Traitement Automatique du Langage : Études et apports aux frontières de l'interdisciplinarité

Richard Dufour

HABILITATION À DIRIGER DES RECHERCHES

Avignon Université

ÉCOLE DOCTORALE 536 «AGROSCIENCES & SCIENCES »
Laboratoire Informatique d'Avignon (LIA)
Spécialité : *Informatique*

Traitement Automatique du Langage :

Études et apports aux frontières de l'interdisciplinarité

Par
Richard DUFOUR

HDR présentée et soutenue à Avignon, le 8 décembre 2020
Unité de recherche : EA 4128

Composition du Jury :

Rapporteurs :

Philippe Langlais	Professeur	RALI, Université de Montréal, Montréal
Sophie Rosset	Directrice de recherche CNRS	LIMSI, Université Paris-Saclay, Orsay
Pascale Sébillot	Professeure	IRISA, Université de Rennes 1, Rennes

Examinateurs :

Jean-François Bonastre	Professeur	LIA, Avignon Université, Avignon
Yannick Estève	Professeur	LIA, Avignon Université, Avignon
Emmanuel Ethis	Professeur	INSEAC, CNAM, Guingamp
Georges Linarès	Professeur	LIA, Avignon Université, Avignon
Emmanuel Morin	Professeur	LS2N, Université de Nantes, Nantes

People talking without speaking
People hearing without listening
People writing songs that voices never shared
No one dared
Disturb the sound of silence

The Sound of Silence
Simon and Garfunkel

Remerciements

Je voudrais tout d'abord remercier l'ensemble des membres du jury de m'avoir fait l'honneur de relire, d'écouter et d'étudier les travaux scientifiques mis en perspective dans ce manuscrit. En premier lieu, je suis extrêmement reconnaissant aux rapporteurs, Philippe Langlais, professeur au RALI de l'Université de Montréal, Sophie Rosset, directrice de recherche CNRS au LIMSI de l'Université Paris-Saclay, et Pascale Sébillot, professeure à l'IRISA de l'Université de Rennes 1, d'avoir accepté cette charge supplémentaire de travail. Je remercie également Emmanuel Morin, professeur au LS2N de l'Université de Nantes, pour m'avoir fourni les premières remarques sur mon manuscrit. Je suis également très honoré d'avoir pu compter dans ce jury Emmanuel Ethis, professeur au CNAM InsEAC de Guingamp, avec qui j'ai eu la chance de collaborer ces dernières années dans plusieurs projets pluridisciplinaires. Il me semblait indispensable de pouvoir compter sur la présence de Georges Linarès, professeur au LIA d'Avignon Université, qui m'a intégré dans l'ensemble de ses projets de recherche et ce, dès mon arrivée au sein du LIA. Je mesure cette chance, qui est loin d'être la norme lors de l'arrivée dans un nouveau laboratoire de recherche. Dans cette continuité, Jean-François Bonastre, professeur au LIA d'Avignon Université, a également choisi de me faire confiance ces dernières années et je l'en remercie vivement. Enfin, il me semblait naturel d'avoir à mes côtés Yannick Estève, professeur au LIA d'Avignon Université, pour passer cette nouvelle étape, lui qui a été à l'origine des travaux de recherche que j'ai pu mener.

Je remercie également toutes les personnes ayant contribué activement à la réalisation de ce manuscrit. J'espère avoir correctement mis en valeur les différents travaux scientifiques réalisés et n'avoir omis personne lors de leur présentation. Je tiens à exprimer tout particulièrement mes remerciements à Mohamed Morchid sans qui ce document n'aurait pu être possible. Je n'oublie pas Damien Malinas, Raphaël Roth et Alexandre Delorme, qui ont largement contribué à cette coloration interdisciplinaire. Je remercie chaleureusement Vincent Labatut et Mickaël Rouvier, qui ont aussi une part importante dans l'élaboration de ce manuscrit. Je conclue ces remerciements en ayant une pensée pour tous les membres du LIA, et plus généralement du Centre d'Enseignement et de Recherche en Informatique (CERI) d'Avignon Université, que j'ai pu côtoyer depuis mon arrivée et qui m'ont tant apporté au quotidien.

Et enfin un très grand merci à Claire pour sa relecture attentive de "*dernière minute*", nécessitant une grande attention et minutie.

Résumé

Le traitement automatique du langage naturel (TALN) est un vaste domaine de recherche intégrant de nombreuses thématiques scientifiques (reconnaissance automatique de la parole, indexation automatique de documents, traduction automatique, synthèse vocale...). Grâce aux avancées obtenues au fil des dernières décennies dans chacune de ces thématiques, de nombreux outils et services ont pris vie hors des laboratoires de recherche pour atteindre le tissu socio-économique, et plus généralement le grand public, avec en fer de lance les assistants vocaux intelligents. Malgré ces grandes avancées, le TALN reste un domaine de recherche très actif, avec de nombreuses problématiques scientifiques ouvertes, considérant le fait que le langage, tout comme les usages de la société, ne sont pas figés mais en constante évolution.

Ce manuscrit propose un panorama des différents travaux de recherche auxquels j'ai pu participer ces dernières années, principalement en tant que maître de conférences au LIA, mettant alors en perspective l'évolution de mes travaux, qui m'ont conduit à travailler en collaboration avec d'autres disciplines scientifiques pour l'avancée du domaine du TALN. Ces derniers travaux sont notamment au centre de mes perspectives de recherche.

La première partie du manuscrit est consacrée à une des problématiques historiques, à savoir la représentation du contenu écrit et parlé. Nous présentons nos travaux sur la robustesse des représentations face au langage, en prenant en considération le contexte des documents, ainsi que l'utilisation et la proposition de représentations latentes pour la classification de documents.

Nous voyons ensuite, dans la deuxième partie, certains des travaux que nous avons menés sur la performance et l'évaluation en traitement du langage, allant de l'analyse et la caractérisation des erreurs de reconnaissance automatique de la parole, à leur correction.

La troisième partie montre l'évolution de mes activités de recherche, qui se sont alors orientées vers des problématiques interdisciplinaires pour le traitement du langage, avec nos travaux sur l'exploration des réseaux sociaux pour l'analyse d'événements, la détection de messages abusifs, et enfin le doublage vocal et la recommandation de voix.

La dernière partie résume mes activités de recherche, précisant les projets dans lesquels je me suis intégré, les différents encadrements d'étudiants auxquels j'ai pu participer, et mon implication dans le rayonnement et la vulgarisation scientifiques. Enfin, je termine ce manuscrit par une synthèse de ces différents travaux, me permettant de dresser une esquisse des activités de recherche futures dans lesquelles je souhaite m'impliquer à court, mais également à long terme, s'orientant assez largement vers des travaux interdisciplinaires.

TABLE DES MATIÈRES

Liste des acronymes

ACP	Analyse en Composantes Principales (Principal Component Analysis)
AFCP	Association Francophone de la Communication Parlée
ANR	Agence Nationale de la Recherche
BLSTM	Bi-directional LSTM (LSTM bi-directionnels)
BoW	Bag-of-Words (Sac-de-mots)
CBOW	Contextual Bag-Of-Words (Sac-de-mots contextuels)
CERI	Centre d'Enseignement et de Recherche en Informatique
CLEF	Conference and Labs of the Evaluation Forum
CNN	Convolutional Neural Network (Réseau de Neurones Convolutifs)
CRF	Conditional Random Fields (Champs Conditionnels Aléatoires)
DNN	Deep Neural Network (Réseau de Neurones Profond)
GMM	Gaussian Mixture Model (Modèle de Mélange Gaussien)
INA	Institut National de l'Audiovisuel
JCJC	Jeunes Chercheuses - Jeunes Chercheurs
LDA	Latent Dirichlet Allocation (Allocation Latente de Dirichlet)
LIA	Laboratoire Informatique d'Avignon
LIUM	Laboratoire d'Informatique de l'Université du Maine
LPC	Laboratoire de Psychologie Cognitive
LSTM	Long Short-Term Memory (Mémoire à long-court terme)
MFCC	Mel-Frequency Cepstral Coefficient
NCE	Normalized Cross Entropy (Entropie Croisée Normalisée)
PLDA	Probabilistic Linear Discriminant Analysis (Analyse discriminante linéaire probabiliste)
PLSTM	Parallel LSTM (LSTM parallèles)

RAP	Reconnaissance Automatique de la Parole
RI	Recherche d'Information
RNN	Recurrent Neural Network (Réseau de neurones récurrents)
RSN	Réseaux Sociaux Numériques
SER	Slot Error Rate
TAL	Traitement Automatique des Langues
TALN	Traitement Automatique du Langage Naturel
TAP	Traitement Automatique de la Parole
TF-IDF	Term Frequency - Inverse Document Frequency (Fréquence du terme-Fréquence inverse du docum
UBM	Universal Background Model (Modèle du monde)
VF	Version Française
VO	Version Originale
WAC	Wikipedia Abusive Conversations
WER	Word Error Rate (Taux d'Erreur-Mot)

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Sommaire

Depuis maintenant plusieurs années, les technologies liées au traitement automatique du langage (TAL) et de la parole (TAP) se sont fait une place dans notre vie quotidienne. En fer de lance, nous retrouvons les assistants vocaux intelligents, mis sur le devant de la scène par les géants du web, créant alors de nouveaux usages, et surtout, de nouvelles attentes de la part des utilisateurs. Le grand public est alors passé d'outils fantasmés, prenant leur origine dans les oeuvres cinématographiques (e.g. *HAL* dans *2001, l'Odyssée de l'espace*) ou encore télévisées (e.g. *Kit* dans *K2000*), à une réalité qui doit néanmoins encore faire face à de nombreux verrous scientifiques et technologiques, mais également à la méfiance et défiance, légitimes, de la société.

Bien entendu, les assistants vocaux ne sont que la vitrine technologique grand public, ceux-ci intégrant différentes briques et avancées issues du TAL. Ce sont donc globalement tous les domaines en TAL (reconnaissance automatique de la parole, traduction automatique, vérification du locuteur, indexation automatique de documents...) qui ont suffisamment progressé, en particulier depuis l'avénement récent des approches par apprentissage profond. La très grande masse des données disponibles, avec l'explosion des documents mis en ligne sur Internet, a également eu une influence majeure sur ces différents domaines de recherche. Internet a également fait émerger de nouveaux modes de communication entre les individus, comme les réseaux sociaux numériques (RSN), qui sont autant de nouvelles problématiques et tâches à explorer.

Travaux de recherche

Ce manuscrit résume les travaux que j'ai pu mener ces dix dernières années en TAL et TAP. Il suit, finalement, une partie des grands questionnements scientifiques présents quasiment depuis l'origine de ces domaines, tout en s'intéressant aux problématiques actuelles dues à l'évolution du langage et de ses usages. Mes travaux de recherche ont débuté pendant ma thèse au Laboratoire d'Informatique de l'Université du Maine (LIUM) dans le domaine de la reconnaissance automatique de la parole (RAP) dans le contexte de parole spontanée. Mes problématiques de recherche se sont ensuite élargies durant mon année de post-doctorat à Orange Labs, et surtout depuis mon arrivée au sein du Laboratoire Informatique d'Avignon (LIA) en tant que maître de

conférences.

Dès mon intégration dans la thématique Langage du LIA, j'ai eu l'opportunité de participer à plusieurs projets financés par l'Agence Nationale de la Recherche (ANR), notamment le projet SuMACC, porté au LIA par Georges Linarès. Dans le cadre de ce projet de recherche, j'ai tout d'abord travaillé sur la problématique de la robustesse des représentations de mots et de documents dans des transcriptions automatiques très bruitées. Les difficultés résidaient, ici, d'une part dans le fait que le langage utilisé était peu conventionnel (registre de langue familier, vocabulaire spécifique, forte spontanéité des échanges...), et d'autre part dans le fait que la transcription automatique de ces documents audio était très fortement erronée, *i.e.* avec un taux d'erreur-mot (WER) très élevé. Ces travaux ont principalement été développés pendant la thèse de Mohamed Morchid, que j'ai encadrée, et qui était dirigée par G. Linarès. Ces travaux sur la robustesse des représentations ont ensuite été continués pendant la thèse de Killian Janod, avec des approches fondées sur des réseaux de neurones, que j'ai co-encadrée avec M. Morchid, et qui était dirigée par G. Linarès.

Quasiment en parallèle de ces travaux sur la robustesse des représentations, j'ai pu participer au projet ANR ContNomina, toujours porté par G. Linarès au LIA, dont la problématique concernait la temporalité, et le caractère diachronique des mots, dans le cadre de la RAP. Des travaux assez proches ont pu être menés dans ce contexte avec la thèse CIFRE de Mohamed Bouaziz, avec l'entreprise EDD, que j'ai co-encadrée avec M. Morchid et G. Linarès, dont le sujet portait sur le traitement de flux de données parallèles diffusés en continu. De même, nous avons proposé, avec K. Janod, une approche permettant la prise en compte de la position dans l'historique des mots pour améliorer une représentation par plongement de mots (*word embeddings*).

Depuis 2015, et le début du projet ANR GaFes, mes travaux de recherche, alors fortement ancrés uniquement dans le domaine du TAL, ont évolué vers une démarche interdisciplinaire qui se poursuit encore actuellement. Ce projet est le fruit d'une collaboration entre informaticiens et sociologues. Dans le cadre de ce projet, nous avons pu travailler sur la mise en place d'un observatoire des festivals, développé pendant l'alternance de Mathias Quillot, permettant de trouver un terrain de dialogue commun entre chercheurs en TAL, chercheurs en sciences humaines, et grand public pour la restitution d'informations issues des RSN. Mes travaux se sont principalement axés sur l'étude des réseaux sociaux, comme par exemple les travaux menés en collaboration avec Mickaël Rouvier, Damien Malinas, Raphaël Roth et Alexandre Delorme, sur les opinions diffusées dans les RSN. Ces travaux interdisciplinaires se sont ensuite poursuivis dans le cadre du projet ANR The Voice, dont Jean-François Bonastre est le responsable scientifique au LIA. Avec la thèse d'Adrien Gresse, que j'ai co-encadrée avec Vincent Labatut et qui a été dirigée par J.-F. Bonastre, nous avons posé des premières bases sur la recommandation automatique de voix pour le doublage, dépassant le simple appariement acoustique puisque devant prendre en

compte des informations liées à des choix artistiques, culturels, et une réception et perception des voix par le public. M. Quillot, actuellement en thèse sur ce sujet, continue ces travaux initiés par A. Gresse en cherchant à mettre en avant, à identifier, et à mesurer l'information *personnage* contenue dans la voix jouée.

Je poursuis également mes travaux de recherche sur l'intégration d'informations issues de la structure des documents, ici en modélisant les interactions des utilisateurs au cours de conversations par messagerie instantanée, en combinaison du contenu textuel échangé. En travaillant sur une tâche commune, à savoir la détection de messages abusifs dans des conversations textuelles sur Internet, nous nous sommes aperçus, avec mes collègues V. Labatut et Etienne Papegniès, de l'importance de la modélisation des comportements des utilisateurs, qui, au contraire du langage, ne peuvent être intentionnellement masqués. Avec Noé Cécillon, qui a débuté sa thèse en septembre 2019, nous poursuivons ces efforts de travaux communs en TAL et réseaux complexes, la nature des informations extraites de ces deux domaines de recherche apparaissant clairement complémentaires.

Enfin, j'ai pu récemment travailler à nouveau sur les problématiques d'évaluation et d'adaptation des systèmes de RAP, qui faisaient alors partie des questionnements de recherche que j'avais pu entrevoir pendant ma thèse et mon année de post-doctorat. J'ai ainsi pu collaborer avec Salima Mdhaffar pendant sa thèse au LIUM, soutenue en 2020, sur la manière de rendre compte de performances des transcriptions automatiques et sur le problème de reproductibilité des résultats dans le cadre d'adaptation de modèles de langage à partir de données diachroniques collectées sur Internet.

Organisation du manuscrit

Ce manuscrit est organisé en cinq parties.

Les trois premières parties, présentant les travaux de recherche principaux dans lesquels j'ai pu m'investir, sont organisées selon les problématiques qui ont guidé mon travail scientifique ces dernières années. La partie I décrit les travaux menés sur la robustesse des représentations du contenu. Dans cette partie, le chapitre 1 se focalise sur l'intégration d'informations issues du contexte pour améliorer la représentation de documents écrits, considérés ici comme bien formés du point de vue des règles du langage. Puis dans le chapitre 2, nous nous intéressons à la représentation de transcriptions automatiques très bruitées, en proposant des approches de plus haut-niveau et une approche multi-vues permettant une réponse au problème des hyper-paramètres nécessaires pour l'entraînement de modèles par espaces de thèmes.

Ensuite, dans la partie II, nous présentons les travaux liés à la performance et l'évaluation en traitement du langage. Nous verrons tout d'abord, dans le chapitre 3, une étude mettant en parallèle la métrique du taux d'erreur-mot en RAP et les performances de classification de

documents transcrits automatiquement. Puis, dans le chapitre 4, nous exposons les travaux menés sur la détection et caractérisation des erreurs de transcription, avant de finir, dans le chapitre 5, sur la correction de ces erreurs et l'évaluation des systèmes de RAP.

La partie III se focalise sur les travaux mêlant TAL et autres domaines de recherche que j'ai entrepris ces dernières années. Le chapitre 6 est l'occasion de présenter plusieurs travaux autour de l'analyse d'événements dans les réseaux sociaux. Ensuite, le chapitre 7 détaille nos travaux conjoints en TAL et réseaux complexes pour la détection de messages abusifs. Enfin, le chapitre 8 termine cette partie en présentant les premiers travaux menés sur la recommandation de voix pour le doublage et la représentation de la voix jouée.

La partie IV est consacrée à la description de mon investissement au niveau de la recherche et de l'encadrement scientifique. Dans le chapitre 9, je donne quelques détails sur les projets de recherche dans lesquels je me suis investi. Puis le chapitre 10 fournit une vue synthétique des étudiants que j'ai pu co-encadrer en thèse et en stage de Master. Enfin, je termine par mon implication dans le rayonnement et la vulgarisation scientifiques dans le chapitre 11.

Dans la dernière partie de ce manuscrit (partie V), je propose un bilan de ces différents travaux présentés, et dont découlent, naturellement, les perspectives de recherche que j'ambitionne de mener dans les prochaines années, donnant une large part à l'évaluation des systèmes automatiques, leur étude, et leur compréhension dans une vision interdisciplinaire appliquée au domaine du TAL.

Première partie

Robustesse de la représentation du contenu écrit et parlé

Chapitre 1

Apports du contexte pour la représentation du contenu

Sommaire

1.1 Introduction

Les approches par sac-de-mots (*Bag-of-Words*, BoW) [Harris, 1954] ont longtemps été les représentations de mots à l'état de l'art dans les domaines du traitement automatique du langage naturel (TALN) et de la recherche d'information (RI). Une des approches les plus simples est de considérer les BoW comme une représentation des mots contenus dans une phrase ou un document sous la forme d'un dictionnaire, où, à chaque entrée (ici, chaque mot), est associé son nombre d'occurrences dans le texte. Avec ce type d'approche, un document peut alors être représenté par un vecteur dont la taille correspond au nombre de mots uniques dans ce document (*i.e.* le dictionnaire), les valeurs contenues dans le vecteur étant ici le nombre d'occurrences de ces mots. Ces représentations souffrent cependant de plusieurs limites, liées par exemple :

— *Au contexte.* Chaque mot est ici considéré indépendamment des autres, *i.e.* le contexte dans lequel le mot apparaît est ignoré. En conclusion, les liens entre les mots et leur position dans la séquence ne sont alors pas pris en compte.
— *À la taille du dictionnaire.* Ce type d'approche conduit à avoir des vecteurs de représentation, pour les documents, potentiellement de grandes dimensions dans le cas où le dictionnaire (vocabulaire) est très grand. De même, ces représentations apparaissent aussi potentiellement très creuses pour un document identifié, celui-ci ne couvrant généralement qu'une partie réduite du dictionnaire.

Pour traiter le problème de contexte et de séquentialité des données, une des méthodes les plus connues, et, sans être la seule, sûrement une des plus largement utilisées, est l'approche *n-gramme*. Le modèle n-gramme est un modèle probabiliste où un mot est pris en compte avec son historique représenté par les $n-1$ mots qui le précèdent. Par exemple, si l'on parle de modèle tri-gramme ($n = 3$), cela veut dire que le modèle prend en compte les deux mots précédant le mot courant. Au niveau de la modélisation du langage, et ce malgré sa simplicité, l'approche n-gramme est encore aujourd'hui, par exemple, largement utilisée en reconnaissance automatique de la parole (RAP).

Pendant la thèse CIFRE de Mohamed Bouaziz, réalisée avec l'entreprise EDD, nous avons entrepris de travailler au niveau du contexte et de la séquentialité des données dans une problématique faisant intervenir des flux continus parallèles. De manière plus précise, nous voulions tirer profit, lorsque nous traitions un flux particulier (ici, une chaîne de télévision), des autres flux diffusés en parallèle (ici, les autres chaînes de télévision). Nous avons alors proposé une architecture s'appuyant sur les réseaux de neurones de type Long short-term memory (LSTM) permettant de prendre en compte des flux parallèles (Parallel LSTM ou PLSTM), améliorant l'approche classique par n-gramme [Bouaziz et al., 2016d]. La proposition ainsi que les résultats que nous avons obtenus sont alors présentés dans la partie 1.2.

Depuis plusieurs années maintenant, de nouvelles approches d'apprentissage automatique permettent de représenter des mots par des vecteurs de nombres réels tout en résolvant, entre autres, certains problèmes évoqués par la taille du dictionnaire. Ici, l'idée est de projeter les mots dans un espace multidimensionnel afin d'obtenir des vecteurs de taille restreinte tout en contenant un maximum d'informations sur le mot dans son contexte d'apparition. Dans la littérature, ce type de représentation est appelé *plongement de mots* ou *plongement lexical* (*word embedding*). Une des caractéristiques de ces plongements lexicaux est que des mots partageant des contextes similaires devraient avoir des vecteurs de valeurs proches. Une des approches les plus connues pour apprendre ces représentations de mots de manière non supervisée, au moyen de réseaux de neurones artificiels, est l'approche Word2vec [Mikolov et al., 2013a]. D'autres méthodes d'apprentissage des plongements de mots existent, parmi lesquelles nous pouvons aussi citer Glove [Pennington et al., 2014], approche également très populaire, permettant dans de

nombreuses applications d'atteindre des performances relativement similaires, bien que souvent légèrement inférieures à Word2vec [Seok et al., 2015, Ghannay et al., 2016, Naili et al., 2017]. Glove est ici une méthode d'apprentissage non supervisé consistant à factoriser une matrice de cooccurence des mots. Notons que toutes ces méthodes ont besoin de très grands corpus de données, non annotées, pour produire des représentations de bonne qualité.

Dans le cadre du projet ANR ContNomina ainsi que de la thèse CIFRE de Killian Janod, en collaboration avec l'entreprise Orkis, une partie de ses travaux s'est intéressée aux méthodes permettant d'obtenir ces plongements de mots, notamment avec l'approche Word2vec, en identifiant une des limites dans l'apprentissage des modèles. En effet, bien que le contexte soit pris en compte, la position des mots à l'intérieur de ce contexte semble en partie ignorée et non directement prise en compte. Nous avons alors proposé une approche par pondération permettant d'intégrer cette information liée à la distance des mots dans le contexte par rapport au mot ciblé [Janod et al., 2016b]. Nous présentons alors ces travaux dans la partie 1.3.

1.2 Intégration d'informations de flux parallèles

1.2.1 Contexte et problématique

Comme nous l'avons vu dans l'introduction, la thèse de Mohamed Bouaziz a porté sur le traitement de flux de chaînes télévisées collectés en continu et en temps réel. D'un point de vue applicatif, il y a un fort intérêt à pouvoir traiter et structurer automatiquement ces types de données pour fournir des informations sur les contenus diffusés, afin d'améliorer leur accessibilité. De nombreux travaux se sont intéressés à structurer les émissions télévisées, comme le prouve, par exemple, les travaux en segmentation thématique [Hearst, 1997, Guinaudeau et al., 2010, Bouchekif et al., 2014]. Ces contenus audiovisuels sont alors vus comme des séquences d'événements chronologiques organisés, dans le sens où les événements, peu importe leur granularité (scènes, thèmes, émissions...), n'apparaissent pas à la suite par hasard, mais font l'objet d'un choix réfléchi. Par exemple, l'enchaînement des sujets dans une émission fait partie d'un choix des producteurs de l'émission, en mettant tel sujet avant tel autre. De même, la chaîne télévisée a un intérêt à diffuser certaines émissions à des moments spécifiques dans la journée afin de maximiser son audimat [Poli, 2007]. L'originalité du contexte des travaux que nous avons menés avec M. Bouaziz réside dans le fait de ne pas considérer un flux de manière isolée, mais de prendre en considération l'idée que la structuration d'une chaîne de télévision peut aussi dépendre des autres émissions (et donc des autres structurations) de chaînes de télévision diffusant des programmes en parallèle.

Le traitement de séquences a été relativement bien étudié dans la littérature (approches n-gramme, Conditional Random Fields (CRF), modèles de markov cachés (HMM), réseaux de neurones...). Ainsi, récemment mis en avant, les réseaux de neurones récurrents LSTM font par-

tie des architectures appliquées avec succès dans de nombreuses applications devant traiter des séquences longues. Une des propositions de la thèse de M. Bouaziz a alors consisté à étendre le paradigme des réseaux LSTM pour traiter des flux parallèles au lieu de flux uniques. Nous présentons l'approche dans la partie 1.2.4. À titre comparaison, nous détaillons juste avant l'approche classique n-gramme et son extension au traitement de flux parallèles, qui ont constitué nos *baselines*, dans la partie 1.2.3. Nous avons enfin dû mettre en place un protocole expérimental et proposer un corpus annoté pour réaliser nos différentes expérimentations, ce que nous décrivons dans la partie suivante.

1.2.2 Annotation d'un corpus de flux télévisés parallèles

Il n'existait pas, à notre connaissance, de corpus d'émissions télévisées multi-chaînes en flux continu sur des mêmes périodes temporelles nous permettant de mener à bien nos expériences. Un corpus audio de flux télévisés a alors été collecté et annoté sur 2 jours à partir de 4 chaînes télévisées – 3 généralistes (TF1, M6 et France 5) et 1 spécialisée (TV5 Monde) – entre le 10 et le 12 février 2014 [Bouaziz et al., 2016c]. L'entreprise EDD, partenaire industriel de la thèse CIFRE de M. Bouaziz, a pu fournir les flux audio, l'entreprise ayant mis en place un système permettant la collecte d'un très grand nombre de flux audio et vidéo en continu et en temps réel. Notre travail se concentrant sur la problématique de la structuration de contenus, nous avons proposé de travailler sur la tâche de prédiction du genre d'une émission en s'appuyant sur l'historique des émissions précédemment déroulées. Cette proposition suit des travaux précédents [Poli, 2008], où les auteurs ont proposé une approche par HMM pour construire des grilles de programmes selon les historiques des chaînes de télévision.

Proposition d'une catégorisation

Un des problèmes qui s'est posé pour l'annotation en genre des émissions télévisées est le choix de la liste des catégories. En effet, nous nous sommes rapidement aperçus que cette tâche n'était pas aussi évidente, chaque chaîne de télévision proposant ses propres métadonnées pour décrire ses programmes. Nous avons donc plutôt choisi de nous appuyer sur des taxonomies existantes proposées par des organismes extérieurs (et donc supposés indépendants) des chaînes de télévision. Nous nous sommes alors intéressés à la catégorisation des genres d'émissions proposée par l'Institut National de l'Audiovisuel (INA). L'INA proposait, au moment de la création du corpus, de classer les émissions en 52 genres différents [Troncy, 2001].

Néanmoins, la liste de l'INA apparaissait trop exhaustive, avec des genres finalement trop proches, rendant la tâche d'annotation trop difficile, avec des risques de confusion importants et potentiellement un grand nombre de catégories finalement sous-représentées car trop précises (rappelons que nous avions une collecte sur 2 jours seulement). Au final, nous avons proposé de nous restreindre à 14 genres d'émissions [Bouaziz et al., 2016c] : *Inter-programmes, Ac-*

tualité, Météo, Dessins animés, Fiction, Documentaire, Téléachat, Plateau/Débat, Magazine de reportage, Autres magazines, Musique, Télé-réalité, Programme court et *Jeu*. Notons que *Inter-programme* est un genre particulier qui n'a d'intérêt que pour représenter des coupures entre et pendant les programmes (publicités, jingles, génériques de début et fin de programme...). Ce genre ne sera pas utilisé dans les expériences que nous présentons dans la suite.

Annotation des données

Les différents flux audio ont alors été segmentés manuellement en genre d'émissions selon la catégorisation proposée précédemment. Les instants précis de début et de fin de programme ont été fournis. Afin de garder une consistance globale dans l'annotation, un seul annotateur a réalisé cette segmentation et catégorisation manuelles (ici, M. Bouaziz). Un exemple d'annotation des 4 flux télévisés est visible dans la figure 1.1.

