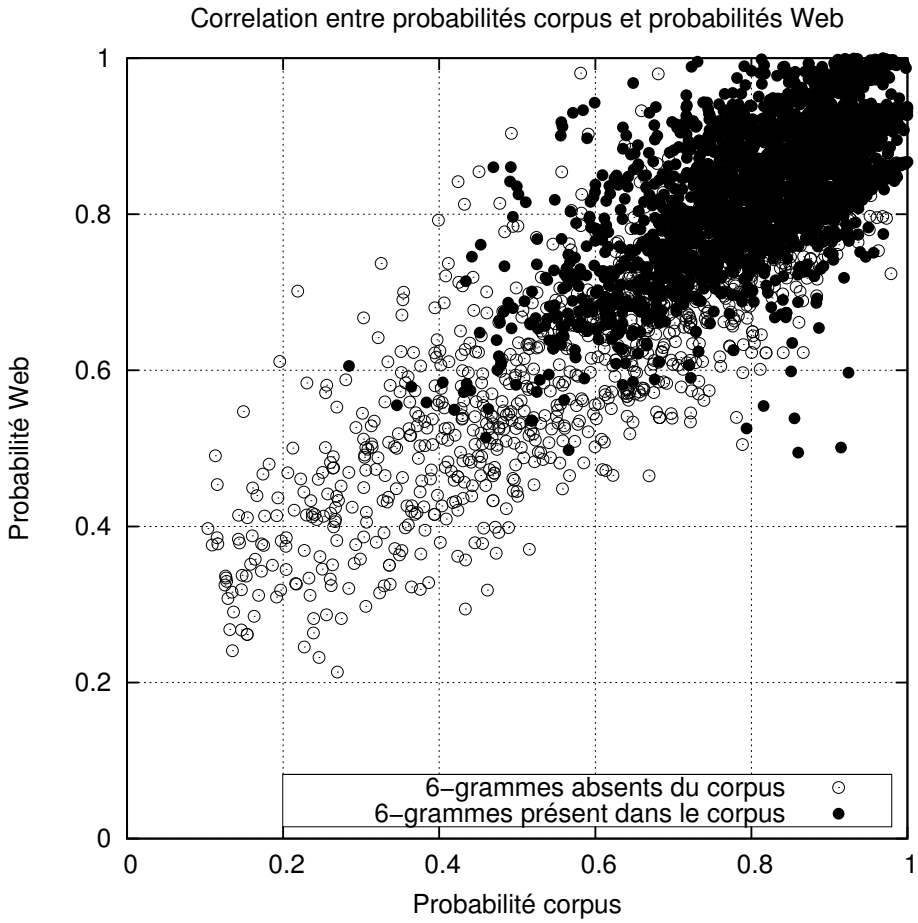

FIGURE 7.3 – *Corrélation entre les probabilités 6-grammes corpus et Web observées sur les données de test du corpus HUB4.*

pas très importante, ce qui indique que cette mesure contient tout de même une bonne partie de l'information utile.

Il faut cependant noter que l'évaluation est effectuée sur la base des 100 meilleurs hypothèses du système de reconnaissance automatique de la parole initial. Un filtrage des hypothèses les plus mauvaises a donc déjà eu lieu, ce qui explique pourquoi la dégradation des résultats n'est pas aussi importante que ce à quoi on pouvait s'attendre.

Le modèle possibiliste estimé sur le Web fournit par contre de meilleurs résultats, ce qui laisse penser que la taille du Web est plus appropriée à ce type de mesure. On note cependant que les performances de la mesure possibiliste Web d'ordre 3 est moins bonne que la mesure possibiliste corpus d'ordre équivalent sur le corpus AVISON. Ce résultat s'explique peut-être par le fait que la faible taille du corpus AVISON permet à la mesure possibiliste d'ordre 3 d'être discriminante alors que sur le Web la plupart des 3-grammes du corpus AVISON y sont présents.

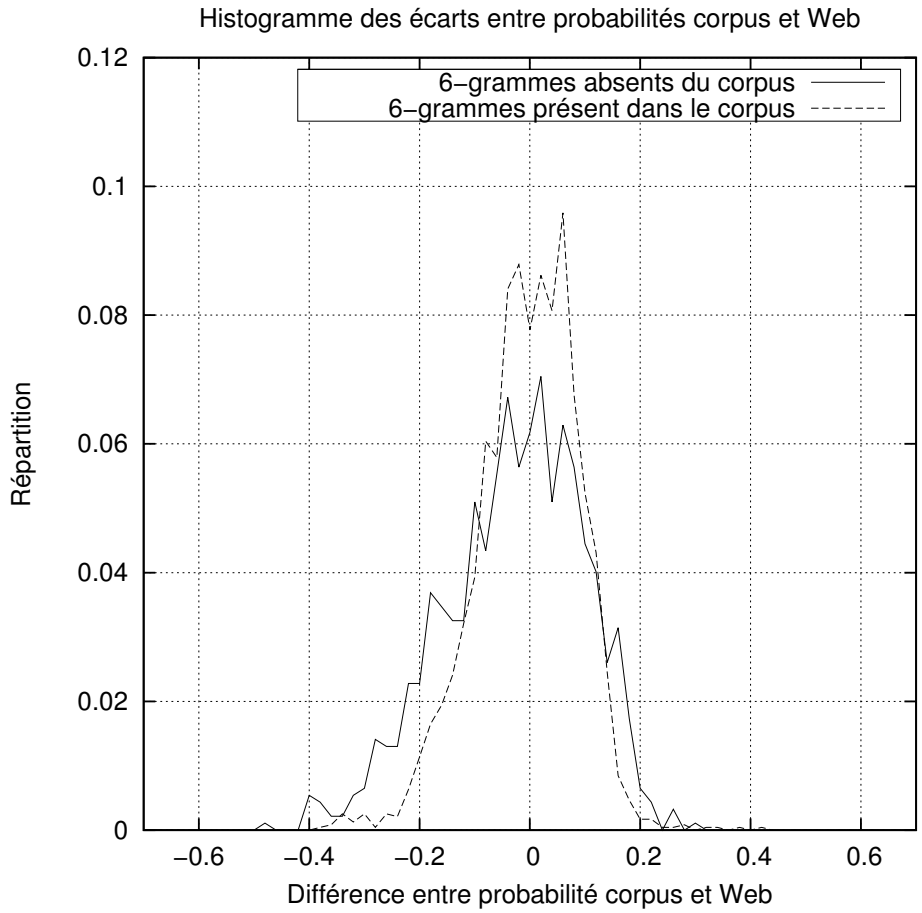

FIGURE 7.4 – *Répartition des écarts entre probabilités 6-grammes corpus et Web, observées sur les données de test du corpus HUB4.*

7.7.2 Possibilités et probabilités Web comme repli du modèle probabiliste corpus

Les lignes intitulées P_w BO P_c, Π_w BO P_c et Π_c BO P_c du tableau 7.4 contiennent respectivement les taux d'erreur mot obtenus avec les modèles probabilistes Web, possibilistes Web et possibilistes corpus utilisés comme repli du modèle de langage probabiliste corpus initial.