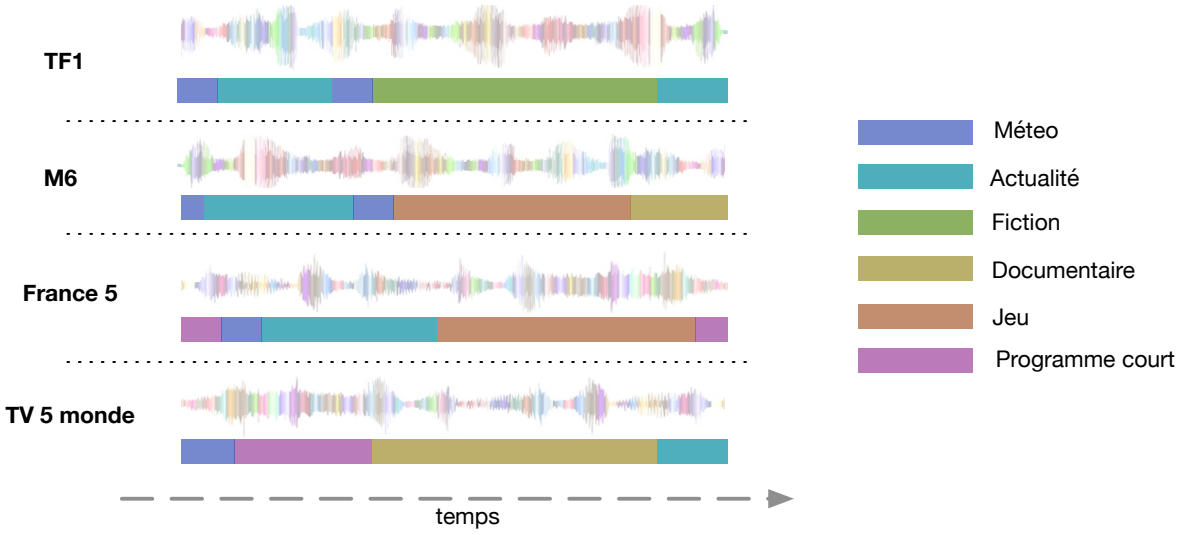

FIGURE 1.1 – Exemple de segmentation des 4 flux télévisés en émissions et catégorisation en genre.

Description du corpus

Afin d'avoir une idée globale de la répartition des catégories pour chaque chaîne télévisée, le tableau 1.1 résume le nombre de segments de parole associé à chaque genre télévisé pour les 4 chaînes ciblées. Nous avons ajouté, entre parenthèses, le pourcentage de chaque genre de programme pour chaque chaîne.

Globalement, les chaînes semblent proposer proportionnellement les mêmes programmes. Nous observons également que certains genres sont sur-représentés, notamment les inter-programmes.

Genre	France 5	TV5 Monde	TF1	M6	Total
Actualité	0	38 (14 %)	4 (2 %)	4 (2 %)	46 (5 %)
Dessins animés	26 (13 %)	0	17 (7 %)	12 (6 %)	55 (6 %)
Plateau/Débat	23 (12 %)	10 (4 %)	4 (2 %)	0	37 (4 %)
Documentaire	35 (18 %)	25 (9 %)	2 (1 %)	0	62 (7 %)
Fiction	1 (1 %)	6 (2 %)	45 (18 %)	39 (19 %)	91 (10 %)
Inter-programme	95 (49 %)	134 (49 %)	124 (50 %)	101 (50 %)	454 (50 %)
Jeu	0	6 (2 %)	12 (5 %)	0	18 (2 %)
Autres magazines	4 (2 %)	5 (2 %)	0	5 (2 %)	14 (2 %)
Magazine de rep.	6 (3 %)	12 (4 %)	7 (3 %)	5 (2 %)	30 (3 %)
Météo	0	18 (7 %)	8 (3 %)	15 (7 %)	41 (4 %)
Musique	0	3 (1 %)	9 (4 %)	4 (2 %)	16 (2 %)
Programme court	3 (2 %)	15 (6 %)	6 (2 %)	0	24 (3 %)
Téléachat	0	0	1	2 (1 %)	3 (0 %)
Télé-réalité	0	0	10 (4 %)	16 (8 %)	26 (3 %)
Total	193	272	249	203	917

Tableau 1.1 – Nombre de segments associés à chaque genre pour chaque chaîne télévisée. Nous avons également fourni la proportion de chaque genre par rapport à la chaîne ciblée lorsque cela était pertinent.

Cela reste cependant cohérent avec cette catégorie (coupure entre les émissions). Les chaînes TF1 et M6 ont clairement des genres de programmes quasiment identiques, alors que France 5 se distingue au niveau des documentaires, des plateaux/débats et des dessins animés. TV5 Monde, de son côté, se distingue par sa proportion élevée d'actualités. Enfin, certaines catégories sont sous-représentées dans toutes les chaînes, comme la télé-réalité, le téléachat ou les autres magazines.

1.2.3 Modèle n-gramme

Afin de vérifier l'intérêt de l'approche PLSTM proposée dans la partie 1.2.4, nous avons choisi d'utiliser l'approche n-gramme, et par extension l'approche 4n-grammes (correspondant à la prise en compte des 4 flux des chaînes de télévision) en tant que *baseline* et dont nous donnons quelques détails ici.

Modèle classique

Pour ce genre de modèle, l'historique d'un élément w est représenté par les $n-1$ éléments qui le précèdent. Dans la pratique, la valeur de n généralement choisie est de 3 ou 4. On parlera alors respectivement de modèles *tri-grammes* et *quadri-grammes*. Etant donné une séquence de

k éléments, les modèles n-grammes attribuent alors une probabilité pour la séquence selon :

$$P(W_1^k) = P(w_1) \prod_{i=2}^{n-1} P(w_i|w_1, \cdots, w_{i-1}) \prod_{i=n}^{k} P(w_i|w_{i-n+1}, \cdots, w_{i-1}) \tag{1.1}$$

Même si ce genre de modèle semble particulièrement réducteur, en ne prenant en compte que des contraintes séquentielles courtes, il contient finalement suffisamment d'informations pour guider efficacement un système de TAL, pour preuve son utilisation encore actuelle en RAP. Il existe plusieurs méthodes pour procéder à l'estimation des paramètres du modèle de langage [Federico and De Mori, 1998]. La plus commune est l'estimation par *maximum de vraisemblance*, dont le nom indique que la distribution des probabilités du modèle de langage obtenue est celle qui maximise la vraisemblance du corpus d'apprentissage. Enfin, des techniques de lissage permettent de compenser l'absence de certain n-grammes dans le corpus d'apprentissage : elles peuvent être vues comme une sorte de généralisation permettant d'attribuer une probabilité non nulle à un événement (n-gramme) non vu. Les principales techniques de lissage sont décrites dans [Chen and Goodman, 1999], où est également présentée une discussion sur leurs performances respectives.

Modèle 4n-grammes

Les modèles n-grammes n'étant pas conçus pour traiter des séquences parallèles, nous avons alors dû faire un choix quant à la fusion des différents modèles n-grammes. Au final, après plusieurs expérimentations, nous avons choisi une approche qui nous semblait le mieux respecter l'idée initiale des modèles n-grammes, en fusionnant les séquences d'événements des 4 chaînes de télévision, afin de constituer un nouvel élément. Ce nouvel élément est alors utilisé pour apprendre un modèle n-gramme classique avec ces nouveaux historiques concaténés. Nous appellerons 4n-grammes cette approche dans nos expériences. Bien entendu, cette approche possède plusieurs problèmes, le principal étant le risque élevé d'historiques manquants dûs à la fusion de 4 éléments en un seul. Néanmoins, au vu du tableau 1.1, la sur-représentation de certains genres ici est un avantage, réduisant le nombre réellement présents dans les grilles de programme.

1.2.4 Réseaux de neurones LSTM parallèles

Réseaux LSTM classiques

Les réseaux de neurones Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] sont une variante des réseaux de neurones récurrents (RNN) [Rumelhart et al., 1986]. L'avantage de ces réseaux est de pouvoir prendre en compte le contexte dans le réseau. Alors que les RNN ont des difficultés à prendre en compte des historiques à long terme, les LSTM permettent de pallier le problème de disparition du gradient exposé dans [Hochreiter and Schmid-

huber, 1997], ne pouvant modifier les poids de l'événement courant en fonction d'événements passés lointains. Ainsi, chaque unité du réseau LSTM contient un état caché h, comme dans les RNN, mais également un état c (cellule mémoire) devant conserver les éléments du passé.

Une cellule dans un réseau LSTM se caractérise alors par un nœud central (mémoire interne), ainsi qu'un ensemble de portes permettant de réguler l'information de l'historique, au travers des portes d'entrée (extraction des informations courantes), d'oubli (conservation ou rejet de certaines informations), et de sortie (les informations conservées et transmises). De plus amples informations sur le fonctionnement des LSTM peuvent se trouver dans [Hochreiter and Schmidhuber, 1997].

LSTM bi-directionnels (BLSTM)

Tout comme les modèles n-grammes classiques, les RNN (et par extension les LSTM) ne prennent en compte que le contexte précédent (ici, en avant). Des approches ont cependant proposé d'utiliser le contexte global, *i.e* le contexte passé et futur par rapport à un instant donné, comme l'architecture RNN bi-directionnels (BRNN) proposée dans [Schuster and Paliwal, 1997]. Les BRNN intègrent alors ces deux directions au moyen de deux couches cachées séparées. Ce réseau contient toujours une couche de sortie qui prend donc des données issues non plus d'une seule couche cachée mais de deux couches cachées. Une description détaillée du fonctionnement de l'architecture BRNN peut se trouver dans [Schuster and Paliwal, 1997]. Finalement, pour obtenir des LSTM bi-directionnels (BLSTM), il suffit de remplacer les cellules du BRNN par les cellules de l'architecture LSTM.

LSTM parallèles (PLSTM)

Les travaux menés par M. Bouaziz durant sa thèse ont alors amené à proposer une architecture s'inspirant des BRNN, mais ici pour traiter des flux parallèles *dans un seul sens* (ici, l'historique), alors que les BRNN prenaient en compte le contexte passé et futur mais d'un seul flux de données séquentielles. La figure 1.2 présente l'architecture RNN parallèle, et par extension l'architecture LSTM parallèle en remplaçant les cellules RNN, comme nous l'avions vu pour les BRNN. Il est possible de retrouver les détails de l'architecture PLSTM dans l'article [Bouaziz et al., 2016d] ainsi que dans la thèse de M. Bouaziz [Bouaziz, 2017]. Au contraire des BLSTM, les PLSTM proposent de traiter n séquences en entrée (x^n), chacune de ces séquences ayant sa propre couche cachée h^n associée.

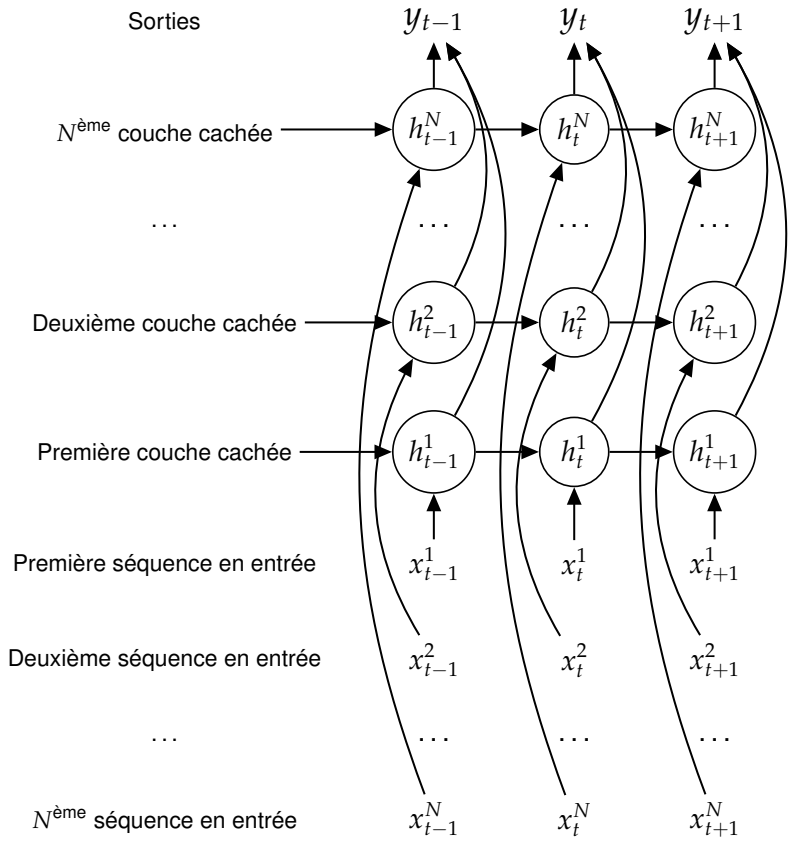

FIGURE 1.2 – Architecture LSTM parallèles (PLSTM) [Bouaziz, 2017].

1.2.5 Prédiction du genre d'émissions télévisées

Corpus de genres d'émissions

Comme expliqué précédemment, nous avons choisi d'évaluer l'approche PLSTM proposée sur la tâche de prédiction du genre d'émissions télévisées sachant l'historique des genres de la chaîne étudiée mais également l'historique de chaînes diffusant des émissions en parallèle. Nous nous sommes appuyés sur le corpus collecté et annoté défini dans la partie 1.2.2. Celui-ci étant finalement trop petit (et plutôt orienté pour des tâches de segmentation de l'audio), nous avons alors choisi de réaliser une collecte plus large pour nos expériences en extrayant les grilles des programmes de 3 années continues (janvier 2013 à décembre 2015) des 4 chaînes que nous avions déjà étudiées. Dans le cadre de ces expériences, nous nous sommes focalisés sur la chaîne M6 comme flux principal de test. Les 3 autres chaînes, à savoir TF1, France 5 et TV5 monde, viennent donc ici en tant que flux parallèles. À partir de notre analyse des genres de la partie 1.2.2, nous avons réduit la liste à 11 genres, les autres n'étant finalement pas suffisamment

présents. Les inter-programmes ne sont ici pas pertinents. Le genre *Autre* permet d'agréger les émissions que nous n'avons pas réussi à associer un des genres.

Les données de 2013 et 2014 ont été utilisées pour l'apprentissage (70 %) et le développement (30 %), l'année 2015 étant conservée pour le test. Le tableau 1.2 résume la répartition des genres d'émissions pour la chaîne M6 selon les corpus d'apprentissage, de développement et de test.

Genre	**Apprentissage**	**Développement**	**Test**
Météo	2 691	1 153	1 712
Fiction	1 890	810	1 478
Actualité	913	392	679
Autre magazine	981	421	466
Musique	461	197	340
Téléachat	421	180	308
Jeu	476	204	287
Dessins animés	361	155	205
Autre	277	119	133
Télé-réalité	83	36	76
Documentaire	29	13	14
Total	8 107	3 680	5 698

Tableau 1.2 – Distribution des genres d'émissions pour la chaîne M6 pour les corpus d'apprentissage, de développement et de test. Les genres sont classés dans le tableau par le nombre décroissant de leurs occurrences.

Systèmes évalués

Nous avons tout d'abord proposé d'évaluer les approches classiques n-gramme et LSTM au moyen d'un seul flux de données (M6 dans notre contexte). Nous avons ensuite comparé ces performances avec les systèmes que nous proposons, à savoir les PLSTM en prenant 1 flux parallèle (TF1) en plus du flux étudié (M6), que nous appelons P2LSTM, ainsi que des PLSTM avec tous les flux parallèles, *i.e.* toutes les chaînes collectées (P4LSTM). De même, à titre de comparaison, nous avons utilisé l'approche 4n-gramme présentée dans la partie 1.2.3.

Nous avons choisi de comparer l'utilisation de séquences d'historique de 1 à 4 émissions précédentes (par exemple pour les n-grammes, évaluation de bi-grammes à 5-grammes). Le LSTM classique ainsi que les P2LSTM et P4LSTM proposés sont composés de 3 couches, à savoir une couche d'entrée x de taille variant de 1 à 4 selon la taille de l'historique, une couche cachée h de taille 80, et une couche de sortie y de la taille égale au nombre de genres (11).

Expériences

Le tableau 1.3 présente les résultats obtenus, en termes de F-mesure, par les différentes approches proposées pour la prédiction des genres d'émissions de la chaîne M6 en faisant varier la taille de l'historique des séquences (de 1 à 4), en considérant soit un seul flux (n-gramme et LSTM), soit des multiflux (4n-gramme, P2LSTM, P4LSTM).

Taille historique	**n-gramme**	**4n-gramme**	**LSTM**	**P2LSTM**	**P4LSTM**
1	19,07	59,39	11,60	47,46	47,24
2	51,38	58,35	34,64	54,17	59,54
3	57,41	57,10	50,74	58,69	59,92
4	56,87	57,26	56,47	58,67	**60,81**

Tableau 1.3 – Performance (en F-mesure) des approches n-gramme, 4n-gramme, PLSTM, P2LSTM et P4LSTM pour la détection de genres d'émissions. La taille des séquences d'historique varie de 1 à 4.

Nous avons tout d'abord observé que sur les approches n-grammes, l'utilisation de plusieurs flux parallèles pour prendre une décision (4n-gramme) permet d'obtenir de meilleurs résultats qu'en ne considérant qu'un seul flux (n-gramme). Ceci a été également vérifié sur les approches à base de LSTM, où les LSTM classiques (un seul flux) obtiennent des résultats bien inférieurs, que ce soit en considérant 2 (P2LSTM) ou 4 (P4LSTM) flux en parallèle. Au final, l'approche P4LSTM a permis d'obtenir les meilleures performances, surtout lorsque l'historique considéré est grand (ici, 4), au contraire de l'approche 4n-gramme, qui semble mieux fonctionner avec un historique très court (1).

1.3 Information temporelle pour les plongements de mots

1.3.1 Contexte

Nous avons évoqué en introduction les approches de représentation par plongement de mots, permettant de représenter un mot par un vecteur de taille réduite dans un espace multidimensionnel. Dans le travail de thèse de Killian Janod, nous avons alors proposé de travailler sur une des méthodes les plus connues pour construire ces représentations, à savoir Word2vec [Mikolov et al., 2013a]. Dans le travail sur la prédiction de genres d'émissions télévisées présenté précédemment, nous avons vu que le contexte est une information importante, et qu'il est nécessaire de la prendre correctement en compte pour obtenir des systèmes robustes et performants.

En TAL, et dans beaucoup de problématiques traitant des séquences, l'ordre des éléments dans cette séquence n'est pas anodine, puisque cette position peut contenir une information. Nous sommes alors partis de l'hypothèse que, pour la modélisation d'un mot, son contexte proche devrait avoir une influence plus importante, puisqu'il a une *action* plus directe, que des mots au

voisinage beaucoup plus lointain. En effet, dans le cadre de séquences très longues, par exemple les mots dans des documents textuels, un contexte large donne une information générale sur le document qu'il ne faut pas négliger pour en extraire une vision globale, mais nous imaginons que son contexte direct caractérise plus fortement un mot. Ces travaux s'inspirent de ceux auxquels j'ai pu participer à mon arrivée au sein du LIA en tant que maître de conférences pour la reconnaissance de noms de personnes dans le cadre du projet ANR PERCOL (voir partie 9.1.2). L'idée était d'utiliser des modèles contextuels continus pour caractériser les noms de personnes, en mettant un poids sur les mots du contexte selon leur position [Bigot et al., 2013b,a].

L'approche Word2vec utilise, lors de son apprentissage, les mots au contexte du mot ciblé comme un sac-de-mots. Chaque mot du contexte est alors traité de manière égale, sa position relative étant alors ignorée (poids de 1 s'il est présent dans le contexte, 0 sinon). La méthode que nous avons proposée a permis d'améliorer la prise en compte du contexte en le pondérant selon sa position dans le voisinage d'un mot. La section suivante présente succinctement les principes de l'approche Word2vec, avec ses deux modèles de plongement de mots : Skip-gram et CBOW. Nous verrons ensuite, dans la partie 1.3.3, l'approche que nous avons proposée pour la pondération du contexte et son intégration dans Word2vec. Nous terminons par une des expériences réalisées pour valider l'approche dans la partie 1.3.4, ici le test de recherche de mots analogues (*Semantic-Syntactic Word Relationship test*) [Mikolov et al., 2013a].

1.3.2 Approche Word2vec

L'approche Word2vec propose un ensemble de modèles permettant de représenter des mots par des plongements lexicaux (*word embeddings*). Ces modèles s'appuient sur une architecture à base de réseaux de neurones artificiels permettant de capturer des relations sémantiques et syntaxiques complexes entre les mots. Ces réseaux sont alors entraînés à reconstruire le contexte linguistique des mots. Deux architectures ont été proposées pour apprendre des plongements de mots [Mikolov et al., 2013a], à savoir les modèles Skip-gram et *Continuous Bag Of Words* (CBOW) présentés dans la figure 1.3. L'apprentissage des modèles Word2vec est non supervisé, mais nécessite des grands corpus de textes en entrée pour apprendre des modèles robustes [Mikolov et al., 2013b]. Ces approches sont capables de capturer plusieurs degrés de similarités entre les mots. Des schémas sémantiques et syntaxiques peuvent alors être reproduits au moyen d'opérations algébriques, ce qui est au centre de la tâche de recherche de mots analogues que nous voyons dans la sous-partie 1.3.4.

L'efficacité des approches Word2vec en TAL a été démontrée dans de nombreuses tâches, comme par exemple en reconnaissance d'entités nommées [Seok et al., 2016], en analyse de sentiments [Rouvier and Favre, 2016] ou encore en classification de documents [Lilleberg et al., 2015]. Dans [Mikolov et al., 2013b], les auteurs ont cependant montré que le modèle CBOW est plus rapide à entraîner et capture plutôt des informations syntaxiques, alors que le modèle

Skip-gram semble donner de meilleurs résultats, capturer des informations liées à la sémantique, et être plus efficace sur les mots peu fréquents. Néanmoins, comme cela dépend généralement de la tâche, nous avons choisi d'évaluer notre proposition sur les deux architectures.

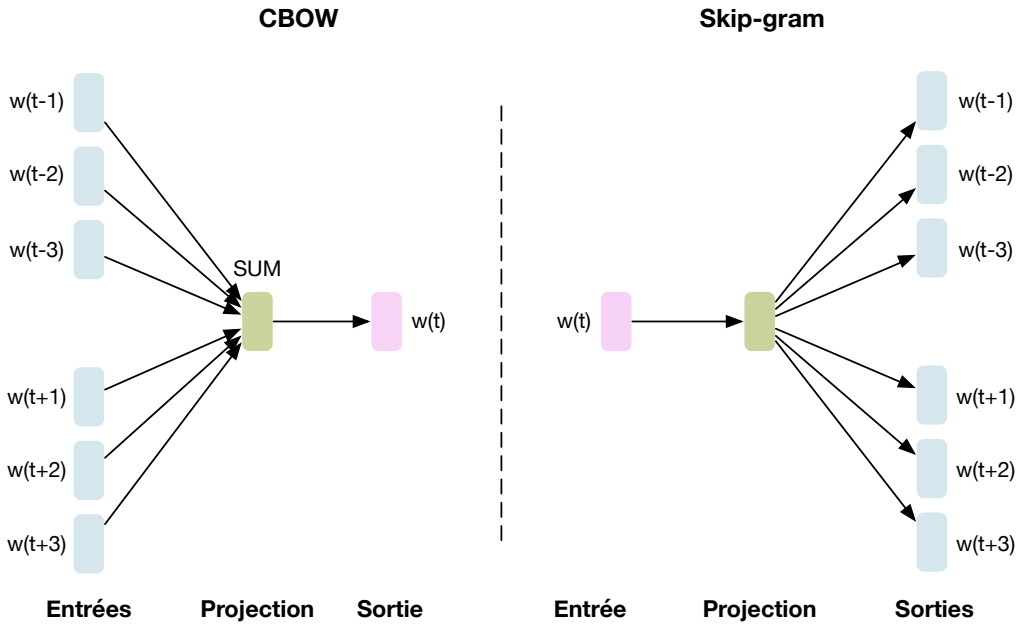

FIGURE 1.3 – Architecture des modèles CBOW et Skip-gram de l'approche Word2vec.

Les deux architectures sont composées de 3 couches (entrée, projection et sortie). Ces deux modèles utilisent un vecteur de caractéristiques binaires en *one-hot encoding* (1 si le mot est présent, 0 sinon) en entrée pour le CBOW et en sortie pour le Skip-gram. Le modèle CBOW cherche à prédire le mot courant ($w(t)$) à partir de la fenêtre du contexte dans lequel il apparaît (dans l'exemple de la figure 1.3, 3 mots avant et 3 mots après). Pour rappel, l'ordre des mots du contexte n'influence pas la prédiction, nous sommes dans un contexte de sac-de-mots ici. À l'inverse, le modèle Skip-gram utilise le mot courant ($w(t)$) pour prédire le contexte autour du mot.

1.3.3 Pondération des mots au voisinage

La méthode que nous avons proposée prend alors en compte la position des mots dans l'optique d'améliorer la représentation des plongements lexicaux au moyen d'une fonction de pondération log-linéaire du contexte δ [Bigot et al., 2013b] en lieu et en place des vecteurs de caractéristiques binaires en entrée (CBOW) ou en sortie (Skip-gram). Le contexte est pondéré

avec $\delta(w)$ pour chaque mot w :

$$\delta(w) = \frac{\alpha}{\gamma + \beta log(d(w))} \tag{1.2}$$

où d est la distance entre le mot au centre du contexte c et du mot w à pondérer. α, γ et β sont les paramètres de la fonction de distance. La fonction log-linéaire est bien adaptée à la pondération des mots car elle fournira des poids élevés aux mots proches du centre (donc dans un contexte proche) et des poids plus faibles aux mots plus éloignés dans le contexte. Ainsi, les mots les plus éloignés devraient avoir un impact moins important sur la représentation du mot ciblé, sans toutefois être ignorés.

1.3.4 Tâche de recherche de mots analogues

Afin de vérifier notre approche de pondération des contextes proposée, nous avons appliqué les modèles CBOW et Skip-gram sur la tâche de recherche de mots analogues, appelée *Semantic-syntactic Word relationship test* [Mikolov et al., 2013a]. Ce test s'appuie sur les propriétés algébriques : en effet, il a été montré que des schémas (*patterns*) sémantiques et syntaxiques peuvent être reproduits avec des opérations algébriques. Par exemple, l'opération sur les vecteurs *Oncle - Neveu + Tante* devrait produire un vecteur de représentation proche de *Nièce*. Ce test contient 19 544 questions (8 869 sémantiques et 10 675 syntaxiques), permettant de vérifier que le modèle a bien appris les relations entre des couples de mots. La tâche consiste alors à prédire un mot sachant la relation entre 3 mots fournis. Par exemple, une des questions est de trouver le mot proche à *Oslo* de la même façon où *France* est proche de *Paris*. Nous nous retrouvons alors à calculer l'opération sur les vecteurs *France - Paris + Oslo*.

La fonction de pondération a été définie comme suit dans nos expériences :

$$\delta(w) = \frac{1 + log(2)}{1 + log(d(w))} \tag{1.3}$$

Les modèles Word2vec avec et sans pondération ont été entraînés sur un corpus de 4 milliards de mots pour un vocabulaire d'environ 1 million de mots, à partir de données provenant de Wikipedia, Gigaword, Brown corpus et Word Language Modeling Benchmark. Le tableau 1.4 présente les résultats obtenus sans et avec pondération des modèles CBOW et Skip-gram en faisant varier la taille des mots dans le contexte (10, 15 et 100 mots avant et après le mot ciblé), alors que le tableau 1.5 présente les résultats des mêmes approches mais en faisant varier la taille des plongements de mots (*i.e.* de la couche cachée), avec des dimensions de 120 et 300.

Ces tableaux montrent que les modèles utilisant l'approche par pondération que nous proposons permettent d'obtenir globalement de meilleurs résultats, peu importe la taille du contexte ou la dimension des plongements de mots. Les meilleurs résultats sont obtenus avec l'approche Skip-gram en prenant un contexte de mots relativement court (10 mots avant et après le mot

	Skip-gram			CBOW		
Taille du contexte	10	15	100	10	15	100
Sans pondération (baseline)	50,0	50,9	43,7	39,0	38,9	36,9
Avec pondération	**55,0**	53,7	51,4	39,9	39,6	**43,9**

Tableau 1.4 – Performances en termes d'exactitude (%) sur la tâche de recherche de mots analogues avec et sans pondération du contexte en faisant varier la taille des mots du contexte (10, 15 et 100). La dimension des plongements de mots est fixée à 300.

	Skip-gram		CBOW	
Taille des plongements de mots	120	300	120	300
Sans pondération (baseline)	43,9	50,0	29,0	39,0
Avec pondération	45,1	**55,0**	30,3	**39,9**

Tableau 1.5 – Performances en termes d'exactitude (%) sur la tâche de recherche de mots analogues avec et sans pondération du contexte en faisant varier la taille des plongements de mots (120 et 300). La taille des mots dans le contexte est fixée à 10.

ciblé) par rapport au modèle CBOW, ce qui reste cohérent avec ce qui a été montré dans [Mikolov et al., 2013b]. Concernant la dimension des plongements de mots (tableau 1.5), l'approche Skip-gram avec utilisation d'un contexte pondéré apporte de meilleurs résultats lorsque celui-ci est élevé (dimension 300).

Nous avons également appliqué ces modèles avec succès sur une tâche d'identification de thématiques dans des conversations orales. Les résultats de ces expériences peuvent se trouver dans l'article [Janod et al., 2016b] ou dans la thèse de K. Janod [Janod, 2017].

1.4 Conclusion

Nous nous sommes intéressés, dans ce chapitre, à la représentation d'éléments en prenant en compte le contexte de séquences. Ici, les éléments ont pris la forme de genres d'émissions télévisées, et, de façon plus classique, de mots. Les premiers travaux présentés ont été réalisés dans un contexte de travail original, avec des données provenant de flux parallèles potentiellement liés entre eux, à savoir ici des flux continus d'émissions télévisées. Nos travaux de recherche sont liés au projet ANR ContNomina et à la thèse CIFRE de M. Bouaziz, ce dernier ayant tout d'abord proposé un corpus annoté en genres d'émissions télévisées à partir de 4 chaînes de télévision collectées en continu sur deux jours. Une catégorisation en genres a été proposée, simplifiée par rapport à celle proposée par l'INA. Les genres des émissions ont ici été annotés manuellement sur le signal audio afin d'avoir le début et la fin précise de chaque catégorie. Dans un second temps, une approche à base de réseaux de neurones récurrents a été proposée pour tirer profit de flux parallèles pour prédire un élément. En effet, les approches qui étaient

actuellement proposées n'étaient conçues que pour traiter des séquences provenant d'un seul flux. L'approche originale PLSTM a pu être évaluée sur la tâche de prédiction de genres d'émissions télévisées. Nous avons montré que la prise en compte des flux parallèles permettait d'améliorer les résultats du flux ciblé, et ce, particulièrement avec l'architecture PLSTM. Bien qu'appliqué à une tâche précise de prédiction, ce type d'approche pourrait être intéressant pour d'autres problématiques, par exemple pour traiter des listes d'hypothèses de transcription (*n-best*) en reconnaissance automatique de la parole (RAP) afin d'améliorer la transcription automatique.

Nous nous sommes ensuite intéressés à un autre problème lié au contexte de séquences, à savoir la position des éléments dans la séquence continue. Ce travail a pris place pendant la thèse CIFRE de K. Janod sur la compréhension du langage. Nous avons ici présenté une méthode permettant d'améliorer la représentation par plongement de mots de l'approche Word2vec, en intégrant une pondération des mots du contexte selon leur position relative au mot ciblé. En effet, le contexte était ici utilisé en tant que sac-de-mots, leur représentation étant en *one-hot encoding* (0 ou 1 selon leur présence), ce qui équivalait à considérer des contextes proches ou éloignés de la même façon. La prise en compte de la position des mots dans le contexte a montré son intérêt sur la tâche de recherche de mots analogues.

Il est intéressant de voir le lien entre recherche académique et problématiques actuelles du monde industriel. Nous l'avons en particulier observé durant le partenariat avec l'entreprise EDD (thèse CIFRE de M. Bouaziz), pour qui, au final, la structuration de flux audio-visuels diffusés en continu, et *a fortiori* en parallèle (plusieurs centaines de chaînes télévisées et radiophoniques à traiter en temps réel), constitue un enjeu stratégique pour l'entreprise tout en faisant émerger des problématiques de recherche originales.

Les expériences présentées dans ce chapitre s'appuient sur des séquences de données bien formées, que ce soit les genres annotés manuellement ou des textes écrits. Or, les documents traités ne sont pas toujours aussi "propres", en particulier en TAL lorsque des tâches s'appuient sur des transcriptions automatiques potentiellement erronées par des erreurs des systèmes de RAP. Le chapitre suivant se concentre sur la robustesse des représentations des mots dans un contexte de données très fortement erronées, et donc bruitées.

Chapitre 2

Représentations latentes pour la classification de documents

Sommaire

2.1 Introduction

Comme nous l'avons vu dans l'introduction du chapitre précédent, les approches par sac-de-mots ont longtemps été la norme en TALN et RI. Elle permettent de mettre un poids sur les mots, l'idée étant d'identifier les plus importants. Nous avons également mis en lumière certaines limites des représentations par sac-de-mots, la première étant l'absence de prise en compte du contexte dans lequel apparaissent ces mots, palliée, par exemple, par des approches de type n-gramme. Le chapitre 1 a alors été l'occasion de décrire des travaux que nous avons pu mener pour améliorer la prise en compte du contexte, que ce soit en intégrant la position des mots

du contexte dans la méthode par plongements de mots Word2vec, ou que ce soit en prenant en compte des informations issues de séquences parallèles avec la proposition d'une approche par réseaux de neurones récurrents (PLSTM).

Les expériences, dans le chapitre 1, ont été réalisées sur des données que nous pouvons considérer comme relativement *propres* (voir partie 1.3), avec des textes correctement formés d'un point de vue grammatical et orthographique. Cependant, ce type de documents *correctement* écrits n'est qu'une forme particulière, ne représentant finalement qu'une partie spécifique (et restreinte) des travaux auxquels j'ai pu participer ces dernières années. Les transcriptions automatiques obtenues par un système de reconnaissance automatique de la parole (RAP), en particulier sur de la parole spontanée, sont un bon exemple de contenus textuels au centre de mes problématiques de recherche, qui posent des problèmes particuliers tels que :

— *Erreurs de transcription* : les mots automatiquement transcrits par les systèmes de RAP peuvent être incorrects. Nous nous trouvons donc face à du contenu textuel potentiellement erroné par ce processus automatique, allant de quelques erreurs de transcription à un nombre très important selon les documents audio traités et leur difficulté inhérente aux conditions acoustiques, aux locuteurs et leur qualité d'élocution, au registre de langue utilisé... Cela aura donc potentiellement un impact négatif sur les mots et leurs représentations, et *a fortiori*, sur la représentation globale des documents.
— *Niveau de langue* : en opposition à des documents bien écrits respectant les règles linguistiques d'une langue, nous pouvons être face à un langage courant, qui ne respecte pas toutes les règles de la langue mais qui a été défini par ses propres usages. Nous parlons alors de niveaux de langue [Authier and Meunier, 1972], pouvant varier d'un média à l'autre (oral ou écrit) mais également entre les individus.
— *Vocabulaire particulier* : Tout comme le niveau de langue, le vocabulaire employé peut être très changeant d'un individu à l'autre, d'un domaine traité...

L'accumulation de tous ces problèmes, qui ne sont ici pas exhaustifs, ajoute à la difficulté de traiter automatiquement ces documents. En effet, en TAL, le contenu textuel constitue souvent le point d'entrée à d'autres tâches, celles-ci permettant en général de valider les méthodes de représentation des mots, comme nous l'avons vu dans le chapitre précédent.

Dans ce chapitre, nous proposons de décrire une partie des travaux réalisés dans le cadre de la thèse de Mohamed Morchid, passant ici du problème de représentation des mots (chapitre 1) à celui de la représentation de documents textuels complets. Ces travaux prennent place dans le cadre du projet ANR SuMACC (voir partie 9.2.1), les expériences étant menées sur la tâche d'identification de thématiques d'appels téléphoniques issue du projet ANR DECODA. La partie 2.2 est tout d'abord l'occasion de décrire le corpus DECODA et de comprendre les enjeux liés aux conversations traitées, conduisant notamment à devoir utiliser des transcriptions automatiques très bruitées. Nous proposons ensuite, dans la partie 2.3, une étude comparative entre

une représentation utilisant le contenu textuel direct, à travers l'approche TF-IDF-Gini, et une représentation dépassant le niveau *mot*, au moyen d'une représentation latente par espaces de thèmes (LDA). L'objectif visé ici est de montrer que cette représentation de plus haut niveau est plus robuste lorsque l'on traite des documents très bruités, ici des transcriptions automatiques avec des taux d'erreur très élevés [Morchid et al., 2014e,d].

Cependant, cette approche par espaces de thèmes souffre de problèmes liés à la méthode elle-même, et en particulier des hyper-paramètres nécessaires pour définir l'espace : selon la configuration choisie, même avec peu de changements dans ceux-ci, une variation très importante au niveau des performances a pu être observée. De ces constats, une approche originale, s'inspirant des travaux en reconnaissance du locuteur, a été proposée [Morchid et al., 2014a, 2015a], cherchant alors à tirer profit non pas d'une seule représentation, mais d'une multitude de représentations, réduites ensuite pour en obtenir une considérée comme plus robuste aux bruits. Nous présentons alors plus en détails le contexte de cette proposition dans la partie 2.5 et détaillons les performances obtenues.

2.2 Corpus DECODA

2.2.1 Données et annotation

Ce corpus a été collecté pendant le projet ANR DECODA [Béchet et al., 2012]. Il consiste en un ensemble de conversations téléphoniques entre humains (dialogues *humain-humain*, en opposition à des dialogues *humain-machine*) réalisées dans le cadre du service client de la régie autonome des transports parisiens (RATP). Dans le cadre de nos expérimentations [1], le corpus récolté est composé de 1 242 conversations téléphoniques, correspondant à environ 74 heures d'audio. Les données, découpées en 3 corpus (apprentissage, développement et test) et annotées en 8 thématiques, sont décrites dans le tableau 2.1.

La classe (thématique) correspond à la raison de l'appel du client (*l'appelant*) et a été annotée manuellement par la personne recevant l'appel (*le conseiller*). Notons que dans le cadre de nos travaux, nous ne nous sommes intéressés qu'à la classe principale associée à un dialogue : dans les faits, de par la variabilité inhérente à la discussion entre le client et le conseiller, des sous-classes (sous-thématiques) ont été fournies dans certaines conversations. Dans la tâche de classification automatique, un dialogue ne peut être associé qu'à une seule classe, et non à plusieurs. Ce côté multi-thématiques par dialogue rajoute donc une difficulté supplémentaire. La figure 2.1 présente un court exemple de dialogue possible entre un appelant, ici le client, et le conseiller de la régie RATP qui doit résoudre son problème. Bien que la raison de l'appel soit liée à un problème sur la classe *Offre spéciale* (thématique principale du dialogue), nous constatons qu'un second

1. Notons qu'au moment de nos expériences, le corpus final DECODA n'était pas encore disponible. Pour avoir des informations sur la dernière version du corpus, le lecteur pourra se référer par exemple à [Lailler et al., 2016].

Classe	Apprentissage	Développement	Test
Problèmes d'itinéraires	145	44	67
Objets trouvés	143	33	63
Horaires	47	7	18
Carte de transport	106	24	47
État du traffic	202	45	90
Tarifs	19	9	11
Procès verbaux	47	4	18
Offre spéciale	31	9	13
Total	**740**	**175**	**327**

Tableau 2.1 – Découpage du corpus DECODA.

thème est possible (*Carte de transport*), ce qui appuie le fait que le contenu linguistique d'une conversation soit multi-thèmes.

2.2.2 Transcription automatique

Nos travaux ne s'appuient ici que sur les transcriptions automatiques du corpus (nous n'avons utilisé aucune information provenant du signal acoustique autre que ce qu'utilise le système de RAP). La transcription de chaque dialogue a été réalisée au moyen du système de RAP du LIA, nommé Speeral [Linares et al., 2007]. Le système de RAP intégrait un modèle acoustique entraîné sur 150 heures de parole dans des conditions téléphoniques. Le dictionnaire était constitué de 5 782 mots, réduit au maximum pour traiter le vocabulaire spécifique lié aux appels client/conseiller de la RATP. Enfin, un modèle de langage tri-gramme a été obtenu en adaptant un modèle de langage générique avec l'ensemble de transcriptions manuelles du corpus d'apprentissage. Le système atteignait un taux d'erreur-mot (WER) très élevé, ce qui est notamment à l'origine des différentes solutions que nous avons proposées : un WER initial de 45,8 % sur l'ensemble d'apprentissage, de 59,3 % sur l'ensemble de développement, et de 58 % sur l'ensemble de test. Ces WER élevés sont principalement dus à un contexte de parole très spontanée [Dufour, 2010] et à des environnements acoustiques très dégradés pour certaines conversations lorsque, par exemple, les utilisateurs appelaient sur leur téléphone portable dans des lieux potentiellement très bruyants (gares, métro, rues...). Afin d'aider les représentations s'appuyant sur le contenu textuel, nous avons appliqué une liste de 126 mots d'arrêt [2] (*stop list*) sur les transcriptions automatiques, résultant en un WER de 33,8 % sur l'apprentissage, 45,2 % sur le développement, et 49,5% sur le test.

2. http ://code.google.com/p/stop-words/

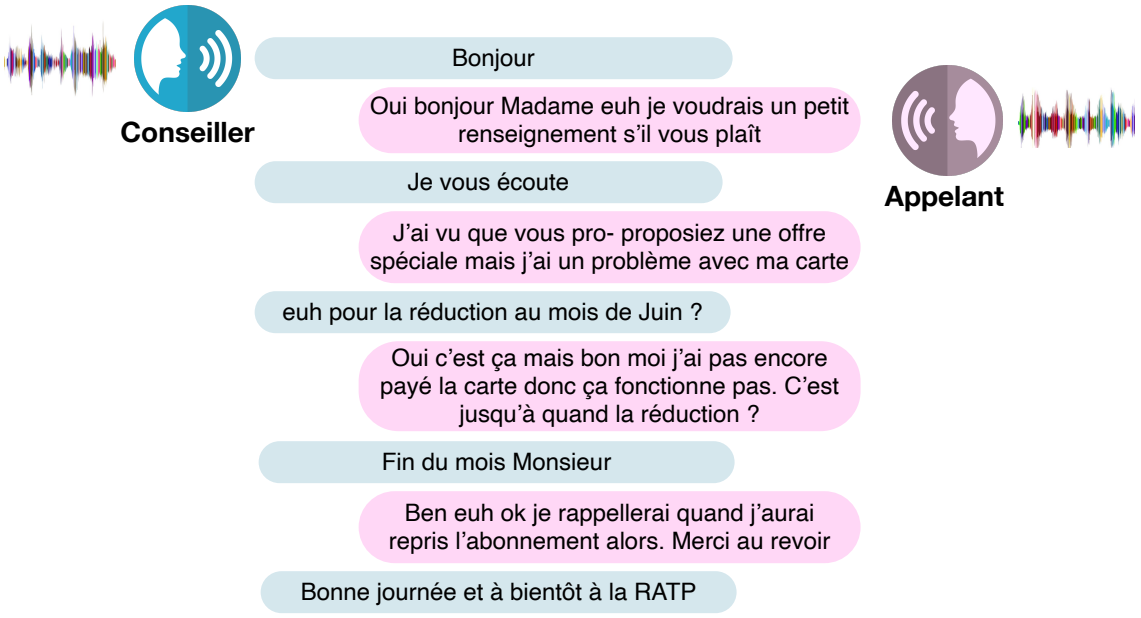

FIGURE 2.1 – Exemple d'un dialogue du projet DECODA entre un appelant et un conseiller de la RATP pour un problème lié à *Offre Spéciale*. Le texte de la conversation correspond à la transcription manuelle.

2.3 Représentation des documents

En RI, la principale caractéristique utilisée pour représenter un mot a longtemps été la fréquence de ce terme (*Term Frequency*, TF), qui permet d'obtenir son importance au sein d'un document, associée à sa rareté au sein du corpus complet de documents (*Inverse Document Frequency*, IDF). Le vecteur des pondérations TF-IDF associées à un document permet alors de le représenter, ce vecteur étant censé permettre de différencier les mots pertinents (*i.e.* qui participent à caractériser le document), des mots non pertinents. Dans notre contexte d'identification de la thématique de conversations téléphoniques, cette représentation par sac-de-mots doit permettre de reconnaître cette thématique. Nous donnons quelques détails sur cette approche, associée avec le critère de pureté Gini, dans la sous-partie 2.3.1.

La représentation TF-IDF, qui s'appuie sur le contenu direct des documents, n'apparaissait cependant pas comme une solution satisfaisante dans notre contexte de travail, puisque nous sommes face à des transcriptions automatiques fortement bruitées : ainsi, par exemple, les erreurs de transcription conduiraient à une représentation incorrecte des mots, et donc, par extension, à une représentation incorrecte des documents.

Ainsi, nous avons supposé que les conversations devraient être considérées dans un espace de représentation plus élevé que le niveau *mot*. Nous nous sommes alors intéressés aux représen-

tations par espaces de thèmes, permettant de projeter le document dans un espace plus grand, et devant également mieux gérer le problème des *multi-thèmes* (une thématique principale, mais également des thématiques secondaires) dans les conversations. La projection des mots transcrits automatiquement dans un espace plus abstrait devrait augmenter la robustesse aux erreurs de RAP. La sous-partie 2.3.2 présente alors la représentation dans un espace de thèmes que nous avons choisi d'étudier, à savoir l'approche par allocation latente de Dirichlet (*Latent Dirichlet Allocation*, LDA).

2.3.1 Représentation directe du contenu

L'approche classique par pondération TF-IDF [Jones, 1972] a longtemps été utilisée pour extraire des mots représentatifs de documents textuels. Dans les derniers travaux, son association avec le critère de pureté Gini [Singh et al., 2010] rend l'approche plus robuste. Dans notre tâche, à savoir l'identification de la thématique d'une conversation, l'approche proposée dans [Morchid et al., 2014h] se déroule en deux étapes. La première étape consiste à extraire, à partir d'un corpus d'apprentissage, un ensemble de mots représentatifs à partir de toutes les thématiques que nous souhaitons identifier dans les conversations. La seconde étape permet de projeter, dans cet ensemble de mots représentatifs, un nouveau document : cela permet d'obtenir une représentation de ce document par rapport aux thématiques développées dans les conversations.

1. Apprentissage. Soit un corpus D de conversations d associé à un vocabulaire de mots $\mathbf{V} = \{w_m\}_{m=1}^{N}$ de taille N où d est ici un sac-de-mots [Harris, 1954]. Un mot (ou terme) w de $\mathbf{V}$ est choisi en fonction de son importance δ dans la thématique t définie comme suit :

$$\delta_t^w = tf_t(w) idf(w) gini(w) \ . \tag{2.1}$$

Le critère de pureté Gini $gini(w)$ est commun à toutes les thématiques $\mathbf{T}$:

$$gini_t(w) = 1 - \sqrt{\sum_{i=1}^{|\mathbf{T}|} p_i^2} \ . \tag{2.2}$$

Ensuite, les mots ayant les scores Δ les plus élevés pour toutes les thématiques $\mathbf{T}$ constituent un sous-ensemble de mots représentatifs $\mathbf{V_\Delta}$. Chaque thématique $t \in \mathbf{T}$ a son propre score δ_t et sa propre fréquence $\gamma_t = p(t)$ qui est la fréquence des conversations $d \in t$ dans le corpus D. Notons qu'un même mot w peut être présent dans différentes thématiques, mais avec des scores différents en fonction de son importance dans la thématique :

$$\Delta(w) = p(w|t, t \in \mathbf{T}) = \sum_{t \in \mathbf{T}} p(w|t)p(t) = \langle \overrightarrow{\delta^w}, \overrightarrow{\gamma} \rangle_{t \in \mathbf{T}} \ . \tag{2.3}$$

2. Extraction d'un vecteur de caractéristiques d'une conversation. Pour chaque dialogue $d \in$

D, un vecteur de caractéristiques V_d^s est extrait. La n^{ieme} $(1 \leq n \leq |\mathbf{V_\Delta}|)$ caractéristique $V_d^s[n]$ est composée du nombre d'occurrences du mot w_n $(|w_n|)$ dans d et du score Δ de w_n dans l'ensemble de mots représentatifs $\mathbf{V_\Delta}$:

$$V_d^s[n] = |w_n| \times \Delta(w_n) \ . \tag{2.4}$$

Ce vecteur de caractéristiques V_d^s est alors la représentation vectorielle de la conversation.

2.3.2 Représentation dans un espace de thèmes

L'autre approche consiste alors à projeter les conversations, et donc leur contenu textuel, dans un espace de plus haut niveau. Nous avons ici choisi le modèle génératif probabiliste par allocation latente de Dirichlet (LDA), initialement utilisé pour la détection de thématiques dans un document [Blei et al., 2003]. L'approche considère qu'un document est constitué d'un mélange de thèmes cachés. LDA intègre le fait qu'un thème est associé à chaque occurence d'un mot dans le document, plutôt que de fournir un seul thème global. Il est alors possible de changer de thème d'un mot à un autre. Dans le cadre de nos documents très bruités, cela permet de ne pas utiliser les mots directement, puisque potentiellement erronés, mais d'utiliser un vecteur de caractéristiques correspondant à la distribution du document dans l'espace de thèmes LDA.

L'échantillonnage de Gibbs est utilisé pour estimer les caractéristiques d'un dialogue d dans l'espace de thèmes LDA. Cet algorithme nous permet d'obtenir des échantillons des paramètres de distribution θ connaissant un mot w d'un document et un thème caché z. Un espace de n thèmes est obtenu avec, pour chaque thème z, la probabilité de chaque mot w du vocabulaire $\mathbf{V}$ sachant z $(P(w|z) = V_z^w)$. Un vecteur de caractéristiques V_d^z peut alors être obtenu pour d. La k^{ieme} caractéristique $V_d^z[k]$ (où $1 \leq k \leq n$) dans le vecteur est la probabilité du thème caché z_k sachant la conversation d :

$$V_d^z[k] = P(z_k|d).$$

Le vecteur de caractéristiques V_d^z représente au final la projection du document d dans l'espace de thèmes. La figure 2.2 présente un exemple d'extraction de ce vecteur.

V_d^z (approche LDA) et V_d^s (approche TF-IDF) seront les représentations de document comparées dans ce travail. De plus amples détails sur la représentation des conversations peuvent se trouver dans [Morchid et al., 2014e].

2.4 Comparaison des méthodes de représentation du contenu

Nous proposons de nous focaliser sur l'évaluation des deux représentations proposées dans la tâche de détection de la thématique principale dans les conversations entre humains du corpus DECODA (voir partie 2.2). Cette première expérience avait pour objectif de mettre en lumière

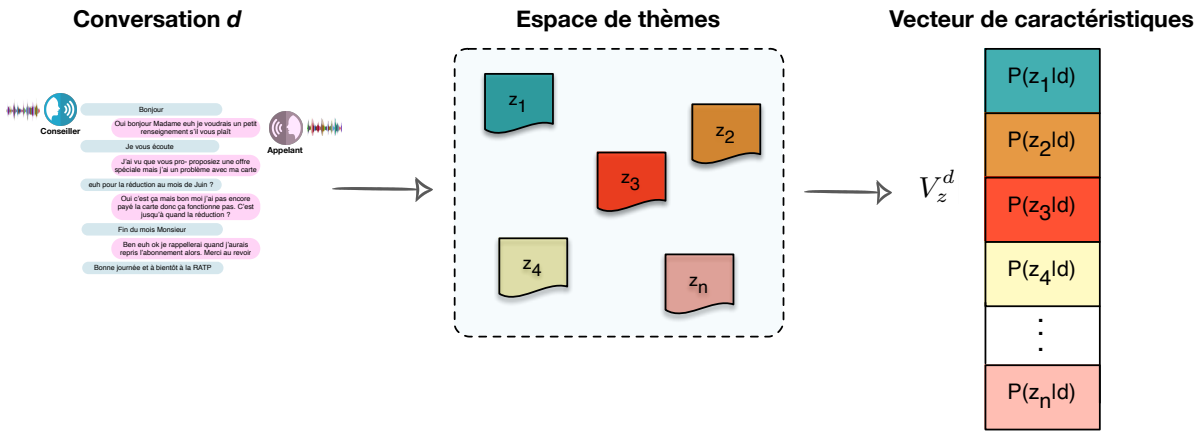

FIGURE 2.2 – Exemple d'extraction d'un vecteur de caractéristiques d'une conversation à partir d'un espace de thèmes.

les limites que pose la représentation de documents ne s'appuyant que sur son contenu direct. Nous présentons le protocole expérimental dans la sous-partie 2.4.1, puis les résultats obtenus selon plusieurs configurations des documents, à savoir les transcriptions manuelles et surtout les transcriptions automatiques, dans la sous-partie 2.4.2.

2.4.1 Protocole expérimental

Pour l'approche TF-IDF, l'extraction de l'ensemble des mots représentatifs des différentes thématiques (ici, les classes) des conversations (phase *1. Apprentissage*) a été réalisée sur le corpus d'apprentissage de DECODA. Les différentes conversations du corpus de test ont ensuite été projetées dans cet ensemble de mots (phase *2. Extraction d'un vecteur de caractéristiques d'une conversation*) afin d'obtenir un vecteur de caractéristiques V_d^s de la taille de l'ensemble des mots représentatifs. Dans ces expériences, nous avons proposé de faire varier le nombre de mots représentatifs des classes de 800 jusqu'au nombre total de mots contenus dans le corpus (7 920 mots) afin d'évaluer l'impact de la taille du vecteur sur les performances.

Pour l'approche par espace de thèmes LDA, un ensemble de 19 espaces de thèmes a été évalué, en faisant varier ici le nombre de thèmes de 5 à 300. Ces espaces ont également été entraînés avec le corpus d'apprentissage de DECODA au moyen de l'implémentation LDA Mallet [McCallum, 2002]. Les conversations du corpus de test ont alors été projetées dans cet espace de thèmes, permettant d'obtenir, pour chaque conversation, le vecteur de caractéristiques V_d^z.

Les deux représentations, TF-IDF et LDA, ont ensuite été fournies en entrée d'un classifieur afin de déterminer la classe (thématique) de la conversation. Un classifieur SVM a alors été entraîné au moyen de la bibliothèque LIBSVM [Chang and Lin, 2011]. Les paramètres du SVM

ont été optimisés par validation croisée sur le corpus d'apprentissage.

Pour comparaison, les expériences présentées ici ont été menées en utilisant les transcriptions manuelles (REF) et les transcriptions automatiques (RAP) obtenues au moyen du système de RAP Speeral présenté dans la sous-partie 2.2.2. Outre des conditions d'apprentissage et de test identiques, avec transcriptions manuelles propres ($REF \rightarrow REF$) et transcriptions automatiques bruitées ($RAP \rightarrow RAP$), plusieurs mélanges entre les conditions ont également été utilisés dans ces expériences, afin de voir l'impact entre des conditions d'apprentissage propres et des conditions de test bruitées ($REF \rightarrow RAP$), ou une concaténation des conditions dans l'apprentissage pour évaluer sur des conditions de test bruitées ($RAP + REF \rightarrow RAP$).

2.4.2 Résultats

Le tableau 2.2 présente les meilleures précisions obtenues sur le corpus de test DECODA pour l'approche de représentation par sac-de-mots TF-IDF et la représentation de plus haut niveau LDA par espace de thèmes.

Configuration		**Précisions maximales (%)**			
Apprentissage	**Test**	**# mots**	**TF-IDF**	**# thèmes**	**LDA**
REF	REF	800	79,7	100	**86,6**
REF	RAP	8 000	69,7	40	**77,0**
RAP	RAP	800	73,5	60	**81,4**
REF+RAP	RAP	2 400	72,2	100	**78,7**

Tableau 2.2 – Précisions maximales (%) obtenues sur la tâche de classification en thématiques de conversations au moyen des deux représentations (TF-IDF et LDA) selon les différentes configurations d'apprentissage et de test considérées.

Comme nous avons pu le constater, les meilleurs résultats de classification sont obtenus par la configuration utilisant les transcriptions de référence pendant l'apprentissage et la phase de test ($REF \rightarrow REF$), et ce dans les 2 représentations des mots. Ce résultat était attendu et constitue plutôt les gains maximaux que nous pouvions espérer avec des transcriptions *propres* (*i.e.* sans bruit lié aux erreurs de transcription). Pour cette première configuration optimale, la représentation LDA est clairement meilleure que la représentation TF-IDF, avec un gain de 6,9 points. En nous concentrant ensuite sur la condition d'évaluation avec des transcriptions automatiques (RAP), les meilleures performances sont obtenues en utilisant des données d'apprentissage également bruitées ($RAP \rightarrow RAP$). Un gain de 7,9 points est noté avec la représentation LDA par rapport à la représentation TF-IDF sur les transcriptions automatiques. Notons que des conditions d'apprentissage différentes ($REF \rightarrow RAP$) ou mixtes ($REF + RAP \rightarrow RAP$) obtiennent toutes les deux des résultats inférieurs, peu importe la représentation de mots considérée.

Il semble clair que l'utilisation de configurations d'apprentissage et de test comparables

permet d'obtenir les meilleures performances de classification, que ce soit sur des transcriptions manuelles ou automatiques. De même, la représentation par espace de thèmes, donc dans un espace plus grand que celui des mots, surpasse l'approche classique TF-IDF, quelque soit la configuration d'apprentissage et de test choisie.

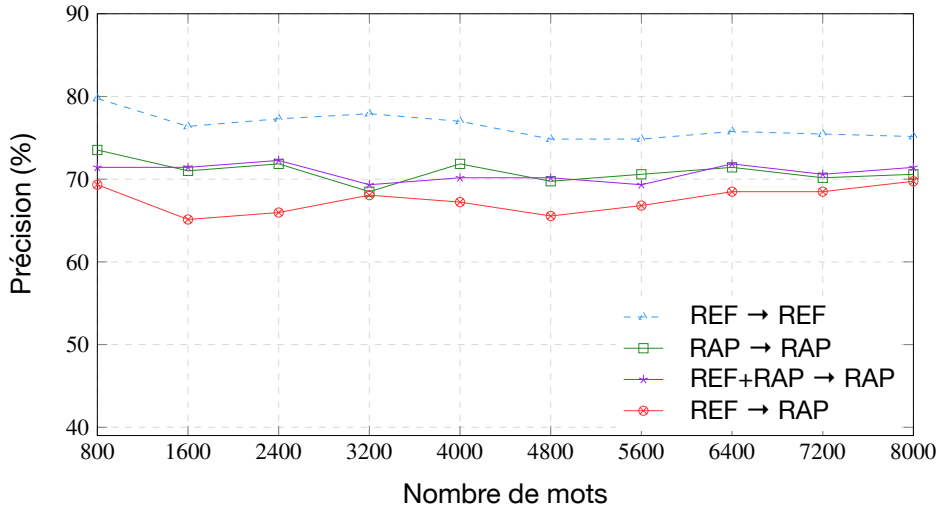

FIGURE 2.3 – Précision (%) sur la tâche de classification en thématiques de conversations au moyen d'une représentation TF-IDF en faisant varier le nombre de mots représentatifs. Différentes configurations d'apprentissage et de test sont reportées.

Les figures 2.3 et 2.4 présentent les résultats obtenus, en termes de précision, sur la tâche de classification en thématiques des conversations du corpus de test DECODA, en utilisant respectivement la représentation par sac-de-mots TF-IDF et la représentation LDA par espace de thèmes. Nous pouvons finalement noter que la performance avec la représentation LDA a tendance à fluctuer lorsque le nombre de thèmes varie, au contraire de l'approche TF-IDF qui semble moins sujette aux variations de performance selon le nombre de mots choisis. Cela pourrait s'expliquer par le taux d'erreur-mot (WER) élevé que nous avions sur le corpus DECODA. En effet, les mots choisis comme pertinents dans certains espaces de thèmes pourraient avoir été mal transcrits dans une proportion importante (et donc conduire plus massivement à des mauvaises représentations). Changer le nombre de thèmes revient à changer l'importance de certains mots, et pourrait expliquer ces variations. Il n'y a pas un changement aussi abrupte avec l'approche TF-IDF, qui conserve, elle, sa base de mots pertinents (nous la faisons ici simplement *grossir*). Cette hypothèse pourrait être soutenue par une baisse de fluctuation plus faible pour la représentation LDA sur les conditions de référence ($REF \rightarrow REF$) par rapport aux conditions automatiques (RAP).