On observe que toutes les techniques de modification du repli aboutissent à un gain en terme de taux d'erreur mots par rapport aux modèles initiaux. Ce résultat indique que les probabilités de repli des modèles de langage classique sont moins bonnes que tous les modèles proposés, y compris les possibilités estimées sur corpus.

L'utilisation des possibilités estimées sur corpus pour pondérer le repli du modèle probabiliste conduit à un gain de 0.6% absolu sur le corpus AVISON et de 0.2% absolus sur le corpus HUB4. Ce résultat indique que la mesure possibiliste apporte de l'information qui n'est pas modélisée par le modèle probabiliste initial.

Les résultats obtenus avec la mesure de possibilité estimée sur le Web suit la même tendance. La réduction du taux d'erreur mots est cependant plus importante pour le corpus AVISON puisqu'elle est de 1.9% absolus.

Le remplacement des probabilités de repli par les probabilités fournies par le modèle de langage probabiliste Web permet une amélioration systématique sur les deux corpus. De plus, c'est cette configuration qui permet d'obtenir les meilleurs résultats. On constate une réduction du taux d'erreur mots maximale de 1.9% absolus sur le corpus AVISON et de 0.3% absolus sur le corpus HUB4. Ce résultat montre que les probabilités Web sont de meilleur qualité que les probabilités fournies par le modèle de repli utilisé.

Lorsqu'il est possible d'utiliser un modèle Web, il semble préférable d'adopter la stratégie consistant à remplacer les probabilités de repli par celles du modèle Web. Par contre lorsque les corpus initiaux sont les seuls données accessibles, il est plus performant d'utiliser une mesure possibiliste estimée sur ces données pour pondérer les probabilités de repli du modèle de langage initial.

7.7.3 Possibilités comme bornes supérieures des probabilités

Les lignes intitulées $P_w \leq \Pi_c$, $P_c \leq \Pi_w$ et $P_w \leq \Pi_w$ du tableau 7.4 contiennent respectivement les taux d'erreur mot obtenus avec les modèles possibilistes corpus comme borne supérieur des modèles probabilistes Web, avec les modèles possibilistes Web comme borne supérieur des modèles probabilistes corpus et avec les modèles possibilistes Web comme borne supérieur des modèles probabilistes Web.

L'utilisation des possibilités estimées sur corpus comme borne supérieure des probabilités Web n'apporte aucune amélioration par rapport au modèle Web seul. Par contre, utiliser le modèle possibiliste Web comme borne supérieure du modèle probabiliste corpus permet une amélioration. On obtient ainsi une diminution du taux d'erreur mot de 0.2% absolus sur le corpus HUB4 et de 2% absolus sur le corpus AVISON.

Comme on pouvait s'y attendre, la combinaison des deux meilleurs modèles, les probabilités Web et les possibilités Web, permet d'obtenir les meilleurs performances. On obtient une diminution du taux d'erreur mot de 0.8% absolus sur le corpus HUB4 et de 2.7% sur le corpus AVISON.

Ces expériences confirment que les modèles probabilistes ont tendance à surévaluer la probabilité de nombreux événements et les modèles possibilistes semblent fournir un bon moyen de corriger ce problème.

7.7.4 Combinaison log-linéaire

A partir des quatre modèles que nous avons proposé, il est possible d'obtenir onze combinaisons linéaires de deux, trois et quatre modèles. Nous ne présentons ici que les plus intéressantes. Les quatre dernières lignes du tableau 7.4 contiennent les taux d'erreur mot obtenus avec pour ces combinaisons.

En observant les combinaisons faisant intervenir deux modèles appliquées au corpus HUB4, on constate que la meilleur combinaison est composée du modèle probabiliste Web et du modèle possibiliste Web. Ajouter le modèle probabiliste corpus au modèle probabiliste Web d'ordre 3 permet une réduction de 0.4% absolus du taux d'erreur mots. Par contre, avec les modèles d'ordre 6, le gain est moins important (0.1% absolus). On en déduit que le modèle de langage Web d'ordre élevé est suffisamment bien estimé et l'ajout des informations provenant du corpus ne modifie pas ses performances. Le modèle Web d'ordre 3 semble par contre moins performant puisque l'ajout des probabilités issues du corpus permet une amélioration. De plus, les informations que recèlent les modèles Web et Corpus sont complémentaires puisque l'on observe un gain systématique à les combiner.

On notera également que la combinaison des modèles possibilistes et probabilistes Web permet d'atteindre le gain maximum qu'il est possible avec ces 100 meilleurs hypothèses sur le corpus HUB4. L'ajout du modèle possibiliste Web a permis d'apporter l'information manquante au modèle probabiliste Web pour atteindre l'optimalité.

Sur le corpus AVISON, on constate que la meilleure combinaison est composée du modèle possibiliste et probabiliste Web. Ces résultats sont meilleurs que ceux obtenus avec le modèle probabiliste Web seul. On peut en déduire que les informations qu'ils contiennent sont complémentaires. Le modèle possibiliste apporte donc une information qui n'est pas présente dans le modèle probabiliste.

La combinaison offrant les meilleurs performances est naturellement celle composée des quatre modèles. Sur le corpus AVISON, on obtient ainsi une réduction absolue du taux d'erreur mots de 3.5% absolus par rapport au modèle n-gramme classique et de 0.7% absolus par rapport au meilleur modèle seul (le modèle probabiliste Web). Sur le corpus HUB4, le gain de cette combinaison par rapport au meilleur modèle seul (le modèle probabiliste Web) est de 0.1% absolus. Cette amélioration paraît faible, mais il s'agit de la solution optimale fournie par l'oracle, il n'est donc pas possible de faire mieux.

Ces expériences montrent que la meilleure combinaison de modèles est celle faisant intervenir toutes les connaissances. Cette combinaison peut être vue comme un modèle numérique s'appuyant sur les modèles probabilistes et possibilistes estimés sur le Web et sur corpus.

7.8 Conclusion

Nous avons proposé une nouvelle manière de considérer l'information contenue dans un corpus en utilisant des concepts de la théorie des possibilités. Cette approche permet de tirer parti de l'information qui est absente du corpus afin d'estimer la possibilité d'un événement. Une formule permettant d'estimer cette mesure de possibilité à partir de corpus textuels ou du Web a été proposée.

Nous avons également proposé une manière d'obtenir un modèle de langage n-

gramme fiable à partir de l'ensemble des documents se trouvant sur le Web, en exploitant les statistiques des moteurs de recherche. Lors de nos expériences, ce modèle s'est avéré meilleur que les modèles n-gramme classiques estimés sur des corpus textuels propres.