Ce problème de variabilité au niveau de la performance des résultats est au coeur d'une partie

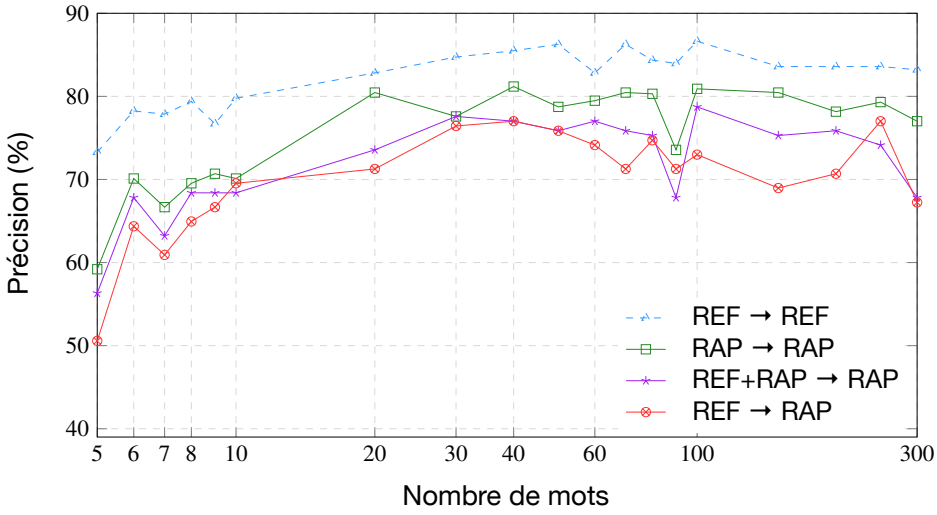

FIGURE 2.4 – Précision (%) sur la tâche de classification en thématiques de conversations au moyen d'une représentation par espace de thèmes (LDA) en faisant varier le nombre de thèmes. Différentes configurations d'apprentissage et de test sont reportées.

des travaux entrepris durant la thèse de M. Morchid. La représentation robuste multi-vues, que nous présentons dans la partie 2.5, constitue une réponse de sa thèse à ce problème.

2.4.3 Autres propositions de représentation haut niveau

La robustesse des représentations face aux erreurs de transcription automatique, et de manière plus générale ce que nous avons considéré comme des transcriptions fortement bruitées (parole spontanée combinée à des taux d'erreur élevés des transcriptions), a été traitée au travers d'autres approches proposées à la fois par M. Morchid et par K. Janod durant leurs thèses respectives. Nous pouvons par exemple citer :

— Modèle *Author-topic* [Morchid et al., 2014c] : au lieu de prendre simplement en compte les mots pour construire les espaces de thèmes, les thématiques des dialogues sont ici intégrées pour construire l'espace de représentation.

— Modèles utilisateurs [Morchid et al., 2014d] : l'idée ici a été de construire des espaces de thèmes, non pas à partir des documents de manière générale, mais de considérer des espaces par type d'intervenant dans la conversation (dans le cadre du corpus DECODA, des modèles pour *Conseiller* et *Appelant*).

— Approches neuronales [Janod et al., 2017] : plusieurs propositions d'architectures neuronales ont été proposées pour contenir le bruit présent dans les transcriptions automatiques, comme des solutions s'appuyant sur des auto-encodeurs débruitants, ou des auto-encodeurs empilés.

Dans la partie suivante, nous proposons de détailler une solution permettant de répondre au problème de la compensation du bruit contenu dans les transcriptions automatiques mais également au problème du choix des paramètres des représentations par espace de thèmes.

2.5 Représentation robuste multi-vues

2.5.1 Contexte

Dans les travaux présentés précédemment, au travers de la thèse de M. Morchid, une approche haut niveau à base d'espaces de thèmes a obtenu de meilleures performances, comparativement à une approche de représentation directe des mots (TF-IDF), dans le cadre d'une tâche de classification automatique. Cela a permis de montrer qu'il était nécessaire de dépasser le niveau *mot* pour obtenir des représentations robustes aux erreurs de transcription. Néanmoins, malgré le niveau de performance élevé atteint, nous avons noté une forte variabilité des résultats selon le nombre de thèmes choisis. Si nous reprenons la figure 2.4, où le nombre de thèmes varie de 5 à 300, les résultats oscillent entre environ 50 % de précision de classification et 85 % environ. Même si elles sont moins importantes, ces variations ont pu être observées entre des nombres de thèmes très proches, comme nous pouvons le voir sur les courbes entre 60 et 100 thèmes.

Ce problème de réglage des hyper-paramètres est un problème récurrent dans beaucoup de méthodes d'apprentissage automatique, l'approche LDA n'y échappant pas. Nous avons vu que le nombre de thèmes est un paramètre déterminant, mais d'autres hyper-paramètres de l'approche LDA peuvent aussi influer sur les performances, par exemple les paramètres α et β [Blei et al., 2003]. Nous avons alors proposé d'estimer tout d'abord un très grand nombre d'espaces de thèmes différents, en faisant varier légèrement la valeur, à chaque fois, d'un des méta-paramètres du modèle. La projection d'un document dans chacun de ces espaces peut alors être considérée comme une vue particulière de ce document. Notre objectif était ici de tirer profit de chacune de ces vues, tout en considérant que celles-ci introduisent une variabilité supplémentaire en raison de leur diversité. Une difficulté inhérente à une approche multi-vues concerne la potentielle grande diversité des vues, introduisant à la fois une variabilité pertinente, nécessaire pour représenter différents contextes du document, et une variabilité non pertinente, voire nuisible, liée aux représentations par espace de thèmes.

Nous avons alors proposé de réduire cette variabilité en fournissant une représentation compacte de ces vues multiples au moyen d'une technique d'analyse factorielle [Morchid et al., 2014a, 2015a], déjà utilisée avec succès dans le domaine de la reconnaissance du locuteur. Dans ce domaine, le paradigme d'analyse factorielle est utilisé comme modèle de décomposition qui permet de séparer l'espace de représentation en deux sous-espaces, contenant respectivement des informations utiles et inutiles. L'approche i-vecteur, permettant d'obtenir un vecteur de caractéristiques compact à partir de données de grandes dimensions, a alors été adaptée par M.

Morchid pour le traitement de documents écrits afin de compenser les variabilités nuisibles liées à la multiplication des modèles LDA. Au final, nous proposons une représentation compacte robuste à partir de multiples espaces de thèmes.

Dans la partie suivante (sous-partie 2.5.2), nous présentons les grands principes de la représentation compacte c-vecteur. Nous présentons enfin les résultats d'une expérience que nous avons menée sur la tâche de détection de la thématique de conversations (sous-partie 2.5.3).

2.5.2 Approche c-vecteur fondée sur les i-vecteurs

Concepts des i-vecteurs

Pendant longtemps, une représentation des locuteurs avec un modèle de mélange gaussien (GMM) a été la référence [Reynolds, 2009]. Plus récemment, l'approche i-vecteur [Dehak et al., 2010] a permis de grandes avancées en termes de performance en reconnaissance du locuteur. Les i-vecteurs ont été conçus comme une technique de réduction de la grande dimension des données fournies en entrée en un vecteur de caractéristiques de petite taille, dont l'objectif est de retenir le maximum d'informations pertinentes liées au locuteur. La technique est une extension des modèles d'analyse factorielle, proposant ici de contenir à la fois la session et le locuteur. Il s'agit ici de capturer l'ensemble des variabilités acoustiques au moyen d'une matrice rectangulaire de faible rang T appelée *espace de variabilité totale*. Le super-vecteur dépendant du locuteur et du canal de transmission M des moyennes GMM concaténées est projeté dans T comme suit :

$$M = m + Tw \tag{2.5}$$

où m est le super-vecteur du Modèle Universel (UBM) [3] et w est le i-vecteur estimé. Outre une plus grande robustesse aux variabilités contenues dans le signal de parole, la représentation i-vecteur permet de représenter des séquences de durées variables par un vecteur de taille fixe.

Proposition : représentation c-vecteur

La proposition a été ici d'adapter le concept des i-vecteurs, orienté locuteur, au problème de représentation de documents textuels. Il s'agissait de pouvoir compresser un ensemble de vues multiples d'espaces de thèmes, construites en faisant varier les hyper-paramètres du modèle, en un seul super-vecteur de taille réduite contenant les informations importantes du document, tout en enlevant la variabilité nuisible liée à ces vues.

En premier lieu, il s'agit de créer un ensemble d'espaces de thèmes à partir d'un corpus de documents $D = \{d_1, d_2, \ldots, d_n\}$ appris avec l'approche LDA en faisant varier les valeurs d'hyper-paramètres. Puis chaque document d est projeté dans chaque espace de thèmes permettant d'obtenir un vecteur de caractéristiques θ_d. Sachant que selon les hyper-paramètres, la

3. L'UBM, ou modèle du monde, est un GMM représentant toutes les observations possibles.

taille du vecteur de caractéristiques peut varier, chaque représentation est projetée à travers un vocabulaire V composé de $|V|$ mots communs sélectionnés selon leur importance [Morchid et al., 2014e]. Ainsi, un ensemble de vecteurs homogènes de caractéristiques y_d est obtenu pour chaque document, en fonction des différents espaces de thèmes construits.

Le vecteur y_d représente alors une session – ou segment – du document d. Ici, (d, r) indique la représentation de d dans l'espace de thèmes r. Dans notre modèle, le super-vecteur de segment $M_{(d,r)}$ de d sachant un espace de thèmes r est modélisé par :

$$M_{(d,r)} = m + Tx_{(d,r)} \tag{2.6}$$

où m est le super-vecteur de l'UBM, T est la matrice de variabilité totale, et $x_{(d,r)}$ contient les coordonnées de la représentation du document dans l'espace réduit de variabilité totale appelé c-vecteur. La représentation c-vecteur est enfin normalisée avec l'algorithme Eigen Factor Radial [Bousquet et al., 2011]. De plus amples détails sur l'approche c-vecteur peuvent se trouver dans la thèse de M. Morchid [Morchid, 2014].

2.5.3 Expériences

Protocole expérimental

Comme pour les travaux présentés dans la partie 2.3 sur la comparaison de méthodes de représentation de documents, nous avons utilisé la tâche d'identification de la thématique principale de conversations téléphoniques humains-humains du corpus DECODA (voir partie 2.2). Cette tâche d'identification peut être vue comme une tâche de classification parmi 8 classes.

Dans le cadre de ce manuscrit, nous reportons les résultats que nous avons obtenus en compactant des espaces de thèmes au moyen de l'approche c-vecteur en faisant varier seulement l'hyper-paramètre du nombre de thèmes dans ces espaces. Le lecteur intéressé pourra se reporter aux travaux que nous avons présentés dans [Morchid et al., 2014a] pour des expériences plus complètes sur la variation des hyper-paramètres de l'approche LDA, à savoir α et β, mais également dans [Morchid et al., 2015a], qui présente de façon plus détaillée les approches et hyper-paramètres, et applique l'étude sur une autre tâche de classification (Reuters-21578). Nous avons alors proposé de faire varier ce nombre de thèmes de 5 à 504 avec un pas de 1, permettant d'obtenir 500 espaces de thèmes différents, composant alors les vues multiples des documents (ici, des conversations) qui y sont projetés. Nous avions choisi ce nombre de vues (500) en imaginant alors qu'il était assez important pour avoir une quantité suffisante de données et d'informations différentes contenues dans chacune d'entre elles, que la représentation compacte c-vecteur pourrait capter.

Un ensemble de mots a été extrait pour chacune des 8 thématiques du corpus DECODA. Tous les mots sélectionnés sont ensuite fusionnés, en gardant les mots uniques, afin d'obtenir un

vocabulaire V de taille 166 dans nos expériences. Comme précédemment, les espaces de thèmes ont été créés avec l'implémentation LDA Mallet [McCallum, 2002].

Enfin, une thématique est associée automatiquement pour chaque conversation (*i.e.* le *c*-vecteur) du corpus de test au moyen de la distance de Mahalanobis [Morchid et al., 2015a].

Résultats

Dans un premier temps, nous avons cherché à évaluer individuellement les 500 espaces de thèmes créés en faisant varier le nombre de thèmes, comme nous le rapportons dans la figure 2.5.

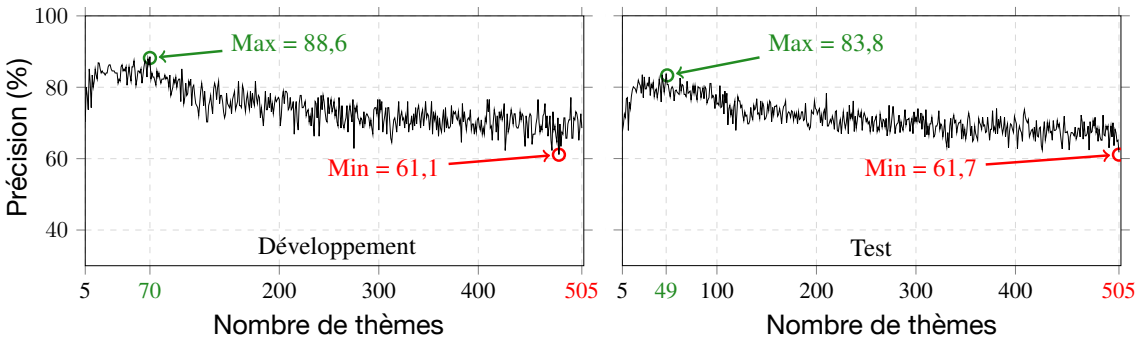

FIGURE 2.5 – Précisions (%) sur la tâche de classification en thématiques de conversations sur les corpus de développement et de test en faisant varier le nombre de thèmes de l'approche LDA. Les performances maximales (Max) et minimales (Min) sont fournies.

Comme nous l'avions vu dans la partie 2.4.2, les représentations obtenues au moyen de l'approche LDA ont tendance à donner des performances très fluctuantes sur notre tâche de classification lorsque l'on change, même un peu, les hyper-paramètres des espaces (des résultats identiques ont été obtenus en faisant varier les paramètres α et β [Morchid et al., 2015a]). De même, l'écart entre la performance maximale et minimale qu'il est possible d'obtenir apparaît très grand (sur le corpus de développement, cette différence est de 27,5 points de précision, sur le test, 22,1 points). Ces résultats ont ainsi justifié l'idée de l'approche *c*-vecteur proposée, de par la difficulté de trouver les hyper-paramètres optimaux pour un espace de thèmes.

Le tableau 2.3 présente les précisions obtenues sur les corpus de développement et de test avec l'approche *c*-vecteur en faisant varier la taille des *c*-vecteurs mais également la taille des gaussiennes du GMM-UBM. Nous avons alors pu observer un gain entre l'approche classique avec un seul espace de thèmes (figure 2.5) et les performances reportées avec l'approche *c*-vecteur dans le tableau 2.3. Ainsi, dans le corpus de développement, la performance maximale passe de 88,6 % à 92 %, et, si l'on se reporte aux performances maximales possibles dans le test, nous passons d'une précision de 83,8 % à 89,3 %. Nous avons également pu observer une variabilité

moindre au niveau des résultats en faisant varier les paramètres des *c*-vecteurs : l'écart entre la performance minimale et maximale atteint, avec les *c*-vecteurs, 7,4 points sur le développement (contre 27,5 points précédemment) et 3,7 points sur le test (contre 22,1 points). L'approche *c*-vecteur a donc clairement montré son avantage, profitant des informations provenant de vues multiples, en comparaison avec l'utilisation d'une vue simple sur laquelle les hyper-paramètres ont été optimisés. Outre un gain en performance, l'instabilité au niveau des résultats en faisant varier les hyper-paramètres n'est plus observée (ou tout du moins plus aussi fortement).

Taille des *c*-vecteurs	Développement				Test			
	Nombre de gaussiennes du GMM-UBM							
	64	*128*	*256*	*512*	*64*	*128*	*256*	*512*
60	89,7	89,7	90,3	90,3	88,0	89,6	88,4	88,0
80	87,4	89,7	89,1	90,3	87,5	89,0	89,3	87,8
100	84,6	89,1	**92,0**	89,7	88,0	91,7	**89,3**	87,5
120	86,3	86,9	89,1	90,9	87,8	88,7	86,2	89,3
140	85,1	86,3	89,1	88,6	85,6	86,5	86,2	87,8

Tableau 2.3 – Précisions (%) sur la tâche de classification en thématiques de conversations sur les corpus de développement et de test en faisant varier la taille des *c*-vecteurs et le nombre de gaussiennes du GMM-UBM.

2.6 Conclusion

Dans le cadre de ce chapitre, nous nous sommes intéressés à la robustesse de la représentation des documents, en nous concentrant sur une représentation par espaces de thèmes au moyen de l'approche LDA. Dans la première partie, nous avons alors comparé cette représentation avec une approche classique de représentation de documents, qui se focalise simplement au niveau du contenu textuel (ici, les mots) avec l'approche TF-IDF. Cette comparaison a été réalisée sur la tâche d'identification de la thématique principale de conversations entre humains issues du corpus DECODA. Plusieurs difficultés avaient été identifiées avant l'étude, et ont motivé le choix des espaces de thèmes : l'utilisation de transcriptions automatiques très bruitées (*i.e.* de nombreuses erreurs de transcription couplées à des conversations dans un registre de langue non standard au niveau linguistique) ainsi que la possible présence de thématiques secondaires multiples. Comme attendu, les performances obtenues au moyen de l'approche LDA, en projetant le document dans un espace de dimension de plus haut niveau, surpassent clairement celles atteintes avec l'approche TF-IDF [Morchid et al., 2014e,d]. Cette étude préliminaire a permis d'identifier une des limites des approches par espace de thèmes, à savoir le choix des hyper-paramètres des modèles.

Cette observation a conduit à la proposition de l'approche *c*-vecteur, inspirée de la représen-

tation *i*-vecteur en reconnaissance automatique du locuteur. L'idée des *c*-vecteurs est d'utiliser de très nombreux espaces de thèmes pour un même document en faisant varier les hyper-paramètres, permettant d'obtenir de multiples vues de celui-ci. Au final, un seul vecteur de caractéristiques du document est obtenu en compactant ces différentes vues, l'objectif étant de tirer profit de l'information pertinente de chaque vue (et donc d'enlever la variabilité nuisible inhérente à chacune d'entre elles). Les résultats atteints ont montré l'intérêt de cette approche, rendant la représentation plus robuste aux variations des hyper-paramètres, tout en améliorant les performances de classification [Morchid et al., 2014a, 2015a].

Ces différentes études concluent cette première partie sur la robustesse de la représentation de mots et de documents dans le cadre de documents écrits et parlés (transcriptions automatiques). Il semble assez clair que, lorsque les documents apparaissent très bruités, des approches projetant les mots dans des espaces de plus haut niveau, que ce soit par exemple au travers d'approches par plongement de mots ou d'espaces de thèmes, permettent d'avoir des représentations plus robustes. Dans les travaux de ce chapitre, se focalisant sur l'utilisation de transcriptions automatiques ayant de très nombreuses erreurs de reconnaissance, nous avons néanmoins pu observer des performances de classification très élevées, oscillant entre 80 et 90 % de précision alors même que les taux d'erreur-mot (WER) globaux pouvaient atteindre les 50 %. Clairement, il semble qu'il y ait un décalage entre la métrique du WER, qui est censée refléter la qualité d'une transcription automatique, et les performances que l'on peut obtenir en utilisant ces transcriptions en entrée d'autres tâches. Nous présentons alors, tout d'abord dans le chapitre 3, l'étude que nous avons faite entre taux d'erreur-mot et performance de classification automatique. Puis, nous nous intéressons aux travaux que nous avons menés sur la détection et caractérisation d'erreurs de transcription ciblées dans le chapitre 4. Nous terminons enfin cette seconde partie par les travaux entrepris sur l'évaluation des systèmes de transcription et le besoin de dépasser cette métrique du WER pour rendre compte de la qualité des transcriptions (chapitre 5).

Deuxième partie

Performance et évaluation en traitement du langage

Chapitre 3

Erreurs de transcription et impact sur les performances de classification

Sommaire

3.1 Introduction

Nous avons pu constater, dans les premiers chapitres de ce manuscrit regroupés au sein de la partie I, que le traitement de documents textuels n'est une tâche ni simple, ni résolue. Les erreurs orthographiques, grammaticales et/ou linguistiques sont autant de problèmes qui rendent l'exploitation automatique des textes difficile. Lorsque l'on doit traiter des transcriptions automatiques issues de documents parlés, ces problèmes apparaissent d'autant plus fortement que le langage *oral* revêt des spécificités qui lui sont propres, en particulier sur de la parole dite *spontanée* [Dufour, 2008] : disfluences dans le discours (pauses, troncations, répétitions, hésitations...), agrammaticalité, état émotionnel du locuteur... Face à ces problèmes, en plus des erreurs commises par les locuteurs eux-mêmes, les systèmes de reconnaissance automatique

de la parole (RAP) peuvent avoir des difficultés à traiter correctement certaines portions de parole, ce qui a pour effet de produire également des erreurs de transcription. Or, de nombreuses applications s'appuient, en entrée, sur ces transcriptions automatiques pour réaliser une tâche (indexation automatique, extraction d'information, classification de documents...). En partant du principe que des erreurs réalisées par les humains ou par les systèmes de RAP seront toujours présentes dans les transcriptions automatiques, nous avons montré, dans les chapitres précédents, que ces erreurs sur les mots constituent alors un bruit qu'il convient de maîtriser et de compenser, en proposant des approches de plus haut niveau dépassant l'utilisation directe du contenu textuel.

De façon assez classique, il semble assez naturel d'imaginer qu'en améliorant la qualité de la transcription automatique, les performances des systèmes les utilisant devraient également augmenter. Les systèmes de RAP s'appuient globalement sur la métrique du taux d'erreur-mot (WER) pour rendre compte de leur performance : plus ce taux est bas, plus le système de RAP est considéré comme performant. La plupart des systèmes de transcription sont ainsi améliorés en fonction de cette métrique. Améliorer la transcription automatique, comme nous le voyons dans le chapitre qui suit (chapitre 4), est relativement coûteux, puisqu'elle peut nécessiter la mise en place d'approches spécifiques et/ou de collecter de nouvelles données potentiellement annotées manuellement. Dans le cadre de la thèse de Mohamed Morchid, dont nous avons présenté certains de ses travaux sur la gestion de transcriptions très bruitées (voir chapitre 2), nous avons alors remarqué que les performances de classification étaient finalement assez élevées, avec des précisions maximales dépassant les 85 %, compte tenu des WER très élevés des transcriptions automatiques utilisées, se trouvant autour des 50 %. Ce constat est également valable pour l'approche classique TF-IDF, qui, bien qu'utilisant directement les mots, permet d'atteindre des performances avoisinant les 80 % de précision.

Ce constat entre le WER et les performances de classification est au coeur du travail de ce chapitre. Nous avons ainsi étudié le lien entre le WER sur les mots identifiés comme représentatifs et non-représentatifs par les deux approches de représentation des documents étudiées dans le chapitre 2, et les performances de classification automatique sur la tâche de détection de thématiques [Morchid et al., 2016b]. Nous avons choisi d'organiser ce chapitre en détaillant tout d'abord la métrique du taux d'erreur-mot, que nous avons déjà évoquée précédemment, mais dont nous n'avons jamais réellement discuté de sa pertinence (partie 3.2). Cela sera l'occasion d'évoquer ce que nous estimons comme avantages et faiblesses de la métrique. Le protocole expérimental est ensuite succinctement décrit dans la partie 3.3 puisqu'il s'appuie sur celui développé dans le chapitre 2. Enfin, nous proposons deux études dans cette partie, à savoir une première étude sur le lien entre le WER et le choix des mots représentatifs par les méthodes de représentation des mots (partie 3.4), et une seconde sur l'impact du choix des mots dans la construction des espaces de représentation dans la partie 3.5.

3.2 Le taux d'erreur-mot (WER)

3.2.1 Formule

Classiquement, les systèmes de RAP sont évalués en termes de taux d'erreur-mot (*Word Error Rate*, WER). Le WER prend en compte les erreurs de :

— *Substitution* : mot reconnu à la place d'un mot de la transcription de référence.

— *Insertion* : mot reconnu inséré par rapport à la transcription de référence.

— *Suppression* : mot de la référence oublié dans l'hypothèse fournie par le système de RAP.

Le WER s'exprime alors par la formule suivante :

$$WER = \frac{nombre\ de\ substitutions + nombre\ d'insertions + nombre\ de\ suppressions}{nombre\ de\ mots\ de\ la\ référence} \tag{3.1}$$

3.2.2 Avantages et limites

Le WER continue de faire consensus dans la communauté lorsque l'on parle de performance en RAP. De mon point de vue, les avantages principaux de cette métrique tiennent à plusieurs facteurs :

— *Simplicité.* Sa large diffusion doit beaucoup à sa simplicité : les trois catégories d'erreurs (insertion, suppression et substitution), puisque basiques, sont compréhensibles par tous. De plus, toute erreur possède le même poids, ce qui est évite la discussion sur une importance relative selon la nature de l'erreur.

— *Transcription manuelle suffisante.* Seule la transcription manuelle est nécessaire pour évaluer n'importe quel système de RAP sur ces données. La métrique n'est donc pas conçue pour traiter un système particulier, mais pour être appliquée de façon assez générale : elle ne requiert donc pas une annotation fine des erreurs (par exemple, par des experts linguistes), souvent coûteuse et difficile à obtenir.

— *Aucune subjectivité.* Ce dernier avantage rejoint le second point, en insistant sur le fait que l'évaluation laisse ici peu de place à la subjectivité. En effet, même s'il peut y avoir certaines difficultés à transcrire manuellement de la parole, en particulier la parole spontanée [Bazillon et al., 2008], une transcription reste plutôt objective, surtout si l'on compare avec d'autres tâches en TAL, comme par exemple dans le domaine du résumé automatique ou de la traduction automatique.

Bien entendu, comme toute métrique, celle-ci est sujette à discussion, et, comme nous le voyons plus particulièrement dans le chapitre 5, des travaux sont menés en vue de l'améliorer ou tout du moins de proposer des alternatives à l'évaluation des systèmes de RAP. Au final, les limites que j'ai souhaité mettre en lumière ici sont liées aux avantages listés précédemment :

— *Poids des erreurs.* Un poids unique des erreurs peut finalement apparaître comme une

limite de la métrique. En effet, il paraît assez peu réaliste que toutes les erreurs aient le même impact d'un point de vue applicatif. Même en considérant les 3 catégories d'erreurs de la métrique, il semble difficile d'imaginer qu'une erreur d'insertion (ajout d'un mot dans la transcription automatique) ait le même impact qu'un mot qui n'est pas présent dans la transcription automatique (suppression).

— *Nature des erreurs.* En lien avec la première limite, nous voulions mettre en avant le faible pouvoir informatif (et qualitatif) des catégories d'erreurs. Un des intérêts est de pouvoir identifier le type de l'erreur, en particulier sur les mots *importants* de la transcription d'un point de vue applicatif.

— *Évaluation orientée RAP.* La métrique est clairement orientée performance des systèmes de RAP, oubliant le côté *applicatif* des transcriptions. En effet, une transcription automatique, en tant que telle, n'a que peu d'intérêt : c'est la manière dont nous allons l'utiliser qui va être importante. Même si l'on considère la simple tâche d'utilisation directe de la transcription, comme par exemple pour le sous-titrage, il semble clair que tous les mots n'auront pas le même impact sur la compréhension des utilisateurs finaux. Alors que pour la RAP, un mot reste une entrée lexicale qu'il convient de reconnaître.

— *Impact humain ignoré.* Finalement, le fait de ne pas intégrer l'humain dans l'évaluation, et notamment comment les humains perçoivent les erreurs de transcription, amplifie le décalage entre des systèmes censés être conçus pour des personnes mais finalement évalués pour des machines. Ce dernier point est, en particulier, au centre de mes perspectives de recherche, que je développe dans la partie V.

Dans ce chapitre, nous n'avons pas cherché à remettre en question la métrique du WER, mais à comprendre si un lien pouvait exister entre ce taux d'erreur et les performances d'une tâche de classification utilisant ces transcriptions en entrée.

3.3 Protocole expérimental

Le protocole expérimental, pour étudier le lien entre WER et performance de classification, s'appuie sur celui proposé dans la partie 2.4.1 du chapitre 2. Pour résumer brièvement ici, nous comparons deux méthodes de représentation des documents, à savoir l'approche classique TF-IDF (voir partie 2.3.1), et l'approche par espaces de thèmes LDA (voir partie 2.3.2).

Les expériences sont menées sur le corpus DECODA, que nous avons décrit dans la partie 2.2. Il s'agit ici de trouver la thématique d'une conversation orale entre un appelant et un conseiller. Les transcriptions automatiques ont été obtenues au moyen du système de transcription Speeral du LIA [Linares et al., 2007]. Dans nos travaux, nous nous sommes appuyés simplement sur le contenu textuel (transcription) de la conversation et n'utilisons pas d'information issue du signal acoustique. La classification automatique (*i.e.* l'association d'une thématique à une conversation)

est alors réalisée en prenant en entrée d'un classifieur SVM une représentation du document, fournissant alors la classe hypothèse en sortie.

Nous avons proposé, dans ce travail, de continuer l'étude de différentes configurations entre les transcriptions manuelles et automatiques, mais en ne considérant pas ici, bien entendu, l'évaluation des transcriptions manuelles (non pertinent pour l'étude du WER). Nous n'aurons donc que les résultats sur les transcriptions automatiques bruitées (RAP), et trois conditions d'apprentissage : transcriptions manuelles propres ($REF \rightarrow RAP$), transcriptions automatiques bruitées ($RAP \rightarrow RAP$), et utilisation conjointe de conditions d'apprentissage propres et bruitées ($RAP + REF \rightarrow RAP$).