Plusieurs techniques permettant de combiner ces différents modèles ont été proposées. La meilleure approche semble être une combinaison linéaire des score linguistiques fournies par l'ensemble des modèles.

Nous avons donc proposé deux modèles de langage qui utilisent le Web comme un corpus ouvert que les moteurs de recherche permettent de manipuler. Ces propositions ont été évaluées dans deux contextes de reconnaissance automatique de la parole caractéristiques : la transcription de données d'informations généralistes, pour lesquels on dispose de gros volumes de données textuelles annotées et nettoyées, et la transcription de documents relatifs à un domaine de spécialité, pour lesquels on ne dispose que de très peu de données textuelles propres.

Dans toutes les situations que nous avons étudiées, l'utilisation de modèles possibilistes en complément des modèles probabilistes permet d'améliorer les performances des systèmes de reconnaissance automatique de la parole.

L'utilisation du Web comme corpus ouvert a systématiquement permis d'améliorer les performances des systèmes, il s'agit donc d'une source d'information non négligeable pour les systèmes de reconnaissance automatique de la parole qu'il est nécessaire de prendre en compte.

Conclusion et perspectives

La généralisation des modèles statistiques en reconnaissance automatique de la parole a induit un besoin de corpus d'entraînement toujours plus volumineux. De plus, ces modèles nécessitent souvent un apport régulier de données, lorsque le contenu des documents à transcrire évolue et qu'il est nécessaire de réaliser leur adaptation.

Dans ce contexte, le Web constitue une ressource très intéressante. D'une part il est de taille sans commune mesure avec celle des corpus conventionnels. D'autre part, les données présentes sur le Web sont mises à jour continuellement, ce qui garantit d'y trouver les données les plus récentes.

Nous avons alors proposé dans une première partie une approche pour adapter localement et dynamiquement le lexique du moteur de reconnaissance automatique de la parole à l'aide de données provenant du Web. L'approche développée consiste à effectuer une première transcription des documents et d'y détecter les mots HV. Ensuite, des descripteurs sont extraits du contexte des mots HV et sont utilisés pour formuler des requêtes soumises à un moteur de recherche Web. Les nouveaux mots présents dans les documents retournés sont alors insérés dans le lexique et dans le modèle de langage. Le document audio est finalement transcrit une seconde fois avec les modèles adaptés.

Cette approche possède plusieurs intérêts. D'une part le caractère local de l'adaptation du lexique permet de ne la réaliser que lorsque c'est nécessaire, c'est à dire quand un mot HV est détecté dans les données à transcrire. De plus, l'utilisation du Web comme source de nouveaux mots nous affranchit de tout travail de collecte, de nettoyage et de mise à jour de corpus. Finalement, le caractère dynamique de l'adaptation fait que le processus fonctionne avec n'importe quel document et ne nécessite aucune étape d'adaptation.

Les résultats expérimentaux ont montré qu'il était possible de récupérer jusqu'à 30% des mots HV sur des données de spécialité. Comme les mots HV sont très largement composés de termes techniques et d'entités nommées, ils ont de plus grande chances de figurer dans les requêtes des utilisateurs. Si le moteur de reconnaissance automatique de la parole est couplé à un processus d'indexation des documents, les mots HV récupérés permettront de ne pas manquer les documents les contenant s'ils font partie des requêtes utilisateur.

Les performances des modèles de langage statistiques sont proportionnelles à la

quantité de données ayant servi à les estimer. Cette caractéristique fait du Web un corpus de choix pour leur estimation et leur adaptation.

La plupart des travaux dans ce domaine consistent à sélectionner un ensemble de documents sur le Web correspondants à la thématique des documents à transcrire et à les exploiter comme un corpus fermé, à l'aide des techniques classiques d'estimation et d'adaptation des modèles de langage.

Nous avons alors proposé de reconsidérer la manière dont l'information disponible sur le Web est utilisée. Nous avons proposé un nouveau type de modèle de langage Web pouvant être estimés dynamiquement grâce aux statistiques des moteurs de recherche. Un modèle reposant sur des concepts de la théorie des possibilités a également été proposé afin de mieux tirer parti des informations disponibles sur le Web. Au lieu d'utiliser les fréquences de séquences de mots pour estimer des probabilités, ce modèle utilise la non existence de séquences de mots pour estimer des possibilités.

Plusieurs stratégies permettant de combiner les modèles possibilistes avec des modèles probabilistes estimés à partir du Web et de corpus ont été proposées. Il a été montré que les modèles Web proposés sont plus performants que les modèles corpus classiques. Les expériences montrent également que les informations modélisées par les modèles probabilistes et possibilistes sont complémentaires. Combiner ces modèles apporte une amélioration systématique des performances du systèmes de reconnaissance automatique de la parole. Il a également été montré que l'amélioration induite par les modèles possibilistes était plus importante dans des situations où le modèle de langage initial était estimé sur peu de données, comme c'est le cas avec les langues peu dotées ou les domaines de spécialité.

Nous avons donc proposé l'utilisation du Web comme corpus ouvert pour améliorer deux aspects de la RAP : l'adaptation dynamique du lexique au contenu des documents et la modélisation du langage. Les modèles proposés s'appuient sur l'utilisation de moteurs de recherche Web comme moyen de manipuler la collection de documents Web, qui est en perpétuelle évolution. Des données peuvent ainsi être récupérées et des mesures calculées dynamiquement à partir des information fournies par les moteurs de recherche.

La continuité de ces travaux pourrait concerner l'intégration du modèle possibiliste directement dans l'algorithme de décodage, ce qui permettrait d'interdire les chemins impossibles très tôt dans le processus. Cette stratégie offrirait un gain de temps d'exécution et probablement de meilleurs performances.

Il serait également intéressant d'adapter le modèle probabiliste pour qu'il soit encore plus complémentaire avec le modèle possibiliste. La théorie des possibilités dispose d'une mesure de nécessité qui est liée à la mesure de possibilité. Elle modélise une information qui peut être rapprochée de celle que mesurent les modèles probabilistes actuels. Il serait donc intéressant de développer une approche pour estimer conjointement un modèle de nécessité et de possibilité Web.

Finalement, ce manuscrit aura permis d'ajouter 56218 mots au contenu déjà énorme du Web et certains seront peut-être capturés par un processus d'adaptation automa-

tique du lexique. De plus, ces données contribueront probablement de manière implicite à améliorer l'ensemble des modèles de langage exploitant cette ressource ouverte.

Liste des illustrations

Liste des tableaux

Bibliographie

(Adda-Decker, 2006) M. Adda-Decker, 2006. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *JEP*, 389–400.

(Allauzen et Gauvain, 2003) A. Allauzen et J. Gauvain, 2003. Adaptation automatique du modèle de langage d'un système de transcription de journaux parlés. *Traitement Automatique des langues 44*(1), 11–31.

(Allauzen et Gauvain, 2005a) A. Allauzen et J. Gauvain, 2005a. Diachronic vocabulary adaptation for broadcast news transcription. Dans les actes de *Ninth European Conference on Speech Communication and Technology*, 1305–1308.

(Allauzen et Gauvain, 2005b) A. Allauzen et J. Gauvain, 2005b. Open vocabulary asr for audiovisual document indexation. Dans les actes de *Proc. ICASSP*, 1013– 1016.

(Allen, 1976) J. Allen, 1976. Synthesis of speech from unrestricted text. *Proceedings of the IEEE 64*(4), 433 – 442.

(Asadi et al., 1990) A. Asadi, R. Schwartz, et J. Makhoul, 1990. Automatic detection of new words in a large vocabulary continuous speech recognition system. Dans les actes de *Proc. Int Acoustics, Speech, and Signal Processing ICASSP-90. Conf*, 125–128.

(Bahl et al., 1986) L. Bahl, P. Brown, P. De Souza, et R. Mercer, 1986. Maximum mutual information estimation of hidden markov model parameters for speech recognition. Dans les actes de *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, Volume 11, 49–52. IEEE.

(Bahl et al., 1983) L. Bahl, F. Jelinek, et R. Mercer, 1983. A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2), 179–190.

(Banko et Brill, 2001) M. Banko et E. Brill, 2001. Scaling to very very large corpora for natural language disambiguation. Dans les actes de *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 26–33. Association for Computational Linguistics.

(Baum et al., 1970) L. Baum, T. Petrie, G. Soules, et N. Weiss, 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics 41*(1), 164–171.

(Bazzi, 2002) I. Bazzi, 2002. *Modelling Out-of-Vocabulary Words for RobustSpeech Recognition*. Thèse de Doctorat, Department of Eelectrical Engeneering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

(Bazzi et Glass, 2000) I. Bazzi et J. R. Glass, 2000. Modeling out-of-vocabulary words for robust speech recognition. Dans les actes de *Proc. ICSLP*, 401–404.

(Beaulieu et al., 1997) M. Beaulieu, M. Gatford, H. Xiangji, S. Robertson, S. Walker, et P. Williams, 1997. Okapi at trec-5. *NIST special publication* (500238), 143–165.

(Bechet, 2001) F. Bechet, 2001. Lia phon : un systeme complet de phonétisation de textes. *Traitement automatique des langues 42*(1), 47–67.

(Bellegarda et al., 1996) J. Bellegarda, J. Butzberger, Y. Chow, N. Coccaro, et D. Naik, 1996. A novel word clustering algorithm based on latent semantic analysis. Dans les actes de *icassp*, 172–175. IEEE.

(Bengio et al., 2006) Y. Bengio, H. Schwenk, J. Senécal, F. Morin, et J. Gauvain, 2006. Neural probabilistic language models. *Innovations in Machine Learning*, 137–186.

(Berger et Miller, 1998) A. Berger et R. Miller, 1998. Just-in-time language modelling. Dans les actes de *Proc. ICASSP*, Volume 2, 705–708.

(Berger et al., 1996) A. Berger, V. Pietra, et S. Pietra, 1996. A maximum entropy approach to natural language processing. *Computational linguistics 22*(1), 39–71.

(Bertoldi et Federico, 2001) N. Bertoldi et M. Federico, 2001. Broadcast news lm adaptation using contemporary texts. Dans les actes de *Proc. ECSCT*, 239–242.

(Biber, 1988) D. Biber, 1988. *Variation across speech and writing*. Cambridge University Press.

(Biber, 1991) D. Biber, 1991. *Variation across speech and writing*. Cambridge Univ Press.

(Bigi et al., 2004) B. Bigi, Y. Huang, et R. De Mori, 2004. Vocabulary and language model adaptation using information retrieval. Dans les actes de *Proc. ICSLP*, 1361–1364.

(Bisani et Ney, 2005) M. Bisani et H. Ney, 2005. Open vocabulary speech recognition with flat hybrid models. Dans les actes de *Proc. Interspeech*, 725–728.

(Bogert et al., 1963) B. Bogert, M. Healy, et J. Tukey, 1963. The quefrency alanysis of time series for echoes : Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. Dans les actes de *Proceedings of the Symposium on Time Series Analysis*, 209–243.

(Boulianne et Dumouchel, 2001) G. Boulianne et P. Dumouchel, 2001. Out-of-vocabulary word modeling using multiple lexical fillers. Dans les actes de *Proc. IEEE Workshop Automatic Speech Recognition and Understanding ASRU '01*, 226–229.

(Brown et al., 1990) P. Brown, J. Cocke, S. Pietra, V. Pietra, F. Jelinek, J. Lafferty, R. Mercer, et P. Roossin, 1990. A statistical approach to machine translation. *Computational linguistics 16*(2), 79–85.

(Brown et al., 1992a) P. Brown, P. Desouza, R. Mercer, V. Pietra, et J. Lai, 1992a. Class-based n-gram models of natural language. *Computational linguistics 18*(4), 467–479.

(Brown et al., 1992b) P. Brown, V. Pietra, R. Mercer, S. Pietra, et J. Lai, 1992b. An estimate of an upper bound for the entropy of english. *Computational Linguistics 18*(1), 31–40.

(Bulyko et al., 2003) I. Bulyko, M. Ostendorf, et A. Stolcke, 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. Dans les actes de *Proc. HLT-NAACL*, Volume 2, 7–9.

(Burget et al., 2008) L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, et J. Cernocky, 2008. Combination of strongly and weakly constrained recognizers for reliable detection of oovs. Dans les actes de *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2008*, 4081–4084.

(Cai et Zhu, 2004) T. Cai et J. Zhu, 2004. Oov rejection algorithm based on class-fusion support vector machine for speech recognition. Dans les actes de *Proc. Int Machine Learning and Cybernetics Conf*, Volume 6, 3695–3699.

(Cao et Li, 2002) Y. Cao et H. Li, 2002. Base noun phrase translation using web data and the em algorithm. Dans les actes de *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 1–7. Association for Computational Linguistics.

(Chelba et Jelinek, 2000) C. Chelba et F. Jelinek, 2000. Structured language modeling. *Computer Speech & Language 14*(4), 283–332.

(Chen et al., 2003) L. Chen, J. Gauvain, L. Lamel, et G. Adda, 2003. Unsupervised language model adaptation for broadcast news. Dans les actes de *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, Volume 1, I–220. IEEE.

(Chen et Goodman, 1999) S. Chen et J. Goodman, 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language 13*, 359–394.

(Church et Gale, 1991) K. Church et W. Gale, 1991. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech & Language 5*(1), 19–54.