3.4 Étude sur les mots pertinents

Dans cette étude, nous avons voulu expérimenter la manière dont les approches de représentation des mots gèrent les erreurs de transcription automatique. Ici, nous savons que les représentations s'appuient sur des mots que nous considérons comme pertinents, *i.e.* qui participent à représenter et à discriminer les thématiques (classes) entre elles. Nous voulions donc savoir si ces représentations avaient tendance à s'appuyer principalement sur des mots correctement transcrits (ou de façon plus raisonnable, avec des WER plus faibles), ou au contraire, si aucun lien n'existait entre WER et mots pertinents, et que les approches prenaient indifféremment les mots transcrits, qu'ils soient corrects ou non. L'objectif a été de comparer les WER des n mots pertinents, des plus importants dans la représentation aux moins importants. Il a donc fallu tout d'abord ordonner la liste de ces mots, selon le protocole défini dans la partie suivante.

3.4.1 Extraction des mots pertinents

Approche TF-IDF

Pour l'approche TF-IDF, l'ordonnancement selon l'importance des mots a été simple, puisque la construction de l'approche permet de le faire sachant que les mots pertinents sont extraits selon leur importance dans toutes les thématiques (voir partie 2.3.1).

Approche LDA

Pour l'approche LDA, il a fallu proposer une solution pour extraire les mots considérés comme importants parmi les espaces de thèmes. Le score $s(w)$, utilisé alors pour retrouver ces

mots, est calculé comme suit [Morchid et al., 2016b] :

$$s(w) = P(w|m) = \int_z P(w|z)P(z|m)\,\mathrm{d}z = \sum_{z \in m} P(w|z)P(z|m) = \sum_{z \in m} V_z^w \times V_m^z = \langle \overrightarrow{V^w}, \overrightarrow{V^m} \rangle \tag{3.2}$$

où $\overrightarrow{V^w}$ est le vecteur de représentation d'un mot w dans tous les thèmes z de l'espace de thèmes m, $\overrightarrow{V^m}$ est le vecteur de représentation dans tous les thèmes z dans m, et $\langle \cdot, \cdot \rangle$ est le produit scalaire.

3.4.2 Analyse

Les figures 3.1 et 3.2 présentent les taux d'erreur-mot (WER) des n mots les plus pertinents obtenus pour chaque approche de représentation des mots sur le corpus d'apprentissage, à savoir respectivement les approches TF-IDF et LDA. Le WER global, calculé sur l'ensemble complet des mots considérés comme pertinents, est également fourni à titre indicatif. Le WER a été calculé de façon classique, en ne prenant donc ici que les n mots représentatifs dans le calcul.

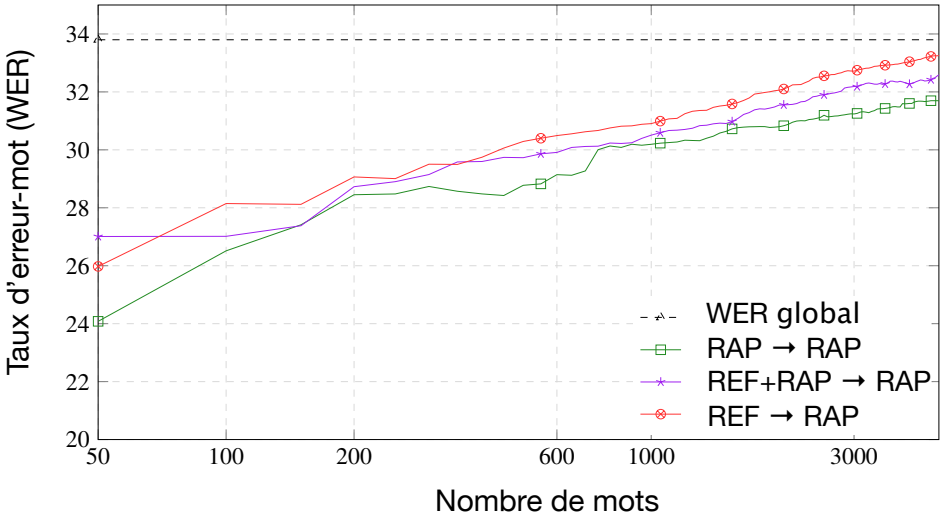

FIGURE 3.1 – Taux d'erreur-mot (WER) des n mots les plus pertinents avec la représentation TF-IDF. Différentes configurations d'apprentissage sont proposées et évaluées sur la transcription automatique (RAP). Le taux d'erreur-mot global (WER global) est fourni pour indication.

Nous avons tout d'abord observé que, peu importe la configuration d'apprentissage (RAP, $RAP + REF$, REF), les mots choisis comme les plus représentatifs par les méthodes de représentation, que ce soit par TF-IDF ou LDA, sont ceux qui ont des taux d'erreurs les plus faibles. En effet, plus nous augmentons la taille des mots pertinents (liste ordonnée), plus le WER aug-

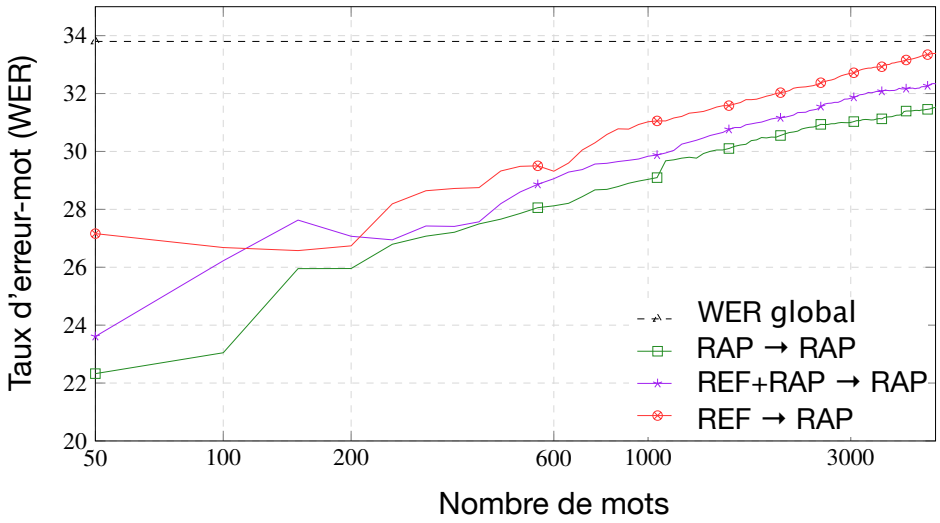

FIGURE 3.2 – Taux d'erreur-mot (WER) des n mots les plus pertinents avec la représentation par espace de thèmes (LDA). Différentes configurations d'apprentissage sont proposées et évaluées sur la transcription automatique (RAP). Le taux d'erreur-mot global (WER global) est fourni pour indication.

mente. En comparant chaque méthode de représentation des mots, nous avons pu observer que les WER obtenus par l'approche par espaces de thèmes (LDA) sont, dans tous les cas, légèrement inférieurs à ceux obtenus par l'approche utilisant les mots directement (TF-IDF). Cela signifie que les mots représentatifs choisis par l'approche LDA sont mieux transcrits par le système de RAP, ce qui pourrait expliquer les performances de classification plus élevées atteintes par la représentation des documents s'appuyant sur LDA.

Les courbes présentées dans la figure 3.2 ont montré que les mots les plus pertinents pour un espace de thèmes donné, sont ceux qui apparaissent comme les mieux transcrits par le système de RAP. Dans l'expérience suivante, nous avons voulu vérifier si, en apprenant des espaces de thèmes simplement avec les mots ayant les WER les plus faibles, cela pouvait améliorer ces représentations, et donc améliorer la précision de notre tâche d'identification de thématiques.

3.5 Sélection de mots pour l'apprentissage de modèles

Pour évaluer l'impact des erreurs de transcription sur la représentation des documents, la seconde expérience que nous avons menée s'intéresse à la construction des représentations sachant la qualité des mots utilisés pour construire les modèles (ici, en nous appuyant sur le WER). La figure 3.3 montre le nombre de mots uniques, ainsi que son pourcentage par rapport au nombre total, ayant un WER inférieur à une valeur donnée (w). Par exemple, il y a environ 4 000 mots

avec un WER inférieur à 50 %, ce qui correspond à 58 % du vocabulaire des transcriptions automatiques du corpus d'apprentissage.

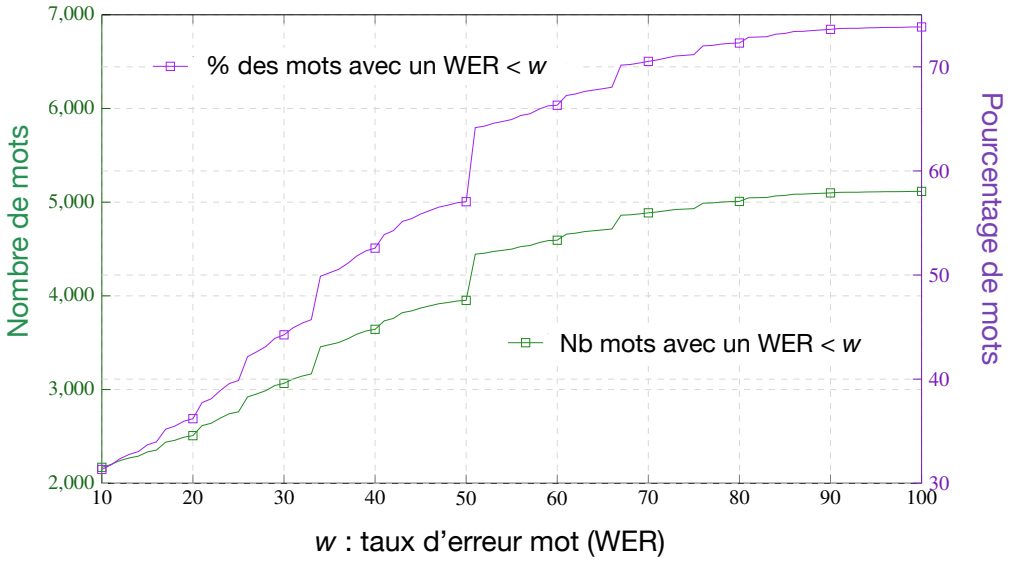

FIGURE 3.3 – Pourcentage et nombre de mots du corpus d'apprentissage DECODA ayant un taux d'erreur-mot (WER) inférieur à w.

3.5.1 Qualité des modèles

Nous avons ensuite proposé d'évaluer l'impact de la sélection de mots les mieux transcrits sur l'apprentissage d'espaces de thèmes. La figure 3.4 montre la perplexité moyenne ainsi que la log-vraisemblance moyenne toutes deux obtenues avec des espaces de thèmes de différentes tailles (60, 80 et 100 thèmes) en considérant les mots ayant un WER inférieur à w. Ces métriques permettent d'estimer la qualité d'un espace de thèmes [Morchid et al., 2016b]. Ces résultats sont cohérents avec ceux de la figure 3.3. Les courbes montrent que la qualité des espaces de thèmes stagne lorsque le vocabulaire contient des mots avec des WER supérieurs à 50 %, signifiant qu'il ne faudrait pas dépasser ce point. De 10 % à 50 %, la qualité de l'espace de thèmes baisse en fonction de l'augmentation du WER. La meilleure qualité est alors observée lorsque le WER est dans l'intervalle de 10 % à 25 %.

Ces mesures de qualité ne signifient cependant pas que des résultats comparables seront observés sur notre tâche de classification automatique. Dans notre contexte, il ne s'agit pas non plus de prendre très peu de mots bien transcrits, sous prétexte que la qualité paraît bonne : il est nécessaire d'avoir un espace de thèmes suffisamment grand (*i.e.* avec un nombre de mots suffisant) pour correctement représenter les différentes thématiques des documents. Il convient

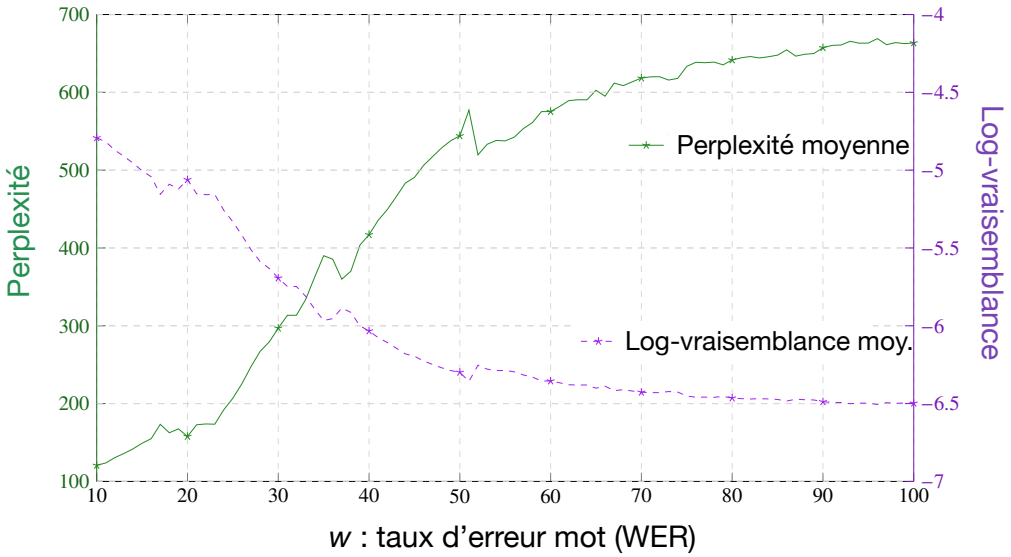

FIGURE 3.4 – Perplexité et log-vraisemblance moyennes des modèles LDA considérant les mots ayant un WER inférieur à w.

donc de trouver un compromis entre un modèle de qualité, utilisant des mots avec un WER bas, et les performances d'une tâche.

3.5.2 Performance de classification

La figure 3.5 présente les performances, en termes de précision (%), obtenues sur la tâche de classification de thématiques des conversations du corpus de test de DECODA au moyen de différentes représentations par espaces de thèmes (60, 80 et 100 classes) entraînées en utilisant les n mots pertinents. Les mots pertinents sont ici sélectionnés selon leur WER. Nous pouvons tout d'abord remarquer que la précision augmente avec une sélection de mots jusqu'à un WER d'environ 40 %. Ensuite, ces précisions stagnent et atteignent une précision maximale de 84,2 % pour l'espace contenant 80 thèmes. Nous pouvons également souligner qu'un espace de thèmes avec un petit nombre de thèmes (60) est plus robuste au WER que les espaces avec un grand nombre de classes (80 ou 100). En comparant les figures 3.4 et 3.5, nous avons alors pu observer que la qualité estimée des modèles n'a pas le même point de fonctionnement que sur notre tâche applicative : comme nous l'avons supposé, nos espaces de thèmes ont besoin d'avoir suffisamment de mots, même fortement bruités (*i.e.* mal transcrits) pour obtenir les meilleures performances. Notons enfin que, lorsque le WER devient trop important (> 60 %), les résultats apparaissent plutôt instables, peu importe le nombre de thèmes utilisés pour construire les espaces de représentation.

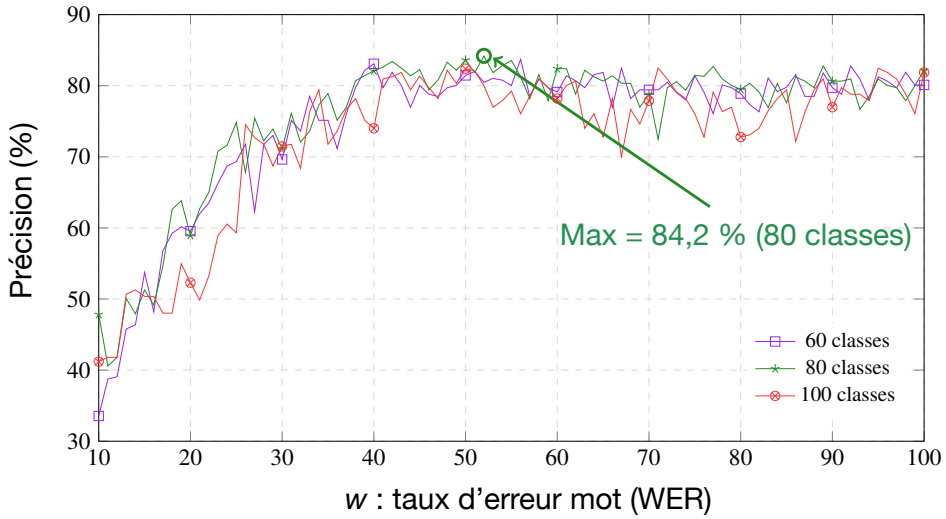

FIGURE 3.5 – Précision de classification (%) sur le corpus de test DECODA au moyen de différentes représentations par espaces de thèmes (60, 80 et 100 classes) entraînées en utilisant les n mots pertinents sélectionnés selon leur taux d'erreur-mot (WER).

3.6 Conclusion

Au sein de ce chapitre, nous avons proposé, durant la thèse de M. Morchid, une étude cherchant le lien entre qualité de représentation des documents et taux d'erreur-mot des transcriptions automatiques [Morchid et al., 2016b]. Nous sommes partis du constat que, bien que les transcriptions étaient fortement bruitées (WER supérieurs à 50 %), les performances de classification sur la tâche de détection de la thématique de conversations apparaissaient très bonnes. Cela ne pouvait pas être simplement dû aux représentations de plus haut niveau proposées, telles que l'approche par espaces de thèmes (LDA), puisque même l'approche classique TF-IDF dépassait les 80 % de précision, comme nous l'avions observé dans le chapitre 2.

La première étude a consisté à analyser les taux d'erreur des mots pertinents utilisés dans deux représentations de documents (TF-IDF et LDA). Nous avons ainsi observé que les mots considérés comme les plus pertinents étaient ceux ayant un WER plus faible, comparativement aux mots les moins représentatifs (*i.e.* les mots ayant le score de pertinence le plus bas dans le vocabulaire choisi), et ce, peu importe l'approche de représentation des mots. Nous avons également pu voir que l'approche LDA avait tendance à choisir comme représentatifs les mots ayant des WER plus bas que ceux choisis par l'approche TF-IDF.

Dans la seconde partie de ce travail, nous avons cherché à étudier le lien entre qualité des représentations par espaces de thèmes et WER des mots utilisés pour les estimer. Nous nous sommes aperçus que les mots ayant les WER les plus faibles permettaient d'obtenir des modèles

LDA de meilleure qualité en termes de perplexité et log-vraisemblance. Nous avons enfin vérifié cette observation sur notre tâche de classification, et, même si le point de fonctionnement entre qualité des modèles et performance de classification diffère, il semble qu'un lien existe entre WER et performance sur une tâche appliquée.

Au final, même si nous avons pu voir l'influence entre WER et performance des modèles utilisant des transcriptions automatiques, il semble que le taux d'erreur-mot global ne soit pas un indicateur fiable sur la qualité des transcriptions. Une étude plus fine au niveau des mots semble indispensable (ici, les mots considérés comme pertinents par les méthodes de représentation) pour mieux rendre compte de la qualité des transcriptions automatiques. Le chapitre suivant est au coeur de ce problème d'erreurs de transcription et d'évaluation, où nous décrivons un travail réalisé sur la catégorisation et détection de régions d'erreurs spécifiques dans les transcriptions automatiques.

Chapitre 4

Caractérisation et détection d'erreurs dans les systèmes de transcription

Sommaire

4.1 Introduction

Les systèmes de reconnaissance automatique de la parole (RAP) font inévitablement des erreurs. Comme nous avons pu le voir dans les chapitres 2 et 3 principalement, même s'il est possible de compenser ces erreurs de transcription, elles peuvent néanmoins avoir un impact négatif sur les systèmes les utilisant en entrée pour réaliser une tâche (par exemple, classification de documents, extraction d'information, reconnaissance de noms de personne...). De plus, certaines applications utilisent directement les sorties des systèmes de RAP, comme cela est le cas pour le sous-titrage automatique. Il est donc nécessaire de fournir une transcription contenant le

moins d'erreurs possibles. D'un point de vue applicatif, nous pouvons identifier deux stratégies principales face à ces erreurs :

1. Masquer ces erreurs afin, selon l'objectif suivi, d'éviter de les afficher ou de les utiliser par les applications s'appuyant sur la transcription.
2. Proposer des solutions pour corriger ces erreurs, soit directement dans les systèmes de RAP, soit dans les transcriptions automatiques en sortie des systèmes.

Dans les deux cas, une détection automatique des erreurs de transcription est utile. Nous avons alors proposé une approche permettant d'identifier les erreurs en sortie des systèmes de RAP tout en les caractérisant afin de fournir une information sur la nature de l'erreur [Dufour et al., 2012a]. Nous voulions aller plus loin qu'une simple classification binaire d'un mot en *correct* / *erroné*, puisque : 1) cette information peut être importante d'un point de vue analytique, pour comprendre le fonctionnement du système et avoir un rapport plus informatif sur les transcriptions et leur performance ; et 2) nous partons de l'hypothèse que des erreurs de natures différentes ont des comportements différents, et qu'il est utile de pouvoir les caractériser pour mettre en place des stratégies spécifiques en vue de leur traitement automatique futur (correction, utilisation dans des applications...). Ces travaux ont été menés durant mon post-doctorat à Orange Labs, en collaboration avec Géraldine Damnati et Delphine Charlet, et dans le cadre du projet ANR PERCOL et du défi REPERE (voir partie 9.1.2).

Plusieurs contraintes ont guidé le choix de détecter et caractériser les erreurs de transcription dans une approche *a posteriori* :

— Un accès seul aux transcriptions en sortie du système de RAP (*i.e.* impossibilité de modifier le système ou les modèles et/ou de récupérer des informations issues du fonctionnement interne du système de RAP).
— Traitement de mots relativement rares (ici, noms de personne) apparaissant épisodiquement au court du temps, ce que les systèmes de RAP génériques traitent mal.

Dans ce chapitre, nous nous intéressons tout d'abord à l'approche que nous avons proposée pour détecter les erreurs de transcription (partie 4.2), en ne les considérant pas de manière isolée mais comme une séquence. Nous proposons ensuite, dans la partie 4.3, une possibilité de caractérisation de ces régions d'erreurs, orientée pour le projet PERCOL.

4.2 Détection des régions d'erreurs

4.2.1 Erreurs isolées *vs.* régions d'erreurs

Contexte

Une des façons les plus simples pour détecter *a posteriori* les erreurs de transcription est d'utiliser les mesures de confiance associées aux hypothèses de mots [Mauclair et al., 2006].

L'application d'un seuil sur ce score permet alors de déterminer automatiquement si ce mot est correct ou non, ce seuil pouvant être réglé par rapport à un point de fonctionnement selon l'application visée. Une des limites concernant la mesure de la performance, et donc de la qualité, des mesures de confiance est qu'elles sont généralement évaluées sur leur capacité à déterminer si un mot est correct. Cependant, pour les tâches où les taux d'erreur-mot (WER) sont faibles, cette tâche de classification binaire en *correct* / *incorrect* traite donc des données fortement déséquilibrées : l'évaluation centrée sur la classe majoritaire (mots corrects) a tendance à masquer la capacité du classifieur à gérer la classe minoritaire (mots erronés). Ceci est clairement notre cas dans ce travail, où nous avons abordé la question de l'évaluation des systèmes de détection des erreurs, cette notion d'évaluation étant poursuivie dans d'autres travaux de recherche (voir chapitre 5).

J'ai déjà pu observer, pendant mes travaux de thèse sur la parole spontanée [Dufour, 2008], que les erreurs des systèmes de RAP n'apparaissent pas forcément isolées, mais possiblement dans des *groupes* d'erreurs. Cela est cohérent avec la manière dont sont construits les systèmes de RAP, qui travaillent sur des séquences (et donc l'historique des mots) pour prendre une décision. Ainsi, certains phénomènes génèrent plusieurs erreurs consécutives, que nous avons appelées dans nos travaux *régions d'erreur*. Cela peut se produire pour de multiples raisons, comme un mot long hors vocabulaire [1], ayant tendance à être segmenté en plusieurs mots courts hypothèses, une substitution qui se propage aux mots adjacents en raison du modèle de langage [Parada et al., 2010], ou encore de mauvaises conditions acoustiques pour lesquelles le décodeur fournit une sortie complètement erronée. Dans [Duta et al., 2006], les auteurs ont analysé les erreurs de transcription sur de la parole spontanée et ont conclu que les deux tiers des erreurs apparaissent dans un groupe ($n \geq 2$ erreurs consécutives). Pour toutes ces raisons, nous avons alors choisi de considérer la détection d'erreurs comme un problème de *région* et non de manière isolée, comme cela est fait habituellement. Nous avons tout d'abord proposé une étude permettant d'appuyer ce choix sur le corpus présenté dans la sous-partie suivante.

4.2.2 Corpus d'émissions télévisées

Description des données

Afin de mener à bien nos expériences, d'étudier les erreurs de transcription et de proposer des approches pour leur détection et caractérisation, nous nous sommes appuyés sur un corpus d'émissions télévisées en français transcrit manuellement et dont les noms de personne y ont également été annotés. Ce corpus, réalisé par Orange Labs et que nous nommerons JT, contient 8 journaux télévisés (informations, interviews, reportages...) issus de 7 chaînes de télévision et

1. Les systèmes de RAP classiques fonctionnant en vocabulaire fermé, un mot hors vocabulaire est un mot présent dans la transcription de référence (manuelle) mais qui n'est pas présent dans le dictionnaire du système de RAP. Cela conduit donc inévitablement à une, ou plusieurs, erreurs de transcription.

collectés entre octobre 2008 et janvier 2009. Ce corpus a été découpé en deux sous-ensembles : 24 émissions sont utilisées pour l'apprentissage (*JT_train*) et 14 émissions pour le test (*JT_test*). Les transcriptions automatiques ont été réalisées au moyen du système de RAP VoxSigma v3.5 de Vocapia Research [Gauvain et al., 2002]. Le dictionnaire contenait environ 65 000 mots, incluant environ 22 000 noms propres. Les corpus sont décrits dans le tableau 4.1, incluant la durée totale, le nombre de mots de référence et les WER associés, ainsi que la taille moyenne des régions d'erreurs.

	JT_train	**JT_test**
Durée	7h45	6h15
Nb mots (taux d'erreur-mots)	84 146 (15,9 %)	70 538 (18,0 %)
Nb régions d'erreurs (taille moyenne)	5 529 (1,8)	4 908 (1,8)

Tableau 4.1 – Description du corpus d'émissions télévisées d'Orange Labs.

La figure 4.1 montre la répartition des erreurs de transcription consécutives calculées en nombre de mots ou en nombre de régions (par exemple, deux erreurs consécutives comptent pour une région). Notons que la figure ne concerne que l'analyse de JT_train. Nous pouvons voir que plus de 25 % des erreurs sont des erreurs isolées (*i.e.* régions de taille 1) et que 55 % des régions n'ont qu'une seule erreur ($n = 1$). En nous focalisant sur les erreurs multiples ($n >= 2$), représentant 75 % des mots mal transcrits, la longueur moyenne de ces régions est de 2,2 mots (globalement, 1,8 comme vu dans le tableau 4.1). Ce résultat conforte l'idée qu'il est intéressant de considérer les erreurs en régions.

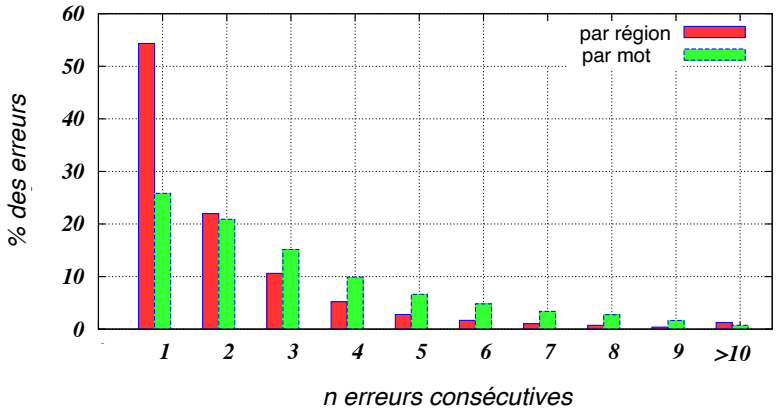

FIGURE 4.1 – Répartition (par mot et par région) des erreurs de transcription en fonction de la longueur des séquences d'erreurs.

Enfin, une mesure de confiance, fournie par le système de transcription, est associée à chaque mot transcrit. Ce sont ici des probabilités *a posteriori* calculées à partir des graphes de mots.

Les scores d'entropie croisée normalisée (NCE) sont de 0,36 sur le corpus JT complet.

4.2.3 Étiquetage de séquences

Nous avons choisi de voir la détection des régions d'erreurs comme un problème d'étiquetage de séquences [Dufour et al., 2012a]. Nous avons alors proposé une segmentation des transcriptions en *région correcte* et *région erronée*. Quatre approches différentes pour la segmentation en régions ont été proposées :

- *Base*. Utilisation d'un seuil θb sur les mesures de confiance *a posteriori* fournies par le système de RAP. Les mots consécutifs détectés comme erreur possible ($< \theta b$) sont alors considérés comme une région d'erreurs.
- *Automate*. L'application d'un seuil unique sur les mesures de confiance peut ne pas être suffisant puisque les erreurs consécutives ne sont pas toutes associées à une mesure de confiance basse. Afin d'assouplir cette contrainte, nous avons introduit deux seuils au lieu d'un dans le cadre d'un automate à états finis à deux états (voir figure 4.2). Le premier seuil θ_{err} permet de passer de l'état *Correct* à l'état *Erreur*, et inversement avec le seuil θ_{cor}. L'automate est utilisé pour chaque segment dans les deux sens de lecture (de droite à gauche et vice versa).
- *CRF isolé*. Utilisation des champs conditionnels aléatoires (CRF) [Lafferty et al., 2001], une méthode statistique permettant de segmenter et d'étiqueter des séquences de données. L'avantage ici est que cette méthode utilise des informations multiples sur les mots environnants : bi-grammes de mots, classes morpho-syntaxiques [2], mesures de confiance, durée des mots courant, précédent et suivant. La mise en œuvre repose sur un formalisme *UIO* (*Unique* pour les erreurs isolées, *Inside* pour les $n > 1$ mots dans les régions d'erreurs et *Outside* pour les mots corrects).
- *CRF intégré*. Comme l'approche CRF peut segmenter mais également étiqueter des séquences, nous avons proposé d'utiliser cette méthode pour directement segmenter et étiqueter les régions d'erreurs selon les catégories définies dans la partie 4.3.2, pour voir si cela influence les performances de détection des régions d'erreurs. Au formalisme *UIO* est ajoutée la classe considérée.