(Clarkson et Robinson, 1997) P. Clarkson et A. Robinson, 1997. Language model adaptation using mixtures and an exponentially decaying cache. Dans les actes de *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Volume 2, 799–802.

(Clarkson et Robinson, 1999) P. Clarkson et T. Robinson, 1999. Towards improved language model evaluation measures. Dans les actes de *Proceedings of EUROSPEECH 99, 6th european conference on speech communication and technology*, Volume 5, 1927–1930.

(Cover et al., 1991) T. Cover, J. Thomas, J. Wiley, et al., 1991. *Elements of information theory*. Wiley Online Library.

(Darroch et Ratcliff, 1972) J. Darroch et D. Ratcliff, 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics 43*(5), 1470–1480.

(Davis et Mermelstein, 1980) S. Davis et P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on 28*(4), 357–366.

(de Cooman, 1997) G. de Cooman, 1997. Possibility theory I : the measure- and integral-theoretic groundwork. *International Journal of General Systems 25*, 291–323.

(Decadt et al., 2002) B. Decadt, J. Duchateau, W. Daelemans, et P. Wambacq, 2002. Transcription of out-of-vocabulary words in large vocabulary speech recognition based on phoneme-to-grapheme conversion. Dans les actes de *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, Volume 1, 861–861. IEEE.

(Deligne et Bimbot, 1995) S. Deligne et F. Bimbot, 1995. Language modeling by variable length sequences : Theoretical formulation and evaluation of multigrams. Dans les actes de *icassp*, 169–172. IEEE.

(Della Pietra et al., 1992) S. Della Pietra, V. Della Pietra, R. Mercer, et S. Roukos, 1992. Adaptive language modeling using minimum discriminant estimation. Dans les actes de *icassp*, 633–636. IEEE.

(Dempster et al., 1977) A. Dempster, N. Laird, et D. Rubin, 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

(Divay et Vitale, 1997) M. Divay et A. Vitale, 1997. Algorithms for grapheme-phoneme translation for english and french : Applications for database searches and speech synthesis. *Computational linguistics 23*(4), 495–523.

(Dubois, 2006) D. Dubois, 2006. Possibility theory and statistical reasoning. *Computational Statistics and Data Analysis 21*, 47–69.

(Dubois et Prade, 1988) D. Dubois et H. Prade, 1988. *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. Plenum Press.

(Dumais et al., 2002) S. Dumais, M. Banko, E. Brill, J. Lin, et A. Ng, 2002. Web question answering : Is more always better ? Dans les actes de *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 291–298. ACM.

(Federico, 1999) M. Federico, 1999. Efficient language model adaptation through mdi estimation. Dans les actes de *Proc. of EUROSPEECH*, 1583–1586.

(Federico et Bertoldi, 2004) M. Federico et N. Bertoldi, 2004. Broadcast news LM adaptation over time. *Computer Speech & Language 18*(4), 417–435.

(Freund et Schapire, 1995) Y. Freund et R. Schapire, 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. Dans les actes de *Computational learning theory*, 23–37. Springer.

(Gales, 1998) M. Gales, 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language 12*, 75–98.

(Galliano et al., 2005) S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. Bonastre, et G. Gravier, 2005. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *INTERSPEECH*, 1149–1152.

(Gauvain et al., 2002) J. Gauvain, L. Lamel, et G. Adda, 2002. The limsi broadcast news transcription system. *Speech Communication 37*(1-2), 89–108.

(Gauvain et al., 1995) J. Gauvain, L. Lamel, et M. Adda-Decker, 1995. Developments in continuous speech dictation using the ARPA WSJ task. Dans les actes de *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Volume 1, 65–68.

(Gauvain et Lee, 1994) J. Gauvain et C. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and Audio Processing, IEEE Transactions on 2*(2), 291–298.

(Geutner et al., 1998) P. Geutner, M. Finke, P. Scheytt, A. Waibel, et H. Wactlar, 1998. Transcribing multilingual broadcast news using hypothesis driven lexical adaptation. Dans les actes de *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

(Good, 1953) I. Good, 1953. The population frequencies of species and the estimation of population parameters. *Biometrika 40*(3-4), 237.

(Goodman, 2006) J. Goodman, 2006. A bit of progress in language modeling extended version. Rapport technique, Microsoft Research.

(Gotoh et Renals, 1999) Y. Gotoh et S. Renals, 1999. Topic-based mixture language modelling. *Natural Language Engineering 5*(4), 355–375.

(Graff, 1995) D. Graff, 1995. North american news text corpus. *Corpus number LDC95T21, Linguistic Data Consortium*.

(Graff et al., 2003) D. Graff, J. Kong, K. Chen, et K. Maeda, 2003. English gigaword corpus. *Corpus number LDC2003T05, Linguistic Data Consortium*.

(Gravier et al., 2004) G. Gravier, J. F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, et K. Choukri, 2004. The ester evaluation campaign of rich transcription of french broadcast news. Dans les actes de *Proc. LREC*, 885–888.

(Grefenstette, 1999) G. Grefenstette, 1999. The world wide web as a resource for example-based machine translation tasks. *Aslib Conference on Translating and the Computer*.

(Gulli et Signorini, 2005) A. Gulli et A. Signorini, 2005. The indexable web is more than 11.5 billion pages. Dans les actes de *Special interest tracks and posters of the 14th international conference on World Wide Web*, 902–903. ACM.

(Hazen et Bazzi, 2001) T. J. Hazen et I. Bazzi, 2001. A comparison and combination of methods for oov word detection and word confidence scoring. Dans les actes de *Proc. ICASSP*, Volume 1, 397–400.

(Hearst et al., 1998) M. Hearst, S. Dumais, E. Osman, J. Platt, et B. Scholkopf, 1998. Support vector machines. *Intelligent Systems and their Applications, IEEE 13*(4), 18–28.

(Heigold et al., 2005) G. Heigold, W. Macherey, R. Schluter, et H. Ney, 2005. Minimum exact word error training. Dans les actes de *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, 186–190. IEEE.

(Hermansky, 1990) H. Hermansky, 1990. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America 87*(4), 1738–1752.

(Ito et al., 2008) A. Ito, Y. Kajiura, S. Makino, et M. Suzuki, 2008. An unsupervised language model adaptation based on keyword clustering and query availability estimation. Dans les actes de *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, 1412–1418. IEEE.

(Ito et al., 2009) A. Ito, Y. Kajiura, M. Suzuki, et S. Makino, 2009. Automatic query generation and query relevance measurement for unsupervised language model adaptation of speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing 1*, 1–13.

(Jelinek, 1969) F. Jelinek, 1969. Fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development 13*(6), 675–685.

(Jelinek, 1976) F. Jelinek, 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE 64*(4), 532–556.