4.2.4 Classification binaire

Nous avons proposé d'évaluer la tâche de détection des régions d'erreurs au moyen de la métrique classique rappel/précision. Nous verrons dans la partie suivante, en particulier pour évaluer le processus complet de détection et de caractérisation des erreurs, que cette métrique est trop limitée lorsque l'évaluation se complexifie. Détecter précisément les régions d'erreurs

2. Lia_tagg : `http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html`

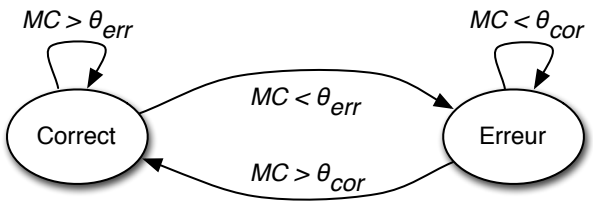

FIGURE 4.2 – Détection des régions d'erreurs au moyen d'un automate à deux seuils (θ_{err} / θ_{cor}) sur les mesures de confiance (MC) des mots transcrits.

étant une tâche difficile, nous avons choisi d'assouplir la détection des régions en considérant comme correcte une région d'erreurs dont les frontières sont erronées à deux mots près.

La figure 4.3 présente les performances obtenues, en rappel et précision, sur le corpus JT_train pour la détection des régions d'erreurs en faisant varier le seuil de décision des 4 approches proposées. Les méthodes *Base* et *Automate* ont permis d'atteindre des taux de précision plus élevés que les méthodes à base de CRF lorsque le rappel est très haut, mais, à l'inverse, les méthodes à base de CRF sont plus précises lorsque le rappel est faible. Cette plus faible précision pour les CRF s'explique par un nombre trop grand d'hypothèses de détection conduisant à de nombreuses insertions ainsi qu'à des régions trop longues. Ces difficultés ont été mieux gérées avec l'utilisation de la mesure de confiance seule. Notons également que le comportement des deux approches à base de CRF différait au niveau de l'évolution du taux de rappel : l'approche *CRF Isolé* ne dépassait pas les 35 % en rappel alors que la variation du seuil de décision permettait à l'approche *CRF Intégré* d'approcher les 60 %.

4.3 Caractérisation des régions d'erreurs

4.3.1 Contexte

Au-delà de la détection d'erreurs, nous voulions dépasser le cadre classique d'analyse des erreurs en caractérisant automatiquement leur nature. Par exemple, des études sur les mots hors vocabulaire [Woodland et al., 2000] ont montré que leur comportement diffère des autres erreurs de transcription. Être capable de mieux connaître les erreurs permettrait de décider d'en ignorer certaines, ou, au contraire, d'appliquer des stratégies adaptées à leur nature pour les corriger. Dans le cadre de notre travail, nous n'avons pas considéré les mots hors vocabulaire comme une catégorie d'erreurs en tant que telle, car spécifique aux systèmes de RAP, mais plutôt comme des classes d'erreurs importantes du point de vue de notre application.

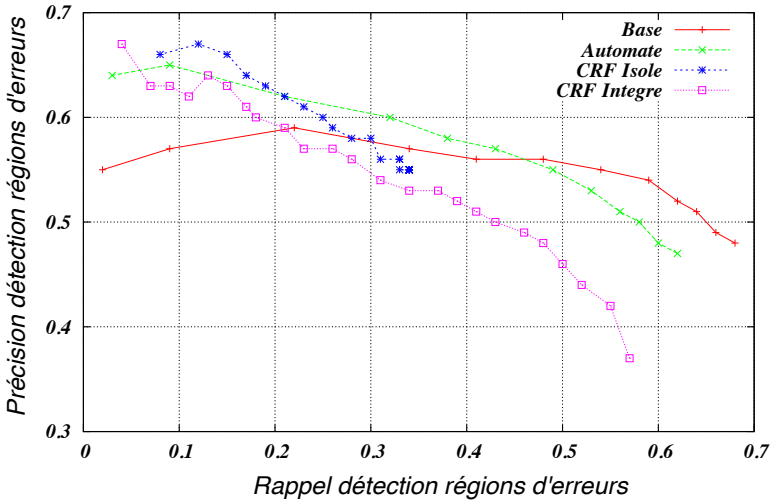

FIGURE 4.3 – Performances (rappel et précision) pour la détection des régions d'erreurs sur le corpus JT_train en faisant varier le seuil de décision des approches étudiées.

4.3.2 Définition des classes d'erreurs

Nous avons alors proposé de définir un ensemble de classes pour les erreurs de transcription selon leur nature, sachant que ce travail prenait place dans le cadre du défi REPERE, avec pour objectif de travailler sur les noms de personne. Nous avons alors défini 4 classes, chaque région d'erreurs étant associée à une d'entre elles (annotation manuelle), à savoir : *Nom de personne* (NP), *Autre nom propre* (ANP), *Homophone* (H), et enfin *Autre* (A) pour toutes les autres erreurs ne rentrant pas dans les catégories précédentes. La figure 4.4 présente la répartition des régions d'erreurs en fonction de leur longueur pour les corpus *JT*. La répartition des erreurs est réalisée par région d'erreurs. Nous avons notamment pu observer des comportements différents en termes de taille de régions d'erreurs, où les classes *NP* et *ANP* génèrent des régions d'erreurs de tailles plus grandes que les erreurs dues à des *H* ou *A*, donc avec un impact négatif potentiellement plus important.

4.3.3 Approches

Une simple catégorisation des erreurs de transcription ne nous a pas semblé une tâche réaliste : les erreurs des systèmes de transcription ne sont pas connues, il n'est pas pertinent de travailler directement sur les régions d'erreurs annotées manuellement. Nous avons donc proposé de classifier les régions d'erreurs obtenues automatiquement avec les approches de segmentation présentées dans la partie 4.2.3.

Les approches *Base*, *Automate* et *CRF isolé* n'étant que de simples approches de segmenta-

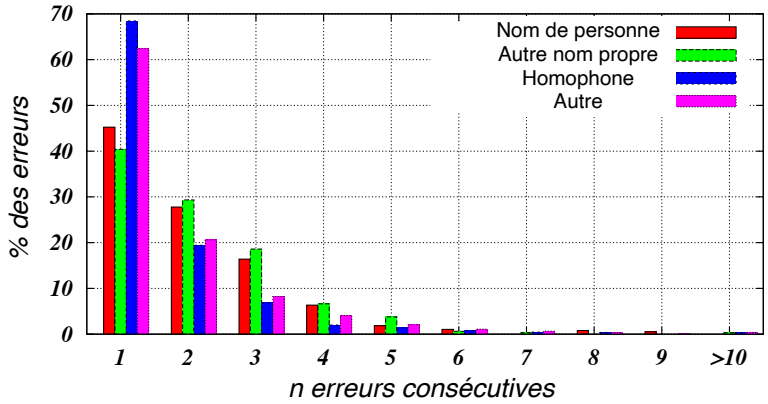

FIGURE 4.4 – Répartition des régions d'erreurs selon leur classe (Nom de personne, Autre nom propre, Homophone et Autre).

tion, nous avons utilisé un classifieur fondé sur l'algorithme AdaBoost (outil Icsiboost [3]), avec lequel nous avons associé automatiquement une des 4 classes aux régions d'erreurs préalablement détectées automatiquement. Plusieurs caractéristiques linguistiques ont été utilisées : les mots des régions (bi-grammes), l'étiquetage grammatical et regroupement en syntagmes (tri-grammes), le nombre des mots de la région, quadri-grammes sur les cinq mots précédents, la durée et la moyenne des mesures de confiance de chaque tour de parole, et enfin le nombre de syllabes par mot. Nous n'avons rien modifié à l'approche *CRF intégré* puisque celle-ci permettait la segmentation et l'étiquetage dans un même processus.

4.3.4 Évaluation globale

Il parait assez simple d'évaluer la tâche de classification automatique seule, en ne se concentrant que sur les régions d'erreurs correctes. Mais une des questions importantes qui s'est posée pendant ce travail a concerné la façon d'évaluer, finalement, la tâche conjointe de caractérisation et segmentation des régions d'erreurs afin de rendre compte d'une performance globale. En effet, comme nous l'avons vu pour la tâche de segmentation des régions d'erreurs, nous avons proposé une évaluation en rappel/précision. Nous avons cependant bien conscience du caractère strict de cette approche : nous avions déjà considéré une *tolérance* sur les frontières des régions (voir partie 4.2.4).

Nous avons alors proposé d'utiliser la métrique Slot Error-Rate (SER) [Makhoul et al., 1999], plutôt utilisée pour évaluer les systèmes de détection des entités nommées. Elle possédait l'avantage de prendre en compte de nombreuses combinaisons d'erreurs potentielles contenues

3. `http://code.google.com/p/icsiboost`

dans notre double problématique de détection et caractérisation de régions d'erreurs :

$$SER = \frac{D + I + S_{all} + 0,5 * (S_{cla} + S_{reg})}{Nombre\ total\ des\ régions\ d'erreurs\ de\ référence} \tag{4.1}$$

où D est le nombre de régions non détectées, I le nombre de régions insérées, S_{cla} le nombre de régions d'erreurs correctement détectées mais mal classées, S_{reg} le nombre de régions d'erreurs dont les frontières ont été mal détectées mais assignées avec la classe d'erreur correcte, et S_{all} le nombre de régions d'erreurs dont les frontières ont été mal détectées et assignées avec une classe d'erreur incorrecte. En fonction de l'application visée, toutes les erreurs n'ont pas le même impact sur le score SER. Ici, les erreurs S_{cla} et S_{reg} ont un coût de 0,5.

4.3.5 Résultats obtenus

Les meilleurs résultats ont été obtenus au moyen de l'approche par *Fusion* [Dufour et al., 2012b]. Cette solution consistait à combiner toutes les propositions en fusionnant les régions d'erreurs au moyen de l'opérateur *OU*. A titre d'information, le SER obtenu par la méthode *Fusion* est de 81,6 % contre 86,7 % pour *Base*. Le tableau 4.2 résume les résultats obtenus sur les tâches de détection des régions d'erreurs et de leur caractérisation (catégorisation), soit pour chacune des tâches, soit de façon combinée avec le SER. Nous avons clairement pu apprécier la difficulté de la tâche au travers de ces résultats. Mais nous avons pu nous confronter aux difficultés concernant l'évaluation de tâches de TAL, en adoptant une métrique existante sur un problème aux caractéristiques similaires.

Détection		**Caractérisation**	**Global**
Rappel	**Précision**	**% classif. correcte**	**SER**
42,2	57,0	78,4	81,6

Tableau 4.2 – Performance en détection seule (Détection), catégorisation seule (Catégorisation), catégorisation et détection combinées (Global) de régions d'erreurs avec la méthode *Fusion* sur les données JT_test.

4.4 Conclusion

Dans ce chapitre, nous avons proposé des solutions pour détecter et caractériser automatiquement les erreurs faites par les systèmes de reconnaissance automatique de la parole (RAP). L'originalité de ces travaux se trouve à la fois dans la manière de considérer les erreurs, ici au travers du concept de régions d'erreurs (*i.e.* ensemble d'erreurs apparaissant consécutivement), mais également dans l'idée de les caractériser non pas d'un point de vue *système* mais plutôt au niveau de la nature de l'erreur (nom de personne, homophone...). Cela nous a conduit à réaliser

une étude qualitative sur les erreurs de transcription qui a conforté l'idée de ne pas traiter une erreur de manière isolée mais en groupe, puisque les système de RAP doivent traiter des séquences : les erreurs ont tendance à se répercuter aux autres mots dans un même segment. Nous avons également montré, au travers de 4 classes que nous avons choisi d'étudier, que selon la nature de l'erreur, l'impact sur les mots alentours n'est pas le même. Par exemple, les erreurs sur les noms propres ont tendance à faire apparaître des régions d'erreurs plus grandes, au contraire d'erreurs liées à l'homophonie. Ces travaux sont issus de mon post-doctorat à Orange Labs et ont naturellement pris leur place dans le cadre du Projet ANR PERCOL, dans lequel j'ai pu continuer à travailler dès mon recrutement en tant que maître de conférences au LIA, puisque le LIA et Orange Labs faisaient tous deux partie du même consortium (voir partie 9.1.2).

Nous avons alors tout d'abord proposé plusieurs approches pour la segmentation en régions d'erreurs, incluant l'utilisation des mesures de confiance, un automate à deux états, et une approche s'appuyant sur les CRF. La classification de ces régions d'erreurs selon une des classes définies (Nom de personne, Autre nom propre, Homophone, Autre) a pu se faire au moyen d'un classifieur SVM, ou avec une approche intégrant la segmentation et l'attribution d'une classe dans le même processus (CRF Intégré).

Outre ces propositions, nous avons surtout réfléchi à la façon d'évaluer cette double tâche, qui combine à la fois un problème de segmentation (détection des régions) et de classification (attribution d'une classe). Nous avons proposé d'appliquer la mesure SER, qui a montré son efficacité en détection d'entités nommées, présentant finalement une problématique d'évaluation très proche. Au cours de ces chapitres, nous avons pu nous apercevoir que l'évaluation est au coeur des réflexions que j'ai pu mener ces dernières années. Le chapitre suivant continue sur la problématique des erreurs de transcription et de leur évaluation, que nous considérons ici du point de vue de leur correction d'une part, et de la manière de rendre compte de la performance des systèmes d'autre part.

Chapitre 5

Correction des erreurs et évaluation des systèmes de transcription

Sommaire

5.1 Introduction

Comme nous avons pu le constater dans les différents chapitres de ce manuscrit, l'évaluation joue un rôle prédominant puisqu'elle permet de rendre compte de la performance d'un système, et souvent de le justifier. Ces métriques sont forcément imparfaites : elles ne peuvent prendre en considération tous les aspects d'une tâche et des applications qui en découlent. Elles sont alors souvent sujettes à discussion et à critique dans les différentes communautés scientifiques. Cela est particulièrement vrai lors d'apparition de nouvelles tâches, où aucun consensus n'a pu avoir lieu, ce que nous développerons dans les chapitres de la partie III lorsque nous parlerons de nos travaux menés dans un contexte interdisciplinaire. Les métriques *historiques* ne sont

cependant pas exemptes de toute critique, d'autant plus que ce choix revêt un enjeu hautement stratégique : les systèmes de TAL étant de plus en plus *industrialisés*, cela devient également un enjeu financier à la fois pour les entreprises, mais également pour les laboratoires de recherche dont les financements peuvent dépendre de la *qualité* de leur système au niveau international.

En TAL, les mesures faisant généralement consensus sont celles qui sont les plus faciles à appliquer largement et qui ne nécessitent pas d'intervention humaine supplémentaire, autre qu'une annotation de référence, pour évaluer un nouveau système. La métrique du taux d'erreur-mot (WER), à laquelle il est impossible d'échapper lors d'évaluation en reconnaissance automatique de la parole (RAP), suit clairement cette idée (voir notre discussion dans la partie 3.2).

Indépendamment de la métrique considérée, l'objectif clairement identifié des systèmes automatiques est que ceux-ci ne fassent aucune erreur. Les travaux en correction d'erreurs des systèmes de RAP suivent généralement deux approches :

— Améliorer le système lui-même, en proposant de nouvelles architectures et/ou d'améliorer la robustesse des modèles, avec par exemple des adaptations au domaine linguistique ou aux conditions acoustiques (conditions d'enregistrement, bruits...).
— Corriger *a posteriori* les erreurs des transcriptions.

Bien entendu, malgré l'objectif affiché du *zéro erreur*, nous ne pouvons prétendre y arriver actuellement en RAP, malgré des avancées certaines, en particulier avec l'avénement des architectures neuronales. Il nous a donc semblé important de joindre ici les travaux que nous avons pu réaliser à la fois sur la correction des erreurs des systèmes de RAP et des réflexions autour de l'évaluation, en particulier devant la multitude de cadres applicatifs utilisant les transcriptions automatiques, comme nous avons pu le montrer dans les chapitres précédents.

Dans ce chapitre, nous présentons tout d'abord, dans la partie 5.2, les travaux sur la correction d'erreurs spécifiques [Dufour et al., 2012c] dans le cadre applicatif du projet ANR PERCOL pour la détection de noms de personne. Ceci est également une problématique que j'ai pu développer durant ma thèse sur la détection et correction d'erreurs liées à l'homophonie [Dufour and Estève, 2008]. Il reste que la correction de ces erreurs spécifiques, peu nombreuses au regard d'une transcription globale, a un impact faible sur le WER. Ces erreurs sont donc souvent peu étudiées, alors même que leur intérêt applicatif peut être important. Nous présentons ensuite, dans la partie 5.3, des travaux auxquels j'ai pu participer pendant la thèse de Salima Mdhaffar, dirigée par Yannick Estève, sur l'adaptation des modèles de langage, qui a proposé des métriques originales pour rendre compte des performances de systèmes de RAP dans un cadre applicatif [Mdhaffar et al., 2019]. Ce chapitre couvre finalement les deux stratégies de correction possible des erreurs, soit *a posteriori*, soit au niveau du système directement.

5.2 Correction *a posteriori* des erreurs

5.2.1 Approche générale

Ce travail a été réalisé dans le cadre du défi REPERE visant à reconnaître des noms de personne dans les émissions audiovisuelles. Nous avions donc un intérêt applicatif réel à corriger les noms de personne dans les transcriptions automatiques puisque cette information apparaissait essentielle dans le contexte du projet. De façon plus générale, plusieurs raisons ont motivé notre choix de traiter les erreurs sur ces mots spécifiques :

— Les noms de personne font partie des mots les plus difficiles à traiter par des systèmes de RAP, de par le fait qu'ils peuvent être rares et/ou n'apparaître qu'à des périodes spécifiques, en particulier dans les émissions d'actualité. Ceux-ci ont donc plus tendance à être hors vocabulaire dans le système de RAP [Sheikh et al., 2015], de nombreuses études s'étant intéressées à ce phénomène, comme dans le cadre du projet ANR ContNomina auquel j'ai eu l'occasion de participer.

— Comme nous l'avions vu dans la partie 4.3.2, les erreurs sur les noms de personne ont tendance à générer des régions d'erreurs plus grandes que les autres erreurs de manière générale : les corriger pourrait avoir un impact positif plus grand sur le WER.

— D'un point de vue applicatif, ils font partie des mots importants à bien transcrire, comme par exemple dans le contexte d'indexation de documents.

Nous avons traité ici la reconnaissance des noms de personne comme un cas particulier de la reconnaissance d'entités nommées. Nous avons alors adopté une stratégie *a posteriori* : retrouver les noms à partir des transcriptions automatiques, et donc du contenu textuel. De façon assez classique, notre système de reconnaissance de noms de personne, partant du signal audio, était composé de 3 modules :

1. Un système de reconnaissance automatique de la parole (RAP) ; dans ce travail, le système de Vocapia Research décrit dans la partie 4.2.2.
2. Un module de détection d'entités nommées (EN) dans la transcription automatique. Nous avons opté pour le système LIA_NE [Béchet and Charton, 2010].
3. Un module permettant le lien entre l'entité nommée et le nom normalisé de la personne. L'objectif est d'associer un identifiant unique à chaque nom détecté, quelle que soit sa forme écrite. Par exemple, l'épouse de Nicolas Sarkozy, ancien président de la République Française, peut se retrouver sous 5 formes différentes : *Carla Bruni-Sarkozy*, *Carla Bruni*, *Madame Bruni-Sarkozy*, *Carla Sarkozy*, *Madame Sarkozy*. Ce module est alors chargé de traduire chacune de ces formes sous l'identifiant normalisé *Carla_BRUNI-SARKOZY*.

La figure 5.1 présente les différents modules nécessaires à notre reconnaissance des noms de personne au travers d'un exemple. Ainsi, à partir de la séquence de mots préalablement transcrite *avec Hollande c'était pas si terrible finalement*, le module de reconnaissance d'EN est

en charge de localiser *Hollande* et d'évaluer qu'il s'agit d'une entité liée à un nom de personne (et non au pays *Pays-Bas*), alors que le dernier module (liaison EN vers nom de personne) permet d'associer *Hollande* à un référentiel normalisé (ici *François_HOLLANDE*). Le dernier module est censé aussi désambigüiser certains noms : *Hollande*, outre *François_HOLLANDE*, aurait aussi pu être *Thomas_HOLLANDE*, son fils, également un personnage médiatique.

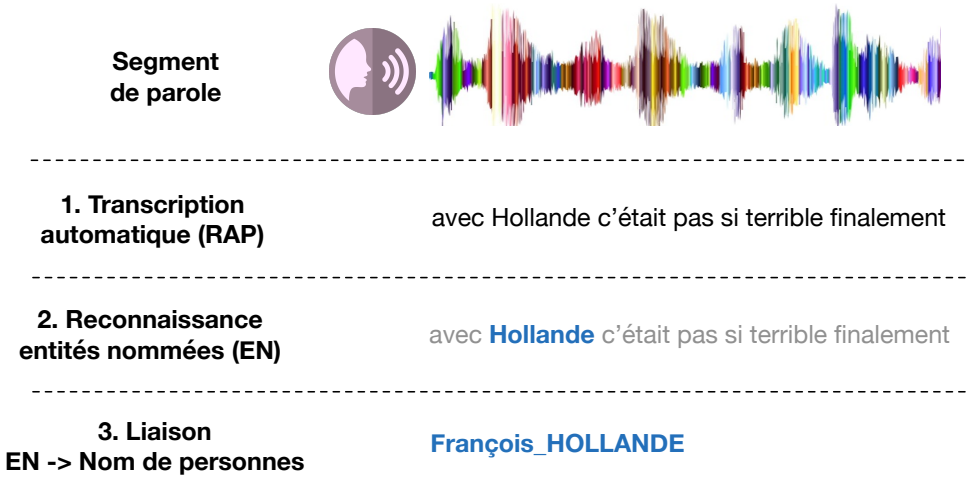

FIGURE 5.1 – Modules pour la reconnaissance des noms de personne à partir du signal audio au travers d'un exemple.

Notons que les trois modules sont alors potentiellement sujets aux erreurs, se répercutant de la transcription au module de détection automatique. Outre ces modules permettant d'extraire les noms de personne (NP), notre travail a consisté à corriger en amont la transcription automatique sur cette catégorie de mots. Nous avons alors proposé de corriger les régions d'erreurs de NP détectées automatiquement (voir chapitre 4) au moyen d'une recherche acoustique dans ces régions, que nous présentons dans la sous-partie suivante.

5.2.2 Correction par correspondance phonétique

Les erreurs de transcription sur les noms de personne, qu'elles prennent leur origine dans des mots hors vocabulaire, dans un contexte inconnu dans les données d'apprentissage, ou dans toute autre cause, sont susceptibles d'entraîner des mots transcrits qui soient phonétiquement proches du nom d'origine. Il peut s'agir d'une simple substitution (par exemple, *Karim examen* au lieu de *Karim Benzema*, *Marc librement* au lieu de *Marc Lièvremont*) ou d'une séquence de mots courts erronés (par exemple, *ou mon tourment* au lieu d'*Uma Thurman* ou *caler les vannes* au lieu de *Cadel Evans*). Nous avons alors choisi de corriger ces erreurs en comparant la représentation phonétique des mots transcrits à un dictionnaire phonétisé de noms de personne.

Notre stratégie de correction d'erreurs se déroule en deux passes :

1. Détecter les régions d'erreurs liées à des noms de personne. Nous nous appuyons sur le travail que nous avons proposé à ce sujet dans le chapitre 4.

2. Rechercher les noms de personne dans les régions d'erreurs sur la base de leur représentation phonétique.

Pour la passe 2, sachant la région d'erreurs automatiquement détectée durant la passe 1, le processus de correction consiste à rechercher dans un dictionnaire de noms de personne l'entité dont la représentation phonétique est la plus proche de la séquence phonétique associée aux mots de la région d'erreurs. Cette approche a été initialement étudiée dans le contexte de la détection de termes parlés à vocabulaire ouvert [Dubois and Charlet, 2008]. Dans le cas général, une telle approche peut être efficace en termes de rappel de détection de termes parlés mais doit être soigneusement utilisée afin d'éviter des taux de faible précision. Nous avons montré ici qu'il peut s'agir d'une approche pertinente pour la recherche de noms, à condition que l'étape de recherche soit guidée par une étape préliminaire de détection d'erreurs. La figure 5.2 présente l'approche de correction *a posteriori* des erreurs de noms de personne. L'algorithme de recherche pour la correction que nous avons proposé est détaillé dans [Dufour et al., 2012c].

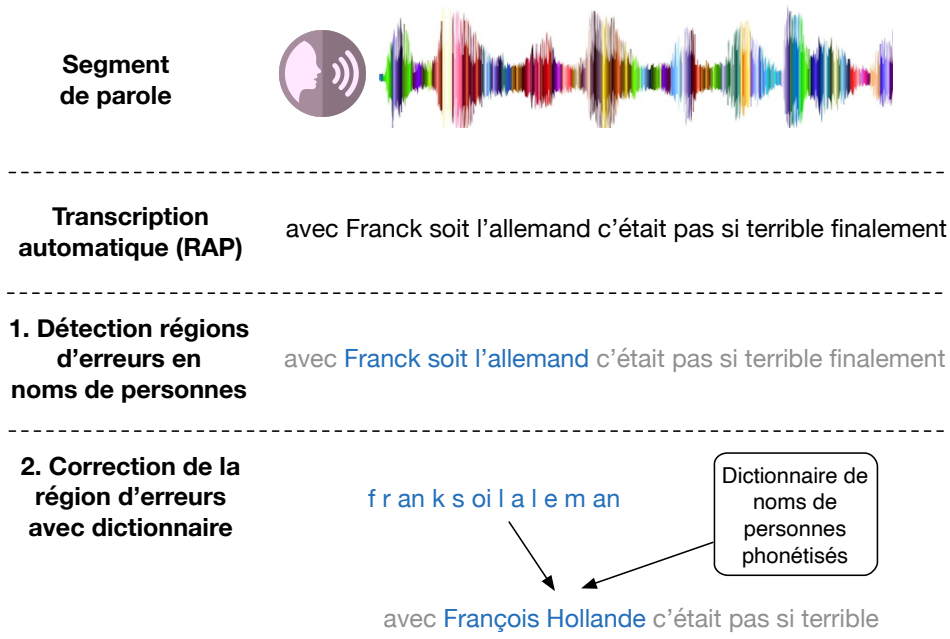

FIGURE 5.2 – Approche en deux passes pour la correction d'erreurs de noms de personne.

5.2.3 Corpus REPERE

Nous avons mené nos expériences sur un corpus d'émissions télévisées de deux chaînes françaises (2 émissions de BFMTV et 5 émissions de LCP) fourni par le défi REPERE [1]. Le corpus se caractérise par une grande variété de thèmes et de types d'émissions (débats, extraits des allocutions parlementaires, actualités...). Un corpus de développement est utilisé pour optimiser l'approche de correction des erreurs de transcription et un corpus de test pour évaluer les performances de reconnaissance des noms de personne. Les deux corpus contiennent chacun 3 heures de parole pour environ 30 extraits d'émissions. Ils ont été transcrits automatiquement, atteignant un taux d'erreur-mot de 20,8 % sur le développement et 24,4 % sur le test. Les noms de personne ont été annotés manuellement et associés à une forme normalisée. Le tableau 5.1 décrit le corpus REPERE utilisé pour notre tâche. Nous avons notamment fait ressortir les informations sur les mots et les régions d'erreurs de manière générale, mais également sur les noms de personne, qui nous intéressent tout particulièrement ici. Cela confirme ce que nous avions déjà observé, à savoir des régions d'erreurs beaucoup plus grandes pour les noms de personne, et très nombreuses, proportionnellement à leur nombre d'occurrences totales.

	Dev.	**Test**
# mots	34 312	34 683
# régions d'erreurs (taille moyenne)	2 296 (2,9)	3 069 (2,8)
# noms de personne (NP)	581	430
# régions d'erreurs des NP (taille moyenne)	234 (4,2)	184 (3,8)

Tableau 5.1 – Description d'une partie du corpus REPERE, au niveau des mots et des régions d'erreurs en général, ainsi que sur les noms de personne.

Enfin, la proportion d'occurrences de noms de personne couverts par notre dictionnaire atteignait 97,2 % pour le corpus de développement et 95,8 % pour le test. 13,9 % des noms de personne dans le corpus de développement étaient hors vocabulaire (16,7 % dans le test). Les mots hors vocabulaire représentaient 38,1 % des régions d'erreurs de noms de personne dans le développement (43,3 % pour le test), confirmant l'intérêt de traiter les régions d'erreurs de noms de personne en général et pas seulement du point de vue des mots hors vocabulaire.

5.2.4 Impact de la correction d'erreurs

Nous avons évalué notre système de correction d'erreurs *a posteriori* sur la tâche de détection de noms de personne de REPERE. Le système de détection des régions d'erreurs, et caractérisation en noms de personne, a été entraîné au moyen du corpus de JT complet (voir partie 4.2.2).

1. Notons qu'au moment de l'étude, cela ne constitue pas le corpus REPERE complet.

Deux systèmes de détection de régions d'erreurs ont été utilisés dans cette expérience : l'approche *Base*, s'appuyant sur les mesures de confiance seules, et l'approche *CRF*. De même, la fusion des deux approches a été explorée (*Fusion*). Enfin, la détection et la caractérisation en classe *noms de personne* ont été étudiées (*Fusion PN*).

L'évaluation a été réalisée en termes de rappel et précision sur les noms de personne (sous leur forme normalisée) détectés. Afin de comprendre l'influence de la qualité de détection des régions d'erreurs, nous avons fait varier le seuil de décision de ces approches sur le corpus de développement et appliqué notre stratégie de correction. Les résultats sur la détection des noms de personne sont reportés sur la figure 5.3.

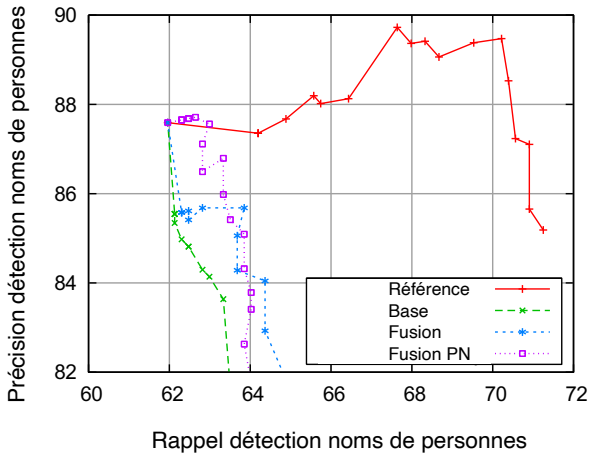

FIGURE 5.3 – Performance (rappel et précision) de la détection de noms de personne après correction des régions d'erreurs en faisant varier le seuil de décision sur le corpus de développement.

La courbe *Référence* correspond au cas optimal où les régions d'erreurs sont segmentées manuellement. Elle illustre le potentiel de la méthode de correction proposée et fournit les performances maximales pouvant être espérées avec notre approche de correction, avec un rappel optimal possible de 70,2 % et une précision de 89,4 %.

Notons que rechercher systématiquement les noms de personne dans les transcriptions (donc sans se concentrer sur les régions d'erreurs) présentait des performances trop basses pour être intéressantes. Sans entrer dans les détails, disponibles dans [Dufour et al., 2012c], nous avons vu que la qualité de détection automatique des régions d'erreurs est extrêmement importante. Il semble également que se concentrer sur la catégorisation des noms de personne, comme attendu, est pertinent (*Fusion PN*). Cette approche a ainsi été utilisée pour la vérification des performances sur le corpus de test, comme présenté dans le tableau 5.2.

La deuxième colonne du tableau 5.2 confirme toujours le potentiel de l'approche de correction d'erreurs lorsqu'elle est appliquée à des régions d'erreurs segmentées manuellement, alors que

	Sans correction	**Reg. err. man.**	**Reg. err. auto.**
Rappel	62,1	68,6	63,3
Précision	80,7	82,6	81,0
Taux d'erreur-mot global	24,4	24,1	24,3

Tableau 5.2 – Performances sur la tâche de détection des noms de personne sur le corpus de test sans correction des transcriptions, avec correction automatique des erreurs de transcription dans les régions d'erreurs manuelles (*Reg. err. man.*) et avec correction automatique dans les régions d'erreurs détectées automatiques (*Reg. err. auto.*).

la troisième colonne confirme également l'intérêt de l'approche *tout automatique* de correction d'erreurs de noms de personne, en améliorant le rappel et la précision par rapport à la transcription automatique sans correction. Il est intéressant de noter ici que l'impact sur le WER de la transcription globale est finalement assez faible en se concentrant sur cette catégorie de mots seule, et donc ne reflète pas réellement l'intérêt applicatif très fort de cette correction. Nous avons également évalué le WER simplement sur les segments ayant subi une modification : le taux d'erreur passe de 23,1 % à 22,1 % pour l'approche de segmentation d'erreurs manuelle, et diminue de 18,3 % à 18,0 % pour l'approche entièrement automatique.

Néanmoins, nous voyons, au travers de ce travail, les limites de l'évaluation : bien que l'intérêt de la correction de ces erreurs spécifiques soit très important d'un point de vue applicatif, la métrique finalement standard en RAP apparaît clairement inadaptée dans ce cadre. La partie suivante prend clairement le parti de réfléchir à cette métrique, et de proposer d'autres manières de rendre compte de la performance d'un système de RAP.

5.3 Correction des erreurs par adaptation des modèles

5.3.1 Contexte

Dans la partie précédente, nous nous sommes intéressés aux travaux que nous avons proposés pour la correction *a posteriori* des erreurs de transcription, *i.e.* sur les transcriptions automatiques déjà fournies par un système de RAP. Comme nous l'avons énoncé en introduction, l'autre possibilité est de proposer directement des améliorations au niveau de l'architecture ou des modèles du système de RAP. Dans le cadre de parole continue avec vocabulaire ouvert, les systèmes sont conçus pour pouvoir fonctionner dans des contextes variés. Néanmoins, il est évident que ces systèmes atteignent les performances les plus élevées lorsqu'ils sont construits et optimisés pour un contexte identifié, que ce soit au niveau acoustique (conditions d'enregistrement, bruits contrôlés, locuteurs connus...) ou linguistique (vocabulaire du domaine, modèles de langage construits pour traiter une thématique identifiée...). L'idéal serait donc de pouvoir

construire un système et des modèles directement dans le contexte d'utilisation ciblé. Plusieurs difficultés apparaissent néanmoins :

— Les données spécifiques nécessaires à l'entraînement des différents modules du système de RAP (modèles acoustiques et linguistiques) peuvent ne pas être en quantité suffisante.
— Il y a un risque de sur-apprentissage sur le contexte ciblé, et donc d'avoir un système incapable de traiter des données qui s'écarteraient, même un peu, des données trop spécifiques utilisées pendant l'apprentissage des modèles.

En prenant en compte ces problèmes, l'approche classique en RAP est d'adapter les différents modèles d'un système généraliste au domaine cible au moyen de données spécifiques. Tous les modules peuvent profiter de cette adaptation. Nous avons notamment pu proposer, pendant ma thèse, une méthode pour l'adaptation des modèles acoustiques dans le cadre de la parole spontanée [Dufour et al., 2010]. Dernièrement, j'ai eu l'occasion de collaborer avec Salima Mdhaffar sur une partie de ses travaux de thèse sur la *Reconnaissance de la parole dans le contexte de cours magistraux : évaluation, avancées et enrichissement* [Mdhaffar, 2020]. S. Mdhaffar a notamment travaillé sur l'adaptation des modèles de langage dans le contexte de cours en ligne en vue de leur exploitation, que ce soit pour le sous-titrage automatique ou encore pour l'exploitation des transcriptions dans d'autres applications, comme l'indexation automatique.

Cela rejoint donc les travaux dans lesquels j'ai pu m'investir au niveau scientifique, S. Mdhaffar ayant continué les réflexions, avec propositions, sur la problématique de l'évaluation des systèmes de RAP. En effet, comme nous l'avons vu dans les différents chapitres, le WER reste une métrique qui, en plus d'être faiblement informative sur les erreurs produites par les systèmes, souffre de la prise en compte identique de chaque erreur de transcription. Dans le cadre de la thèse de S. Mdhaffar et l'indexation de cours en ligne, nous pouvons imaginer que certains mots-clés de ces cours sont bien plus importants que d'autres mots, ce que le WER ne peut rendre compte. Une partie des travaux présentés dans cette partie sont issus de son travail, et un peu de nos réflexions et rédactions communes, sur l'adaptation des modèles de langage et l'évaluation qualitative des systèmes de RAP dans le contexte applicatif de l'indexation de documents [Mdhaffar et al., 2019]. La sous-partie suivante présente le corpus PASTEL, réalisé dans le projet ANR du même nom, puis nous présentons dans la partie 5.3.3 l'adaptation des modèles de langage proposée, et la définition de métriques pour l'évaluation de systèmes de RAP dans la partie 5.3.4. Enfin, une évaluation extrinsèque, orientée *tâche*, est décrite dans la partie 5.3.5.

5.3.2 Corpus PASTEL

Bien que n'ayant pas participé à la constitution et la réalisation de ce corpus, cette partie permettra de bien comprendre le contexte du travail de thèse de S. Mdhaffar, qui en est à l'origine. Le corpus a été collecté durant le projet ANR PASTEL (*Transcription Automatique de la Parole pour l'Apprentissage et la Formation*). Les données sont issues :

1. du projet CominOpenCourseware (COCo) fournissant des vidéos de cours accompagnées éventuellement des supports projetés pendant le cours ainsi que l'alignement temporel entre le discours de l'enseignant et le changement de diapositive ;
2. de la plateforme en ligne Canal-U agrégeant des ressources audiovisuelles liées à l'enseignement supérieur.

Tous les cours sélectionnés ont été transcrits manuellement. Les cours ont également été segmentés manuellement en thématiques selon deux niveaux de granularité :

— La granularité 1 (G1) signifie qu'une nouvelle notion est initiée tout en restant dans le même sujet global du cours.

— La granularité 2 (G2) est utilisée lorsqu'un changement global de sujet se produit, l'objectif étant de diviser le cours en chapitres (chaque chapitre est alors composé d'au moins une granularité G1).

Les mots considérés comme faisant partie du domaine spécifique pour chaque cours ont été extraits manuellement des transcriptions manuelles ainsi que des supports (diapositives) de présentation. L'objectif sous-jacent était de déterminer dans quelle mesure ces mots spécifiques étaient reconnus avec et sans adaptation des modèles de langage. Les mots du domaine sont considérés comme des expressions linguistiques, nommées dans les travaux *expression-clé*, qui se réfèrent à des concepts, objets ou entités essentiels à la compréhension de la diapositive actuelle ou d'une transcription donnée. Cette annotation a été faite pour les cours pour lesquels des diapositives ont été fournies.

Le corpus comprend au final 9 vidéos de cours pour une durée totale d'environ 10 heures. Le tableau 5.3 décrit le corpus, avec, pour chaque cours, le nombre de sujets identifiés (colonnes G1 et G2), le nombre d'expressions-clés dans les transcriptions automatiques (colonne $\mathbf{EC_t}$) et les diapositives du cours (colonne $\mathbf{EC_d}$), le nombre de diapositives (colonne #d), et enfin la durée. À noter que les cours issus de la plateforme Canal-U n'ont pas de supports de cours associés (donc aucune annotation manuelle en expressions-clés). Une description plus détaillée du corpus est disponible dans [Mdhaffar et al., 2020].

5.3.3 Adaptation du modèle de langage

Comme vu dans la description du corpus PASTEL, les cours ont une thématique spécifique qui rendent l'utilisation d'un système de RAP générique peu performant sur ces données, en particulier sur les mots ou expressions du domaine. Afin d'obtenir des données spécifiques aux domaines de chaque cours, l'approche suivie a été de prendre Internet comme source de données. L'approche proposée dans [Mdhaffar et al., 2019] a été d'utiliser les titres des diapositives des supports de cours comme requêtes sur des moteurs de recherche sur Internet (dans le cadre de ce travail, Google). Le contenu textuel des différents liens vers les sites web renvoyés par chaque requête est alors extrait (maximum de 400 pages par requête). L'adaptation des modèles de

Intitulé du cours	G1	G2	EC_t	EC_d	#d	Durée	Source
Introduction à l'informatique	31	2	47	54	75	1h04	COCo
Introduction à l'algorithmique	38	10	25	35	62	1h17	
Les fonctions	35	3	109	78	137	1h14	
Réseaux sociaux et graphes	43	7	54	84	64	1h05	
Algorithmique distribuée	72	5	232	146	73	1h16	
Langage naturel	52	5	120	100	55	1h09	
Architecture de la république	49	7	-	-	-	1h21	Canal-U
Méthode traditionnelle	12	7	-	-	-	0h41	
Imagerie	57	1	-	-	-	1h08	
Total	389	46	587	497	466	10h19	-

Tableau 5.3 – Description des données du corpus PASTEL (G : Granularité de la segmentation thématique, EC : expression-clé, t : transcription manuelle, d : diapositive) [Mdhaffar, 2020]

langage est, au final, assez classique : une interpolation linéaire est réalisée entre un modèle de langage générique et un modèle de langage entraîné à partir des données Internet, afin de fournir le nouveau modèle adapté.

Le système de transcription est celui développé au LIUM, et qui s'appuie sur la boîte à outils Kaldi [Povey et al., 2011]. Le modèle de langage générique est un modèle de langage n-gramme entraîné sur des transcriptions manuelles, mais également sur des textes écrits issus d'articles de presse, pour un total de 1,6 milliards de mots. Le vocabulaire du modèle de langage générique contient environ 160 000 mots. Plus de détails sur la construction du modèle de langage peuvent se trouver dans [Rousseau et al., 2014].

5.3.4 Évaluation de la transcription automatique

La métrique Individual Word Error Rate (IWER)

Le Individual Word Error Rate (IWER) est ici une variante du taux d'erreur-mot (WER) classique en RAP. Introduit dans [Goldwater et al., 2010], le IWER se focalise sur l'évaluation d'un mot identifié. Il modifie principalement la façon de pénaliser les insertions : si l'on se focalise sur le taux d'erreur d'un mot particulier, il est impossible de savoir si l'insertion provient de ce mot ou des mots adjacents. Une responsabilité équivalente est alors supportée par les mots adjacents. Ainsi, pour le i^{eme} mot de référence, son IWER est calculé comme suit :

$$IWER(w_i) = sup_i + sub_i + \alpha.ins_i \tag{5.1}$$

où $sup_i = 1$ si w_i est supprimé, $sub_i = 1$ si w_i a été substitué par un autre mot, et $ins_i =$ le

nombre d'insertions de mots adjacents à w_i. Le paramètre α est calculé comme suit :

$$\alpha = \frac{I}{\sum_{w_i} ins_i} \tag{5.2}$$

où I est le nombre total d'insertions dans tout le corpus. Au final, un IWER peut être calculé pour un ensemble de mots n :

$$IWER(w_1...w_n) = \frac{1}{n}\sum_{i=1}^{n} IWER(w_i) \tag{5.3}$$

S. Mdhaffar a proposé dans sa thèse d'évaluer des transcriptions, en plus du WER classique, au moyen de l'IWER, en l'adaptant à son problème de transcription pour des cours en ligne. L'avantage de l'IWER est de pouvoir fournir une évaluation pour un ensemble de mots ciblés : il s'agit ici de ne pas évaluer de manière globale la transcription, mais de ne s'intéresser qu'à des mots estimés comme informatifs. Les mots considérés du domaine des cours font alors partie de l'ensemble de mots évalués. Mais au lieu d'avoir un seul ensemble de mots, elle a choisi de considérer plusieurs ensembles (les différents cours) et de faire la moyenne des IWER obtenus sur ces différents ensembles :

$$IWER_{Average} = \frac{1}{\sum_{y=1}^{m} n_m}\sum_{j=1}^{m}\sum_{i=1}^{n_m} IWER(w_i) \tag{5.4}$$

où m est le nombre de transcriptions de cours et n_m le nombre de mots du cours m.

Même si je n'ai pas participé à la proposition de la métrique IWER, il me semblait pertinent de la présenter pour comprendre le cheminement des travaux exposés dans ce chapitre.

Évaluation de l'adaptation des modèles de langage

L'application de la métrique IWER a été vérifiée dans la tâche d'adaptation des modèles de langage pour l'évaluation de la transcription automatique des cours en ligne. Le tableau 5.4 présente les performances obtenues sur le corpus PASTEL en termes de WER et IWER avec le système de RAP générique et le système avec modèle de langage adapté.

Les résultats ont montré que, outre le fait que le WER ait baissé significativement avec le modèle de langage adapté (ce qui est classique), la métrique IWER reflète une baisse beaucoup plus importante, peu importe les expressions-clés considérées. Il semble donc qu'une évaluation globale au moyen du WER masque l'apport réel de la correction, et donc la performance intrinsèque des systèmes de RAP. Afin de vérifier cette observation, la partie suivante s'intéresse à l'évaluation de l'adaptation sur une tâche d'indexation de documents.

Métrique	Mots sélectionnés	RAP générique	RAP adapté
WER	Tous les mots	19,46	**16,42**
IWER	Expressions-clés des titres	29,52	**14,05**
	Expressions-clés des diapositives	32,31	**14,52**
	Expressions-clés de la transcription manuelle	31,00	**17,30**

Tableau 5.4 – Performances (en WER et IWER) des systèmes de RAP *générique* et *adapté* sur le corpus PASTEL. Le IWER a été calculé sur des ensembles différents d'expressions-clés (titres, diapositives et transcription manuelle).

5.3.5 Évaluation sur la tâche d'indexation de documents

L'intérêt était de confirmer les observations précédentes en évaluant l'adaptation sur une tâche ciblée (évaluation extrinsèque), ici leur indexabilité. En d'autres termes, l'idée était de déterminer si la qualité de la transcription (et donc le fait d'avoir un WER ou IWER plus faible) joue un rôle dans son indexation. Les segments des transcriptions ont été indexés à l'aide du moteur de recherche Lemur. Trois ensembles de segments ont été comparés : ceux des transcriptions manuelles, ceux des transcriptions automatiques sans adaptation, et ceux des transcriptions automatiques avec adaptation. L'objectif était d'obtenir une liste de segments ordonnée en fonction des requêtes fournies en entrée de Lemur (ici, des mots du domaine sélectionnés manuellement). Est considérée ici comme référence la liste ordonnée de segments obtenue avec la transcription manuelle. Le coefficient de Spearman [Gautheir, 2001] est ensuite utilisé pour mesurer la corrélation de rang entre les transcriptions manuelles et automatiques (sans et avec adaptation).

Requête	RAP générique	RAP adapté
Expressions-clés des titres	0,458	**0,588**
Expressions-clés des titres de la transcription manuelle	0,288	**0,516**

Tableau 5.5 – Performance sur la tâche d'indexabilité des transcriptions (coefficient de Spearman) sur le système de RAP générique (sans adaptation) et le système de RAP adapté.

Le tableau 5.5 présente les scores de corrélation moyens obtenus au moyen des transcriptions automatiques sans adaptation (générique) et des transcriptions adaptées. Les résultats indiquent une meilleure indexabilité en faveur de l'adaptation. D'autres expériences ont été menées dans [Mdhaffar et al., 2019], en particulier sur une tâche de recherche de documents.

5.4 Conclusion

Nous nous sommes tout d'abord intéressés, dans ce chapitre, à la correction des erreurs faites par les systèmes de RAP selon deux approches différentes. Premièrement, nous avons proposé

une approche de correction *a posteriori* des transcriptions automatiques. Ce travail prenant forme dans le cadre du défi REPERE, nous nous sommes concentrés sur la correction des noms de personne, généralement mal transcrits par les systèmes de RAP génériques, et dont l'impact applicatif peut être grand malgré un impact faible sur le WER. Pour ce faire, nous avons choisi une approche complètement automatique, en détectant tout d'abord les régions d'erreurs, avec l'approche présentée dans le chapitre 4, puis en appliquant une recherche phonétique de noms de personne dans cette région d'erreurs. Le nom de personne le plus probable remplaçait alors les mots contenus dans la régions d'erreurs. Bien qu'ayant eu une implication moindre, l'approche de correction par adaptation des modèles de langage réalisée pendant la thèse de S. Mdhaffar a été décrite, ces travaux permettant de comprendre mes intérêts de recherche ces dernières années.

Plus généralement, ce chapitre a continué l'ouverture des discussions et propositions d'évaluation de la transcription automatique, en particulier d'un point de vue applicatif. Il s'agissait de montrer, dans un premier temps, que le taux d'erreur-mot classique ne reflète pas l'intérêt applicatif de la correction de mots spécifiques, tels que les noms de personne. De même, certaines métriques semblent mieux rendre compte de la qualité des transcriptions, que ce soit au travers de métriques nouvelles sur sa qualité intrinsèque, comme ce qui a été proposé avec le IWER par S. Mdhaffar, ou que ce soit dans le cadre d'une application utilisant ces transcriptions, comme la tâche d'indexation de documents.

La problématique de l'évaluation reste néanmoins toujours ouverte et compliquée : la communauté s'accorde à dire que les métriques sont clairement imparfaites, mais, de ce qui transparaît de ces études, aucune n'arrive à tout prendre en compte. Il faut néanmoins toujours être critique par rapport à des scores, qui ne reflètent au final qu'une performance au niveau des systèmes automatiques proposés. Il reste bien des portes à explorer sur la problématique de l'évaluation. Gardons cependant à l'esprit que les domaines historiques en TAL, tels que la RAP, restent assez simplement *évaluables* malgré les limites des métriques, et ce, grâce à un consensus acquis durant des décennies. Dans les travaux que nous présentons dans le chapitre suivant, nous verrons que certaines tâches récentes, intégrant des problématiques de plusieurs disciplines, se doivent de construire des protocoles d'évaluation complets, actuellement complètement inexistants.

Troisième partie

Interdisciplinarité et traitement du langage

Chapitre 6

Exploitation des réseaux sociaux pour l'analyse d'événements

Sommaire

6.1 Introduction

Comme nous avons pu le voir et le détailler dans les parties et chapitres précédents, traiter automatiquement le langage, au travers de documents souvent bruités et ne respectant que peu les règles linguistiques, est difficile et requiert des approches dépassant le simple niveau *mot*. De même, l'évaluation des systèmes automatiques est un questionnement permanent, les tâches historiques en TAL (reconnaissance automatique de la parole, traduction automatique...) n'échappant pas à des critiques, toujours actuelles, sur les métriques utilisées.

Les travaux que j'ai menés pendant plusieurs années sur des corpus parlés, en particulier sur des enregistrements d'actualités radiophoniques et télévisées, m'ont amené assez naturellement à travailler sur des données issues des réseaux sociaux numériques (RSN). En effet, j'ai surtout travaillé sur des sorties de systèmes de transcription automatique dans le cadre de parole continue et spontanée, et donc sur des représentations clairement bruitées (agrammaticalité, erreurs de transcription, répétitions de mots, mots hors-vocabulaire...), mises en lumière dans le chapitre 1. Dans le contexte d'échanges de messages sur Internet, même si ce bruit est différent, nous sommes face à des difficultés que des méthodes classiques ne peuvent simplement traiter (messages courts, absence possible de contexte, erreurs grammaticales et orthographiques...).

De façon générale, les RSN permettent aux communautés d'utilisateurs d'échanger et de partager des ressources dans le monde entier (idées, opinions, données...) avec un public qui ne cesse de grandir. À titre d'exemple, en juillet 2020, Facebook comptait 2,6 milliards d'utilisateurs mensuels actifs dans le monde, Youtube 2 milliards, Instagram 1,1 milliards, et Twitter 326 millions [1]. Les chercheurs, en particulier dans les domaines du TAL et de la recherche d'information (RI), ont saisi ce phénomène sans précédent par le nombre d'utilisateurs que ces réseaux agrègent, ainsi que par la taille des données multimédias échangées (textes, vidéos, audio...), ouvrant alors de nouveaux enjeux de recherche.

De façon assez classique, les chercheurs en TAL et RI se sont alors employés à proposer des solutions pour aider à résoudre des problèmes nouveaux apparus avec les RSN, tels que des tâches de classification finalement courantes [Sriram et al., 2010, Lee et al., 2011, Pennacchiotti and Popescu, 2011], mais qui intègrent souvent une difficulté supplémentaire liée à une subjectivité, que ce soit dans leur définition, leurs données et/ou leur annotation par des experts humains (analyse de sentiments [Rouvier and Favre, 2016], détection d'opinions [Khan et al., 2014]...). Finalement, les RSN apparaissent comme un objet d'étude pluridisciplinaire, ceux-ci étant étudiés en sociologie [Java et al., 2007], en droit [Strahilevitz, 2005], en marketing [Bolotaeva and Cata, 2010]...

Le cadre de travail du projet ANR SuMACC (voir partie 9.2.1), mais surtout mon implication dans le projet ANR GaFes (voir partie 9.3.1), m'ont amené à faire évoluer ma manière d'appréhender mon travail de recherche et mes problématiques scientifiques. Les questions que j'ai pu précédemment me poser, comme tout ce qui concernait l'évaluation (et surtout leurs limites) dans la partie II, prennent tout leur sens dans ces projets de recherche à la coloration clairement interdisciplinaire. Il convient ici de travailler en collaboration avec des chercheurs d'autres thématiques pour arriver à exploiter, analyser, concevoir et évaluer des systèmes de traitement du langage conçus ici dans le contexte d'étude des réseaux sociaux.

Un besoin de compréhension des approches proposées est nécessaire pour tirer profit des résultats obtenus avec des systèmes de TAL, l'attente des sciences humaines étant extrêmement

1. `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`

forte sur ces sujets. Dans un premier temps, nous présentons, dans la partie 6.2, un travail qui a été réalisé dans le cadre de la thèse de Mohamed Morchid concernant une étude qualitative de différentes caractéristiques extraites automatiquement, afin de rendre compte de l'importance, et ainsi de leur influence, de chacune d'entre elles dans le phénomène de relais massif d'une information sur les réseaux sociaux [Morchid et al., 2014f]. Nous voyons ensuite une étude, dans la partie 6.3, que nous avons réalisée dans le cadre du projet GaFes et de l'alternance de Mathias Quillot sur les plongements lexicaux et temporels, et en particulier sur la manière dont il est possible de les exploiter et de rendre compte des informations, pour qu'elles soient exploitables par des chercheurs en sciences humaines [Quillot et al., 2017]. Enfin, nous mettons en lumière des travaux réalisés en collaboration entre chercheurs en sociologie et chercheurs en informatique (partie 6.4) pour l'extraction d'une grande diversité d'opinions dans des messages argumentés [Dufour et al., 2018]. Il s'agira enfin de montrer les enjeux, mais aussi les limites et les difficultés rencontrées, tout en montrant la richesse d'une recherche à vocation interdisciplinaire.

6.2 Étude de caractéristiques liées à la diffusion massive de messages sur Twitter

L'arrivée des RSN a introduit de nouvelles façons d'échanger entre humains, et avec elles, de nouveaux usages. Parmi eux, la diffusion massive d'une information, ici extrêmement rapide et mondialisée, est devenue un phénomène important. Nous proposons ici une étude qualitative de ce phénomène, souvent dénommé *buzz*, et en particulier des caractéristiques importantes en vue de sa détection automatique mais également de sa compréhension, pouvant alors intéresser de nombreux domaines de recherche, en particulier en sciences humaines et sociales (SHS).

6.2.1 *Buzz* et TAL sur les réseaux sociaux numériques

Les SHS se sont, bien entendu, emparées de ce phénomène nouveau de *buzz*, cherchant par exemple à le comprendre [Le Deuff, 2006], l'étudier [Chabi, 2008], ou encore évaluer son impact [Delcroix et al., 2016]. Son traitement et son analyse d'un point de vue automatique est également apparu très tôt, les chercheurs en TAL y voyant un sujet original apportant de nouvelles problématiques de recherche (données massives, messages courts, language *atypique*...). De par son côté relativement plus ouvert et plus simple pour la collecte de données, la plateforme Twitter est rapidement devenue privilégiée pour étudier le *buzz*. Les messages textuels envoyés via cette plateforme (*tweets*) ont la particularité d'être composés d'un nombre réduit de caractères (140 caractères maximum pendant notre étude, 280 lors de l'écriture de ce manuscrit). La puissance de ce service réside dans le fait que les utilisateurs peuvent relayer ou transmettre un tweet (cette possibilité est appelée *retweet*). Ainsi, les informations diffusées par un utilisateur n'ont évidemment pas le même impact sur son environnement si elles sont très peu ou alors

beaucoup, voire massivement, relayées en peu de temps (ici, un très grand nombre de retweets).

Le mécanisme de retweet avait déjà été étudié dans une optique de prédiction. Dans [Suh et al., 2010], les auteurs ont proposé d'extraire un ensemble de caractéristiques sur le contenu et le contexte pour expliquer ce mécanisme (urls et mots-clés en particulier), et également être capable de fournir un modèle prédictif sur le nombre de retweets. D'un point de vue prédictif, plusieurs travaux se sont concentrés sur le phénomène de retweets massifs [Zaman et al., 2010, Hong et al., 2011]. Les caractéristiques qui influencent la probabilité de retweet, telles que l'âge, le taux d'envoi de messages par les utilisateurs ou la taille des tweets, ont été analysées par [Comarela et al., 2012]. Le phénomène du retweet massif est toujours un problème actuel, comme le prouvent les publications récentes dans ce domaine [Hoang and Mothe, 2018].

Finalement, il y a quelques années, une des préoccupations principales des chercheurs en TAL était de pouvoir fournir des outils et approches originales pour la détection de ce phénomène. Cependant, peu d'études mettaient en évidence l'influence des caractéristiques des tweets dans la propagation massive d'informations. Nous avons alors proposé un travail préliminaire visant à étudier un ensemble de caractéristiques liées à l'environnement d'un utilisateur et au contenu du message, avec un focus spécial sur le *buzz*. Cette étude permet d'évaluer l'impact de chaque fonctionnalité dans le phénomène de retweet massif. Nous avons ensuite proposé de vérifier notre étude au moyen d'une classification en utilisant les caractéristiques de retweets les plus corrélées choisies dans l'étude qualitative. Cette étude dépasse alors le cadre du TAL, en apportant des premières réponses, et pistes, dans la compréhension de la propagation de l'information.