(Jelinek, 1990) F. Jelinek, 1990. Self-organized language modeling for speech recognition. *Readings in speech recognition*, 450–506.

(Jelinek et Mercer, 1980) F. Jelinek et R. Mercer, 1980. Interpolated estimation of Markov source parameters from sparse data. Dans les actes de *Proceedings of the workshop on Pattern Recognition in Practice*, 381–397. Amsterdam.

(Ji et al., 2006) G. Ji, J. Bilmes, K. Kirchhoff, et C. Manning, 2006. Graphical model representations of word lattices. Dans les actes de *Spoken Language Technology Workshop, 2006. IEEE*, 162–165. IEEE.

(Jiampojamarn et al., 2008) S. Jiampojamarn, C. Cherry, et G. Kondrak, 2008. Joint processing and discriminative training for letter-to-phoneme conversion. Dans les actes de *Proc. ACL*, 905–913.

(Kajiura et al., 2006) Y. Kajiura, M. Suzuki, A. Ito, et S. Makino, 2006. Generating search query in unsupervised language model adaptation using www. *The Journal of the Acoustical Society of America 120*(5), 3043–3044.

(Kaplan et Kay, 1994) R. Kaplan et M. Kay, 1994. Regular models of phonological rule systems. *Computational linguistics 20*(3), 331–378.

(Katz, 1987) S. Katz, 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on 35*(3), 400–401.

(Keller et Lapata, 2003) F. Keller et M. Lapata, 2003. Using the web to obtain frequencies for unseen bigrams. *Computational linguistics 29*, 459–484.

(Kemp et Waibel, 1998) T. Kemp et A. Waibel, 1998. Reducing the oov rate in broadcast news speech recognition. Dans les actes de *Fifth International Conference on Spoken Language Processing*, 757–761.

(Kneser et Ney, 1995a) R. Kneser et H. Ney, 1995a. Improved backing-off for m-gram language modeling. Dans les actes de *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, Volume 1, 181–184. IEEE.

(Kneser et al., 1997) R. Kneser, J. Peters, et D. Klakow, 1997. Language model adaptation using dynamic marginals. Dans les actes de *European Conference on Speech Communication and Technology*, 1971–1974.

(Kneser et Steinbiss, 1993) R. Kneser et V. Steinbiss, 1993. On the dynamic adaptation of stochastic language models. Dans les actes de *icassp*, 586–589. IEEE.

(Kuhn et De Mori, 1990) R. Kuhn et R. De Mori, 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 570–583.

(Kullback et Leibler, 1951) S. Kullback et R. Leibler, 1951. On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

(Langlois et al., 2003) D. Langlois, A. Brun, K. Sma
"ıli, et J. Haton, 2003. Événements impossibles en modélisation stochastique du langage. *Traitement Automatique des Langues 44*(1), 33–61.

(Lapata et Keller, 2005) M. Lapata et F. Keller, 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing 2*, 3.

(Lau et al., 1993) R. Lau, R. Rosenfeld, et S. Roukos, 1993. Trigger-based language models : A maximum entropy approach. Dans les actes de *icassp*, 45–48. IEEE.

(Lauritzen, 1996) S. Lauritzen, 1996. *Graphical models*. Oxford University Press.

(Lecorvé et al., 2008) G. Lecorvé, G. Gravier, et P. Sebillot, 2008. An unsupervised web-based topic language model adaptation method. Dans les actes de *Proc. ICASSP*, 5081–5084.

(Lecouteux et al., 2009) B. Lecouteux, G. Linarès, et B. Favre, 2009. Combined low level and high level features for Out-Of-Vocabulary Word detection. Dans les actes de *International Conference on Speech Communication and Technologies, Interspeech*, 1187–1190.

(Lin et al., 2007) H. Lin, J. Bilmes, D. Vergyri, et K. Kirchhoff, 2007. Oov detection by joint word/phone lattice alignment. Dans les actes de *Proc. ASRU Automatic Speech Recognition & Understanding IEEE Workshop*, 478–483.

(Liu et al., 2007) C. Liu, K. Thambiratnam, et F. Seide, 2007. Online vocabulary adaptation using limited adaptation data. Dans les actes de *Eighth Annual Conference of the International Speech Communication Association*, 1821–1824.

(Manning et al., 2008) C. Manning, P. Raghavan, et H. Schütze, 2008. *Introduction to Information Retrieval*. Cambridge University Press.

(Manning et al., 1999) C. Manning, H. Schütze, et MITCogNet, 1999. *Foundations of statistical natural language processing*. MIT Press.

(Markel et Gray, 1974) J. Markel et J. Gray, 1974. Linear prediction of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing 3*, 207–217.

(Martins et al., 2007a) C. Martins, A. Teixeira, et J. Neto, 2007a. Dynamic language modeling for a daily broadcast news transcription system. Dans les actes de *Proceedings of the ASRU*, 165–170.

(Martins et al., 2010) C. Martins, A. Teixeira, et J. Neto, 2010. Dynamic language modeling for European Portuguese. *Computer Speech & Language 24*(4), 750–773.

(Meng et al., 2010) S. Meng, K. Thambiratnam, Y. Lin, L. Wang, G. Li, et F. Seide, 2010. Vocabulary and language model adaptation using just one speech file. Dans les actes de *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 5410–5413. IEEE.

(Munteanu et al., 2007) C. Munteanu, G. Penn, et R. Baecker, 2007. Web-based language modelling for automatic lecture transcription. Dans les actes de *Proceedings of 8th Annual Conference of the International Speech Communication Association*, 2353–2356.

(Myers, 1980) C. Myers, 1980. *A comparative study of several dynamic time warping algorithms for speech recognition*. Thèse de Doctorat, Massachusetts Institute of Technology.

(Nakov et Hearst, 2005b) P. Nakov et M. Hearst, 2005b. Using the web as an implicit training set : application to structural ambiguity resolution. Dans les actes de *Proceedings of HLT/EMNLP'05*, 835–842.

(Ney et Essen,) H. Ney et U. Essen. On smoothing techniques for bigram-based natural language modelling. Dans les actes de *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 825–828. IEEE.

(Ney et al., 1994) H. Ney, U. Essen, et R. Kneser, 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language 8*(1), 1–38.

(Niesler et al., 1998) T. Niesler, E. Whittaker, et P. Woodland, 1998. Comparison of part-of-speech and automatically derived category-based language models for speech recognition. Dans les actes de *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Volume 1, 177–180. IEEE.

(Nocéra et al., 2004) P. Nocéra, C. Fredouille, G. Linarès, D. Matrouf, S. Meignier, J. Bonastre, D. Massonié, et F. Béchet, 2004. The LIA's french broadcast news transcription system. Dans les actes de *SWIM : Lectures by Masters in Speech Processing*.

(Nocera et al., 2002a) P. Nocera, G. Linarès, et D. Massonié, 2002a. Principes et performances du décodeur parole continue speeral. Dans les actes de *Proc. JEP*.

(Nocera et al., 2002b) P. Nocera, G. Linarès, D. Massonie, et L. Lefort, 2002b. Phoneme lattice based a$*$ search algorithm for speech recognition. 83–111. Springer.

(Ohtsuki et al., 2005) K. Ohtsuki, N. Hiroshima, M. Oku, et A. Imamura, 2005. Unsupervised vocabulary expansion for automatic transcription of broadcast news. Dans les actes de *Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1021–1024.

(Paul, 1992) D. Paul, 1992. An efficient a* stack decoder algorithm for continuous speech recognition with a stochastic language model. Dans les actes de *Proceedings of the workshop on Speech and Natural Language*, 405–409. Association for Computational Linguistics.

(Ponte et Croft, 1998) J. Ponte et W. Croft, 1998. A language modeling approach to information retrieval. Dans les actes de *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 275–281. ACM.

(Povey et Woodland, 2002) D. Povey et P. Woodland, 2002. Minimum phone error and i-smoothing for improved discriminative training. Dans les actes de *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, Volume 1, I–105. IEEE.

(Rabiner, 1989) L. Rabiner, 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

(Rama et al., 2009) T. Rama, A. Singh, et S. Kolachina, 2009. Modeling letter-to-phoneme conversion as a phrase based statistical machine translation problem with minimum error rate training. Dans les actes de *The NAACL Student Research Workshop*.

(Rao et al., 1995) P. Rao, M. Monkowski, et S. Roukos, 1995. Language model adaptation via minimum discrimination information. Dans les actes de *icassp*, 161–164. IEEE.

(Rastrow et al., 2009) A. Rastrow, A. Sethy, et B. Ramabhadran, 2009. A new method for oov detection using hybrid word/fragment system. Dans les actes de *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2009*, 3953–3956.

(Resnik, 1999) P. Resnik, 1999. Mining the web for bilingual text. Dans les actes de *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 527–534. Association for Computational Linguistics.

(Rigau et al., 2002) G. Rigau, B. Magnini, E. Agirre, P. Vossen, et J. Carroll, 2002. Meaning : A roadmap to knowledge technologies. Dans les actes de *COLING-02 on A roadmap for computational linguistics-Volume 13*, 1–7. Association for Computational Linguistics.

(Rijsbergen, 1979) C. J. V. Rijsbergen, 1979. *Information Retrieval* (2nd ed.). Butterworth-Heinemann.

(Ristad, 1998) E. Ristad, 1998. A natural law of succession. *International Symposium on Information Theory*, 445–467.

(Rogina et Schaaf, 2002) I. Rogina et T. Schaaf, 2002. Lecture and presentation tracking in an intelligent meeting room. Dans les actes de *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, 47–52. IEEE.

(Rosenfeld, 1994) R. Rosenfeld, 1994. A hybrid approach to adaptive statistical language modeling. Dans les actes de *Proceedings of the workshop on Human Language Technology*, 76–81. Association for Computational Linguistics.

(Rosenfeld, 1995) R. Rosenfeld, 1995. Optimizing lexical and n-gram coverage via judicious use of linguistic data. Dans les actes de *Fourth European Conference on Speech Communication and Technology*, 1763–1766.

(San-Segundo et al., 2001) R. San-Segundo, B. Pellom, K. Hacioglu, W. Ward, et J. Pardo, 2001. Confidence measures for spoken dialogue systems. Dans les actes de *icassp*, 393–396. IEEE.

(Sarikaya et al., 2005a) R. Sarikaya, Y. Gao, M. Picheny, et H. Erdogan, 2005a. Semantic confidence measurement for spoken dialog systems. *Speech and Audio Processing, IEEE Transactions on 13*(4), 534–545.

(Sarikaya et al., 2005b) R. Sarikaya, A. Gravano, et Y. Gao, 2005b. Rapid language model development using external resources for new spoken dialog domains. 573–576.

(Schwartz et Austin, 1991) R. Schwartz et S. Austin, 1991. A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses. Dans les actes de *icassp*, 701–704. IEEE.

(Schwartz et Chow, 1990) R. Schwartz et Y. Chow, 1990. The n-best algorithms : an efficient and exact procedure for finding the n most likely sentence hypotheses. Dans les actes de *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, 81–84. IEEE.

(Sejnowski et Rosenberg, 1987) T. Sejnowski et C. Rosenberg, 1987. Parallel networks that learn to pronounce english text. *Complex systems 1*(1), 145–168.

(Senay et al., 2011) G. Senay, G. Linarès, et B. Lecouteux, 2011. A segment-level confidence measure for spoken document retrieval. Dans les actes de *Proceedings of ICASSP*, 5548–5551.

(Shannon et Weaver, 1959) C. Shannon et W. Weaver, 1959. *The mathematical theory of communication*. University of Illinois Press.

(Singh-Miller et Collins, 2007) N. Singh-Miller et C. Collins, 2007. Trigger-based language modeling using a loss-sensitive perceptron algorithm. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing*, Volume 4, IV–25. IEEE.

(Soong et Huang, 1991) F. Soong et E. Huang, 1991. A tree-trellis based fast search for finding the n-best sentence hypotheses in continuous speech recognition. Dans les actes de *icassp*, 705–708. IEEE.

(Soricut et Brill, 2006) R. Soricut et E. Brill, 2006. Automatic question answering using the web : Beyond the factoid. *Information Retrieval 9*(2), 191–206.

(Spärck Jones, 1972) K. Spärck Jones, 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28*(1), 11–21.

(Stern, 1997) R. Stern, 1997. Specifications of the 1996 hub-4 broadcast news evaluation. Dans les actes de *Proc. DARPA Speech Recognition Workshop*, 7–14.

(Sun et al., 2003) H. Sun, G. Zhang, F. Zheng, et M. Xu, 2003. Using word confidence measure for OOV words detection in a spontaneous spoken dialog system. Dans les actes de *Eighth European Conference on Speech Communication and Technology*, 2713–2716.

(Suzuki et al., 2006) M. Suzuki, Y. Kajiura, A. Ito, et S. Makino, 2006. Unsupervised language model adaptation based on automatic text collection from WWW. Dans les actes de *Ninth International Conference on Spoken Language Processing*, 2202–2205.

(Taylor, 2005) P. Taylor, 2005. Hidden markov models for grapheme to phoneme conversion. Dans les actes de *Proceedings of the 9th European Conference on Speech Communication and Technology*, 1973–1976.

(Taylor et al., 1998) P. A. Taylor, A. Black, et R. Caley, 1998. The architecture of the festival speech synthesis system. Dans les actes de *Proceedings of the third ESCA/COCOSDA Workshop on Speech Synthesis*, 147–152.

(Timothy J. Hazen et Seneff, 2002) J. P. Timothy J. Hazen, Theresa Burianek et S. Seneff, 2002. Recognition confidence scoring for use in speech understanding systems. *Computer Speech & Language 16*, 49–67.

(Torkkola, 1993) K. Torkkola, 1993. An efficient way to learn english grapheme-to-phoneme rules automatically. Dans les actes de *International Conference on Acoustics, Speech, and Signal Processing*, Volume 2, 199–202. IEEE.

(Tsiartas et al., 2010) A. Tsiartas, P. G. Georgiou, et S. S. Narayanan, 2010. Language model adaptation using www documents obtained by utterance-based queries. Dans les actes de *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5406–5409.

(Valtchev et al., 1997) V. Valtchev, J. Odell, P. Woodland, et S. Young, 1997. Mmie training of large vocabulary recognition systems. *Speech Communication 22*(4), 303–314.

(Vaufreydaz et al., 1999) D. Vaufreydaz, M. Akbar, et J. Rouillard, 1999. Internet documents : a rich source for spoken language modelling. Dans les actes de *IEEE Workshop ASRU'99 (Automatic Speech Recognition and Understanding)*, Keystone - Colorado (USA), 277–281.

(Wan et Hain, 2006) V. Wan et T. Hain, 2006. Strategies for language model web-data collection. Dans les actes de *Proc. ICASSP*, Volume 1, 1069–1072.

(Wessel et al., 2001) F. Wessel, R. Schluter, K. Macherey, et H. Ney, 2001. Confidence measures for large vocabulary continuous speech recognition. *Speech and Audio Processing, IEEE Transactions on 9*(3), 288–298.

(White et al., 2008) C. White, G. Zweig, L. Burget, P. Schwarz, et H. Hermansky, 2008. Confidence estimation, oov detection and language id using phone-to-word transduction and phone-level alignments. Dans les actes de *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2008*, 4085–4088.

(Yang et Pedersen, 1997) Y. Yang et J. Pedersen, 1997. A comparative study on feature selection in text categorization. Dans les actes de *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, 412–420.

(Yazgan et Saraclar, 2004) A. Yazgan et M. Saraclar, 2004. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. Dans les actes de *Proceedings of the 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Volume 1, 745–748.

(Young et al., 1994) S. Young, J. Odell, et P. Woodland, 1994. Tree-based state tying for high accuracy acoustic modelling. Dans les actes de *Proceedings of the workshop on Human Language Technology*, 307–312. Association for Computational Linguistics.

(Yu et al., 2000b) H. Yu, T. Tomokiyo, Z. Wang, et A. Waibel, 2000b. New developments in automatic meeting transcription. Dans les actes de *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, 310–313.

(Zadeh, 1978) L. Zadeh, 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems 1*(1), 3–28.

(Zahariev, 2004) M. Zahariev, 2004. *A (acronyms)*. Thèse de Doctorat, Simon Fraser University.

(Zhan et Waibel, 1997) P. Zhan et A. Waibel, 1997. Vocal tract length normalization for large vocabulary continuous speech recognition. Rapport technique, Carnegie Mellon University.

(Zhu et Rosenfeld, 2001) X. Zhu et R. Rosenfeld, 2001. Improving trigram language modeling with the world wide web. Dans les actes de *Proc. ICASSP*, Volume 1, 533–536.

(Zipf, 1949) G. Zipf, 1949. *Human behavior and the principle of least effort.* Addison-Wesley Press.

Bibliographie personnelle

(Lecouteux et al., 2011) B. Lecouteux, G. Linarès, et S. Oger, 2011. Integrating imperfect transcripts into speech recognition systems for building high-quality corpora. *Computer Speech and Language 1*, à paraître.

(Oger et al., 2008a) S. Oger, G. Linarès, et F. Béchet, 2008a. Local methods for on-demand out-of-vocabulary word retrieval. Dans les actes de *the International Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, 118–122.

(Oger et al., 2008b) S. Oger, G. Linarès, F. Béchet, et P. Nocera, 2008b. Enrichissement dynamique du vocabulaire à partir du web. Dans les actes de *Journées d'Étude sur la Parole (JEP)*, 341–344.

(Oger et al., 2008c) S. Oger, G. Linarès, F. Béchet, et P. Nocera, 2008c. On-demand new word learning using the world wide web. Dans les actes de *the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4305–4308.

(Oger et al., 2009a) S. Oger, V. Popescu, et G. Linarès, 2009a. Probabilistic and possibilistic language models based on the world wide web. Dans les actes de *INTERSPEECH*, 2699–2702.

(Oger et al., 2009b) S. Oger, V. Popescu, et G. Linarès, 2009b. Using the world wide web for learning new words in continuous speech recognition tasks : Two case studies. Dans les actes de *Speech and Computer Conference (SPECOM)*, 76–81.

(Oger et al., 2010a) S. Oger, V. Popescu, et G. Linarès, 2010a. Combination of probabilistic and possibilistic language models. Dans les actes de *INTERSPEECH*, 1808–1811.

(Oger et al., 2010b) S. Oger, V. Popescu, et G. Linarès, 2010b. Modèles de langage probabilistes et possibilistes basés sur le web. Dans les actes de *Journées d'Etude sur la Parole (JEP)*, 261–264.

(Oger et al., 2010e) S. Oger, M. Rouvier, N. Camelin, R. Kessler, F. Lefèvre, et J.-M. Torres-Moreno, 2010e. Système du lia pour la campagne deft'10 : datation et localisation d'articles de presse francophones. Dans les actes de *DEFT*, 69–83.

(Oger et al., 2010c) S. Oger, M. Rouvier, et G. Linarès, 2010c. Classification du genre vidéo reposant sur des transcriptions automatiques. Dans les actes de *Traitement Automatique des Langues Naturelles (TALN)*, 341–344.

(Oger et al., 2010d) S. Oger, M. Rouvier, et G. Linarès, 2010d. Transcription-based video genre classification. Dans les actes de *the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5114–5117.

(Senay et al., 2010a) G. Senay, G. Linarès, B. Lecouteux, S. Oger, et T. Michel, 2010a. Décodage intéractif de la parole. Dans les actes de *Journées d'Etude sur la Parole (JEP)*, 253–256.
english

(Senay et al., 2010b) G. Senay, G. Linarès, B. Lecouteux, S. Oger, et T. Michel, 2010b. Transcriber driving strategies for transcription aid system. Dans les actes de *the International Language Resources and Evaluation Conference (LREC)*, Valletta, Malta. European Language Resources Association (ELRA).

(Senay et al., 2011) G. Senay, S. Oger, R. Rubino, G. Linarès, et T. Parent, 2011. Audio indexing on a medical video database : the avison project. Dans les actes de *International Conference on BioMedical Engineering and Informatics (BMEI)*.