6.2.2 Analyse de caractéristiques liées aux *retweets* massifs

Corpus et caractéristiques

Notre étude a nécessité la collecte d'un grand corpus de données. Au travers de l'API publique Twitter [2], un corpus de 6 millions de messages (*tweets*) a pu être récupéré. La date d'émission des tweets se situe entre le 7 avril et le 2 mai 2011, soit des données sur environ 1 mois. Cela nous permet de conserver une cohérence temporelle dans la collecte. En effet, nous pourrions imaginer que les comportements changent, par exemple, d'une année sur l'autre, avec l'augmentation du nombre d'utilisateurs, les usages de la plateforme par les utilisateurs... De cet ensemble de données, nous avons choisi de conserver 30 903 tweets pour notre étude, ayant chacun été relayé (*retweeté*) de 1 à plus de 100 fois. À noter qu'au moment de la collecte, Twitter ne fournissait pas le nombre exact de retweets supérieurs à 100 (le nombre était alors noté 100+ sur la plateforme), alors qu'au moment de l'écriture du manuscrit, le nombre exact est fourni. Le corpus est constitué à 44 % de messages avec retweets massifs (> 100), 43 % entre 0 et 30, et 13 % entre 30 et 100. Nous nous sommes appuyés sur des caractéristiques déjà prises en compte dans des travaux

2. https ://developer.twitter.com/

précédents sur l'analyse des retweets [Suh et al., 2010, Luo et al., 2013] :

— Caractéristiques liées au contenu du message :

1. *Retweet* : nombre de fois où le message a été relayé.
2. *Hashtag* : nombre de *hashtags* dans le message [3].
3. *Mention* : nombre d'utilisateurs (avec caractère @) mentionnés dans un message.
4. *URL* : présence ou non d'une adresse vers un site Internet dans le message.

— Caractéristiques liées à l'utilisateur :

1. *Ancienneté* : ancienneté du compte (en jours).
2. *Favoris* : nombre de tweets que l'utilisateur a mis en favoris (*j'aime*).
3. *Abonné* : nombre d'utilisateurs qui suivent le compte.
4. *Abonnement* : nombre de comptes que l'utilisateur suit.
5. *Tweet* : nombre de tweets écrits par l'utilisateur.

Étude des facteurs explicatifs pour l'analyse des retweets massifs

Nous avons ensuite proposé d'étudier la corrélation entre les différentes caractéristiques précédemment décrites sur les messages ayant un retweet massif (*i.e.* les messages avec nombre de retweets supérieur à 100) au moyen d'une analyse en composantes principales (ACP), permettant alors de transformer ces caractéristiques choisies en un ensemble de variables décorrélées, nommées axes principaux. Un ensemble de facteurs est alors obtenu, évitant la redondance de l'information tout en réduisant le nombre de variables. Chaque facteur représente une certaine partie de la variance totale de l'ensemble de données.

Le tableau 6.1 présente les coefficients de corrélation linéaire entre les variables initiales et les 4 premiers facteurs. Nous pouvons voir ici que le Facteur 2 permet de distinguer des caractéristiques liées au contenu (*url* et *mention*) de celles liées à l'utilisateur. De même, nous pouvons observer l'opposition sur les axes entre les caractéristiques de *retweet*, qui nous intéresse tout particulièrement ici, et *url*/*mention*. Durant ces expériences, nous avons également observé que plus de 18 % de la variabilité (Facteur 1) est portée par une corrélation forte entre les caractéristiques liées à l'utilisateur. La caractéristique *hashtag* est, quant à elle, portée par le Facteur 4, alors que celle de *abonné* est portée par le Facteur 3, considérant alors l'indépendance de ces caractéristiques. Enfin, la plupart des caractéristiques liées aux utilisateurs semblent former un même cluster.

L'analyse factorielle nous a permis d'analyser l'importance de chaque caractéristique, selon son impact sur la variabilité totale et sa corrélation avec la caractéristique *retweet*, du plus au moins important : *hashtag*, *abonné*, *mention*, *tweet*, *url*, *status*, *abonné* et *ancienneté*.

3. Un *hashtag* est ici un mot-clé qu'un utilisateur choisit de mettre en avant dans un message. Il est précédé par le caractère # (ex : #keyword).

Contenu	**Facteur 1**	**Facteur 2**	**Facteur 3**	**Facteur 4**
Retweet	-0,3295	**0,7009**	-0,0072	0,1481
Hashtag	0,0112	-0,1757	-0,2089	**0,9306**
Mention	0,2569	**-0,5927**	-0,1982	-0,2175
URL	0,1682	**-0,5240**	0,1599	0,1431
Utilisateurs	**Facteur 1**	**Facteur 2**	**Facteur 3**	**Facteur 4**
Ancienneté	**0,5071**	0,0901	0,4011	0,1717
Favoris	**0,5389**	0,1974	0,3391	-0,0007
Abonné	0,3745	0,1678	**-0,6694**	-0,0795
Abonnement	**0,5133**	0,2025	-0,4886	-0,0062
Tweet	**0,7092**	0,2178	0,1964	0,0232

Tableau 6.1 – Corrélations des variables-facteurs pour l'analyse de caractéristiques liées au retweet massif.

6.2.3 Détection automatique des *retweets* massifs

Grâce à l'analyse qualitative réalisée dans la partie précédente, nous avons une idée plus précise des caractéristiques qui semblent avoir un impact dans le phénomène de retweet. Nous avons alors proposé de vérifier cette étude en appliquant nos observations précédentes sur une tâche de détection automatique de retweets massifs, l'objectif étant de montrer que les caractéristiques *choisies* ont un impact plus important dans la décision de retweet (*i.e.* permettent ici de mieux classifier les messages selon leur niveau de retweet).

Dans notre contexte d'étude, nous nous sommes concentrés sur deux classes : les tweets étant très faiblement relayés (ici, de 0 à 30 fois) et ceux massivement retweetés (plus de 100 fois). Notons qu'il n'est pas ici question de prédire le nombre de retweets, mais de comprendre le phénomène de retweet massif : les messages ayant donc un nombre de retweet entre 30 et 100 ont volontairement été retirés. Nous avons aussi fait ce choix pour pallier un problème lié à ces données et qui concerne leur durée de vie. En effet, ceux-ci peuvent être relayés tout au long de leur existence (*i..e.* jusqu'à ce que l'auteur décide de le retirer par exemple) : lors de leur collecte à un instant t, ce nombre peut ne pas être le même qu'à l'instant $t + 1$. Le *buzz* étant un phénomène de diffusion massive et rapide d'une information, si celle-ci n'est pas rapidement massivement relayée, alors elle est en dehors de l'étude que nous considérons.

Afin de classifier les messages selon leur niveau de retweet (faible ou massif), nous avons utilisé une approche par apprentissage supervisé avec les machines à vecteurs de support (SVM) [Vapnik, 1999]. Un classifieur SVM est alors entraîné avec les caractéristiques définies dans la partie 6.2.2 afin de détecter automatiquement si un tweet a été massivement relayé ou non. Pour éviter un déséquilibre entre les données, nous avons choisi, pour chaque classe considérée, un ensemble d'entraînement de 12 195 tweets et un ensemble de test de 1 355 tweets.

Le tableau 6.2 présente les résultats de classification globaux obtenus sur les deux classes

considérées, en termes de rappel, précision et F-mesure. Afin d'évaluer la pertinence des caractéristiques utilisées, nous avons fait varier leur nombre de 1 à 8. Plus précisément, le classifieur 1 utilise uniquement la caractéristique dont notre étude a jugé qu'elle est la plus pertinente (ici, *hashtag*), le classifieur 2, les deux plus pertinentes (*hashtag + abonné*) et ainsi de suite en suivant l'ordre défini dans la partie 6.2.2.

# Caractéristiques	**Rappel**	**Précision**	**F-mesure**	**Gain (%)**
Hashtag	50,0	24,9	33,3	–
+ Abonné	57,4	58,0	57,7	+73,3
+ Mention	63,0	66,0	64,9	+12,5
+ Tweet	62,7	66,8	65,7	+1,2
+ URL	**64,1**	**67,8**	**65,9**	+0,3
+ Status	62,8	66,8	64,8	-1,2
+ Abonné	62,8	66,8	64,8	–
+ Ancienneté	62,4	64,7	63,5	-2,0

Tableau 6.2 – Performance (précision, rappel et F-mesure) de la classification de tweets selon le nombre de leur retweet (faible ou massif) en faisant varier le nombre de caractéristiques utilisées.

Nous pouvons observer que les performances de classification les plus élevées sont obtenues en utilisant les 5 caractéristiques considérées comme les plus pertinentes pour définir le retweet massif dans notre étude qualitative. Cela confirme l'intuition initiale qui a motivé ce travail, à savoir que certaines caractéristiques jouent un rôle dans le phénomène de retweet, et en particulier le relais massif de messages, mais que toutes les caractéristiques ne sont pas pertinentes, ou tout du moins ne semblent pas déterminantes pour qu'un message soit massivement relayé.

Plus précisément, en considérant le meilleur classifieur (5 caractéristiques), nous avons observé que les performances de la détection de messages fortement relayés (massif), avec un rappel de 86,9 % et une précision de 59,8 %, sont bien meilleures que la détection des messages faiblement relayés (faible), avec un rappel de 41,2 % et une précision de 75,8 %. Cela semble cohérent puisque l'étude proposée s'est concentrée sur les caractéristiques corrélées avec le phénomène de retweet massif. Ces performances plus faibles des messages peu relayés pourraient également s'expliquer par l'instant de collecte : un tweet ayant un faible nombre de retweets à l'instant t ne signifie pas qu'il ne sera pas retweeté massivement à l'avenir, même si nous avons pris la précaution d'éliminer les tweets avec un niveau de retweet *intermédiaire*.

Cette étude préliminaire constitue une proposition pour comprendre la diffusion des messages dans les réseaux sociaux. Nous avons pris le parti de ne pas prendre le problème de la diffusion massive de messages (le *buzz*) comme une simple tâche de classification, mais plutôt de fournir une analyse permettant d'expliquer les facteurs importants permettant de donner une première compréhension de ce phénomène. Au final, la classification, et les performances associées, restent secondaires, notre étude étant plutôt pensée comme un travail dépassant le cadre du TAL.

6.3 Plongements lexicaux et temporels dans le cadre d'événements culturels

6.3.1 Contexte d'étude

Bien qu'exploratoire, le travail précédemment présenté sur l'étude du phénomène du retweet massif se déroulait, comme tous les travaux présentés jusqu'alors, dans un cadre expérimental bien défini, incluant un processus d'évaluation clair : les résultats obtenus par notre système de classification automatique ont pu confirmer nos hypothèses de départ.

Le travail de cette partie, réalisé pendant le projet ANR GaFes et l'alternance de M. Quillot, apparaît dans un contexte d'étude beaucoup plus difficile : la problématique et les hypothèses initiales sont connues et clairement définies, au contraire de son évaluation. Il s'agissait ici d'être capable d'extraire des informations de gigantesques bases de données textuelles issues du web afin de permettre à des chercheurs en sociologie d'en tirer des informations pertinentes pour leurs études. Il a donc été nécessaire de proposer des approches originales, mais également de travailler avec des chercheurs en sociologie et en RI pour définir un protocole d'évaluation pouvant rendre compte de la pertinence des méthodes proposées.

Nous nous plaçons dans le contexte d'événements culturels récurrents (ici, des festivals) et de leur occupation dans l'espace des RSN pendant et en dehors du festival. Pour des événements particuliers (concerts, festivals, élections présidentielles...), les personnes (grand public, experts, journalistes...) sont de plus en plus enclines à s'exprimer au travers des messages sur différentes plateformes en ligne. Twitter, qui constitue une nouvelle fois notre plateforme d'expérimentation, est un formidable objet d'étude pour de nombreux domaines de recherche, en TAL, bien entendu, mais également en sociologie [Murthy, 2012]. Dans ce travail clairement exploratoire, nous cherchons à rendre compte du contenu partagé par les utilisateurs du monde entier à travers le contenu des tweets [Quillot et al., 2017]. En plus du problème d'évaluation, une des difficultés majeures de ce type d'analyse réside dans la durée des événements considérés : bien qu'un festival se déroule sur une période définie (de quelques jours à plusieurs semaines), l'activité des utilisateurs sur les RSN peut intervenir à tout moment (avant, pendant ou après un festival).

6.3.2 Corpus multilingue de très grande taille de messages courts (*tweets*)

Dans le cadre de notre étude, nous avons utilisé le corpus fourni par le lab MC2 CLEF 2017 [4], contenant 70 millions de tweets. Ces tweets ont été collectés sur Twitter en utilisant un ensemble prédéfini de mots-clés liés aux festivals dans le monde. Ils couvrent une période allant de mai 2015 à novembre 2016 et incluent 134 langues différentes [Ermakova et al., 2016].

Aucune donnée manuellement annotée pour notre étude n'est fournie avec ce corpus. De par la

4. `http://mc2.talne.eu/`

taille énorme de ces données, nous avons dû restreindre notre champ d'étude pour la comparaison des modèles de plongements de mots à l'analyse d'une liste de mots nous apparaissant comme intéressants. Cette sélection de mots-clés a été effectuée manuellement par des experts : cette liste a été établie en partie par des sociologues spécialistes des publics des festivals. Avec un total de 119 mots-clés différents, cette liste a été construite selon trois stratégies différentes :

— Des noms de ville clairement liées à des festivals.
— Des mots génériques liés au concept de *festival*, tels que *théâtre*, *musique*, *films*...
— Des marques commerciales participant aux noms de festivals, telles que *Deezer*, *Apple*...

Ce corpus est en fait une version réduite du corpus collecté par Éric San Juan, maître de conférences au LIA, dans le cadre du projet GaFes.

6.3.3 Plongements de mots et représentation temporelle

Nous avons alors posé comme hypothèse qu'il est difficile d'extraire, et donc de rendre compte, des informations (sujets, idées...) véhiculées à travers les tweets, sans s'intéresser à l'aspect temporel de ceux-ci, qui serait ici leur date d'émission. En effet, un modèle global aurait tendance à ne révéler que les informations fréquemment partagées, en ignorant celles peu communes qui pourraient néanmoins être importantes sur une période donnée. Sur la base de cette hypothèse de travail, notre étude préliminaire, inspirée des travaux initiés dans [Basile et al., 2014, Hamilton et al., 2016], propose d'étudier le comportement de deux modèles s'appuyant sur des approches différentes de plongement de mots : une approche ignorant l'aspect temporel des messages, avec le modèle état de l'art Word2Vec [Mikolov et al., 2013a], et une seconde par plongement temporel s'appuyant sur la date d'émission des tweets.

Plongements de mots avec Word2Vec

Les modèles neuronaux Word2Vec [Mikolov et al., 2013a] ont déjà été étudiés précédemment dans la partie 1.3 pour le problème de la prise en compte du contexte d'un mot. Des détails sur Word2Vec peuvent se trouver dans la partie 1.3.2. Ces modèles se fondent sur l'hypothèse que des mots sémantiquement similaires ont tendance à avoir des distributions contextuelles proches. Dans ce travail, nous avons utilisé la méthode CBOW qui cherche à prédire un mot sachant un contexte d'apparition. Nous utilisons ici la couche cachée de ces réseaux de neurones, chaque mot étant alors représenté par un vecteur. L'approche Word2Vec a été choisie ici car, comme nous l'avons spécifié, celle-ci a été éprouvée dans de nombreuses applications, et correspond à une représentation des mots à l'état de l'art, mais sans prise en compte de l'aspect temporel.

Plongements temporels de mots

Nous avons ensuite considéré un modèle de plongement de mots pouvant intégrer l'information temporelle contenue dans notre ensemble de données. Ainsi, au lieu de prendre en compte tous les documents, comme ce qui est réalisé dans les modèles Word2Vec, nous comptons d'abord le nombre d'occurrences par unité de temps (*e.g.* année, mois, jour...) pour chaque mot de la liste. Il en résulte une matrice temporelle d'occurrences $n \times m$, où n représente le nombre de mots du vocabulaire, et m le nombre d'unités de temps. Nous effectuons ensuite une ACP sur cette matrice, qui nous fournit les plongements temporels des mots, qui sont ici les valeurs des composantes principales [Quillot et al., 2017].

6.3.4 Évaluation des modèles

Propositions

L'évaluation des modèles constitue le coeur du problème ici, puisque nous voulions connaître l'apport de l'information temporelle en comparant deux représentations différentes. Cette comparaison est généralement un problème difficile, la solution la plus fréquente étant, comme vu précédemment, de comparer leurs performances sur une tâche ciblée avec référence manuelle (par exemple, en RAP avec le WER, en classification automatique avec la précision...). Pour évaluer correctement l'impact de chaque modèle, une vérité terrain objective serait la solution idéale, c'est-à-dire ici savoir quels mots représentent clairement un événement culturel particulier.

Puisqu'aucune référence n'est disponible, nous proposons une évaluation sous deux formats. En premier lieu, nous effectuons une analyse détaillée à travers une comparaison subjective par des experts humains fondée sur une interprétation visuelle humaine à partir de représentations graphiques des mots (dendrogrammes et projection en 2 dimensions sur les deux premières composantes principales), dans le but d'étudier la contribution de l'information temporelle à la représentation d'événements culturels. En complément de cette analyse subjective humaine, nous proposons deux tests statistiques (test de Wilcoxon-Mann-Whitney et tau de Kendall).

Résultats

Les tests statistiques appliqués sur les deux modèles que nous proposons ont tout d'abord mis en évidence que le modèle Word2Vec apparaît statistiquement différent du modèle de plongements temporels de mots, ce qui semble signifier que chacun encode une information différente. Sans données de référence, il apparait difficile d'aller plus loin sur une évaluation quantifiable.

L'évaluation qualitative manuelle a alors permis de fournir une étude plus détaillée sur les mots extraits pour représenter des événements culturels. De cette étude, nous avons pu mettre en évidence que l'information temporelle jouait clairement un rôle pour la description d'un événement culturel, les modèles Word2Vec encodant naturellement la sémantique globale d'un

mot. En prenant l'exemple de l'*Apple Music Festival*, qui se déroule à Londres, les modèles temporels ont réussi à capturer cet événement, les vecteurs de ces trois mots étant très proches, ce qui n'était pas le cas pour Word2Vec, qui avait tendance à rapprocher des mots liés à des plateformes musicales, telles que *Spotify* [Quillot et al., 2017].

Bien conscient des limites de cette étude, elle montre les difficultés que nous rencontrons dans les recherches à la frontière de plusieurs disciplines. Nous ne sommes pas ici dans des tâches classiques en TAL, qui supposent souvent un cadre d'étude clair et bien défini, ni sur des recherches purement sociologiques, qui supposent un contrôle dans la collecte des informations pour pouvoir en extraire des conclusions solides. Dans la partie suivante, nous avons continué sur des travaux proches, avec un cadre de travail mieux structuré par la communauté scientifique, qui a pris conscience du besoin d'un *framework* pour traiter ces types de problématiques.

6.4 Argumentation et diversité des opinions par les utilisateurs de réseaux sociaux

6.4.1 Campagne d'évaluation CLEF

Les travaux préliminaires présentés dans les parties 6.2 et 6.3, et réalisés dans un contexte interdisciplinaire, ont finalement soulevé des questionnements, en particulier sur la manière de rendre compte des performances des systèmes de traitement automatique proposés. Le cadre des RSN, de par la nature de ces données massives, rend l'extraction de connaissances beaucoup plus difficile que sur des tâches traditionnelles en TAL et RI. Cette difficulté est principalement liée au manque de cadres de travail, ce que les tâches historiques en TAL possèdent depuis de très nombreuses années [Jones and Galliers, 1995] grâce aux campagnes d'évaluation organisées régulièrement, fournissant, entre autres, données et protocoles d'évaluation communs (par exemple en TAL, les évaluations NIST en reconnaissance du locuteur, les campagnes ESTER en RAP...).

Dans le domaine de la RI, la conférence annuelle CLEF, organisée depuis 2000, fait partie des références, en plus de la conférence principale, sur l'organisation de campagnes d'évaluation, avec cadres de travail communs, sur des tâches nouvelles. Ces dernières années, des tâches à la coloration RI et sciences sociales sont apparues, telles que la contextualisation de micro-blogs dans les événements culturels [Ermakova et al., 2016, 2017]. Il s'agissait ici de fournir un cadre expérimental sur des tâches en manquant cruellement. Dans le projet GaFes, avec l'équipe de sociologues d'Avignon Université sur le projet GaFes (Emmanuel Ethis, Damien Malinas, Raphaël Roth, Stéphanie Pourquier-Jacquin et Alexandre Delorme), nous avons proposé un travail original commun, entrant dans le cadre de la tâche *Mining Opinion Argumentation*, incluse dans le lab *MC2-Multilingual Cultural Mining and Retrieval*, de la conférence CLEF 2018 [Hajjem et al., 2018]. Cette tâche nous intéressait pour deux raisons principales : 1) travailler sur un cadre

d'évaluation complet pour comparer nos approches avec d'autres laboratoires internationaux ; 2) travailler sur les données du projet GaFes, que nous avons déjà manipulées dans la partie 6.3.2, et dans lequel nous étions tous impliqués.

Cette tâche de fouille d'opinions avec argumentation visait à identifier automatiquement, dans des messages courts exprimés via la plateforme Twitter, des positions des internautes sur un événement culturel. L'idée était de trouver des opinions sur un festival de manière générale, ou sur un sujet particulier, à partir d'une énorme collection de tweets. Nous devions fournir, à partir d'une requête, des informations pertinentes exprimées sous la forme d'un résumé de tweets argumentés devant refléter un maximum de points de vue différents (*i.e.* éviter la redondance d'arguments dans les tweets extraits automatiquement). Une des difficultés résidait dans le fait qu'aucun ensemble de données annotées manuellement pour entraîner des modèles n'était fourni.

6.4.2 Approche non supervisée pour l'extraction des messages pertinents

Nous avons proposé un système qui comprend quatre étapes [Dufour et al., 2018]. De façon assez classique en TAL, la première étape réalise un pré-traitement sur les messages *bruts* afin de les rendre plus facilement interprétables et généralisables par un processus automatique. La deuxième étape, dont nous donnons quelques détails dans la suite de cette partie, prend en entrée les données propres et propose une méthode pour extraire deux ensembles de données de façon non supervisée (*argumenté* et *non argumenté*) alors qu'aucune donnée étiquetée n'est fournie. À partir de ces deux ensembles, un réseau de neurones convolutifs (CNN) est entraîné à l'étape 3 pour reconnaître les messages argumentés et non argumentés. Enfin, la dernière étape cherche à extraire, à partir d'un ensemble de messages liés à une requête, ceux contenant les éléments les plus argumentés tout en intégrant un maximum de diversité dans les opinions véhiculées au moyen de notre système à base de CNN.

Sélection dans les données non-annotées

Puisqu'aucune donnée de référence annotée n'était disponible, nous avons proposé de *déduire* ces données au moyen d'une approche semi-supervisée. Nous avons utilisé des listes de mots définissant des opinions, comme par exemple le lexique français d'émotions FEEL [Abdaoui et al., 2017]. Cela reste finalement assez classique pour traiter ce type de problème (détection d'opinions, de sentiment...).

L'originalité de cette partie se situe plutôt dans l'introduction de connaissances supplémentaires pour la définition du corpus en deux ensembles *argumenté* et *non argumenté*. Nous nous sommes appuyés sur l'expérience de sociologues pour introduire de nouvelles informations sur la définition de messages argumentés dans le contexte de festivals (par exemple, l'utilisation de pronoms particuliers, des mots-clés identifiés...). L'ensemble de ces règles ont permis de séparer les messages argumentés de ceux non argumentés.

Apprentissage par réseau de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) représentent l'un des modèles de réseaux neuronaux profonds les plus utilisés, à l'origine, en reconnaissance d'images. De nombreux travaux récents ont montré que les CNN peuvent être adaptés pour traiter des problèmes de classification de documents à partir de leur contenu, ayant permis, dans plusieurs domaines, d'atteindre des performances devenues état de l'art, comme par exemple dans l'analyse de sentiments dans le texte [Tang et al., 2014, Rouvier and Favre, 2016].

La différence entre les CNN appliqués aux images et leur équivalent en TAL réside dans la dimensionnalité et le format d'entrée. En vision par ordinateur, les entrées sont généralement des matrices 2D ou 3D à canal unique (par exemple, en niveaux de gris) ou à canaux multiples (par exemple, RVB), souvent de dimension constante. Dans la classification de messages, chaque entrée consiste en une séquence de mots de longueur variable. Chaque mot w est représenté par un vecteur à n dimensions (souvent, un plongement de mot) de taille constante. Toutes les représentations des mots du message sont ensuite concaténées dans leur ordre respectif et complétées avec des vecteurs nuls (*padding*) à une longueur fixée (longueur maximale possible du message). Un CNN est alors entraîné, dans notre contexte, à reconnaître les messages *argumentés* de ceux *non argumentés*.

Diversité des messages

Enfin, la liste des messages argumentés, non incluse dans les données d'apprentissage non supervisé, est constituée en deux étapes :

1. Un score est attribué à chaque message grâce au CNN préalablement entraîné. Une première liste classée peut alors être obtenue avec ce processus de classification.
2. Un score de similarité cosinus est calculé entre un message et le reste des messages de la liste, en partant des messages considérés comme les plus argumentés dans la liste précédente. Les messages de la liste considérés comme trop similaires au message initial (*i.e.* score trop proche) sont alors retirés.

Cela permettait, au final, d'obtenir une liste de messages ordonnés selon leur niveau supposé d'argumentation, tout en espérant une diversité importante au niveau des idées véhiculées dans les messages.

6.4.3 Évaluation par des experts humains

Métrique

Les organisateurs de la tâche ont fourni une métrique commune, s'appuyant à la fois sur une annotation humaine et sur des métriques. Devant la difficulté d'annoter manuellement des

millions de tweets, les organisateurs ont proposé de prendre en considération un ensemble réduit composé de l'ensemble des résultats fournis par les participants à la campagne. De cet ensemble, ils ont donné une liste, pour chaque requête, des tweets les plus informatifs. La métrique choisie est le Normalized Discounted Cumulative Gain (NDCG), une mesure commune dans les tâches de RI. Cette mesure part du principe que les messages les plus intéressants (dans notre cas, ceux considérés comme les plus argumentés) devraient apparaître en premier dans la liste tandis que les messages non pertinents (moins ou pas argumentés) ne devraient pas apparaître (ou au rang le plus bas possible). Globalement, plus la mesure est élevée, meilleurs sont les résultats.

Discussions

Cette campagne d'évaluation CLEF a permis de confronter notre système à d'autres participants, dans un cadre de travail identique et contrôlé. Les résultats obtenus ont permis de mettre en évidence que le système conçu, entre approches statistiques et apports d'experts en sciences humaines, paraît proposer une plus grande diversité dans les messages extraits, au contraire des approches des autres participants. En effet, par rapport à la référence humaine, notre système se classe premier, tout en ayant des résultats différents des autres participants.

La structuration de ces travaux interdisciplinaires, concrétisés ici au travers d'une campagne d'évaluation, a été une étape importante dans le projet GaFes, et dans mon travail de recherche. Bien entendu, la façon d'évaluer ces travaux reste ouverte, et les métriques proposées ici sont tout autant sujettes à discussion et critique. Néanmoins, il semble qu'une étape supplémentaire soit franchie dans les communautés TAL et RI, que ce soit de par les types de projets financés, dont nous reparlerons dans la partie 8, ou de par ces campagnes d'évaluation.

6.5 Conclusion

Ce chapitre résume les travaux que nous avons réalisés ces dernières années sur des problématiques interdisciplinaires, avec en point d'orgue, la difficulté d'évaluer des approches de TAL et RI dans un contexte de fouille de données textuelles massives et d'extraction d'informations qualitatives pour enrichir les travaux de recherche en sciences humaines. L'étude qualitative sur la recherche de caractéristiques importantes pour comprendre la mécanique de la diffusion massive de messages sur le réseau social Twitter a constitué un premier changement dans les problématiques en traitement du langage sur lesquelles je travaillais auparavant. Cette étude, proposée dans le cadre de la thèse de M. Morchid, a été réalisée en deux étapes. La première partie a consisté à étudier la corrélation entre différentes caractéristiques, liées au message lui-même ainsi qu'à l'utilisateur, déjà proposées dans plusieurs autres travaux comme devant avoir une influence dans le mécanisme de relais de l'information (*retweet*). Cela a permis de mettre en lumière certaines caractéristiques apparaissant comme plus descriptives et informatives sur

le retweet massif. Cette partie purement descriptive a pu ensuite être confirmée sur une tâche de classification automatique des messages, devant déterminer si le message faisait partie de la classe *faiblement relayé* ou *massivement relayé*.

Les travaux réalisés dans le cadre du projet ANR GaFes ont amené de nouvelles réflexions, et surtout de nouvelles difficultés, que je n'avais encore que peu explorées. En effet, ces premiers travaux sur le mécanisme du retweet massif se trouvaient, au final, dans un contexte bien structuré, avec analyse et évaluation sur une tâche classique en TAL. Le travail sur l'analyse de l'intérêt d'intégrer la temporalité dans les plongements lexicaux pour décrire des événements culturels a bien montré les limites de travaux interdisciplinaires lorsque le cadre expérimental est inexistant : bien que nous ayons pu mettre en lumière, au moyen d'analyses humaines, l'intérêt de la prise en compte d'informations temporelles, nous n'avons pu réellement conclure de la manière dont nous pouvions le faire dans les travaux présentés auparavant, c'est-à-dire sur une tâche avec jeu de test sur lequel nous pouvions évaluer plusieurs modèles.

Ce problème de reproductibilité des résultats a pu être en partie résolu avec le dernier travail présenté sur la tâche d'extraction de tweets argumentés. Ce travail a eu lieu durant la campagne d'évaluation internationale issue du projet GaFes et qui s'est déroulée pendant la conférence CLEF 2018. Nous avons proposé une approche non supervisée, intégrant des connaissances sociologiques des festivals, grâce au travail de chercheurs d'Avignon Université en sociologie des publics des festivals. Même si cela est une avancée, ces cadres de travail restent néanmoins encore peu nombreux, mais montrent la volonté d'une communauté scientifique pluridisciplinaire de structurer des recherches à la frontière de plusieurs thématiques. Il convient donc de soutenir ces problématiques et ces cadres expérimentaux, toujours fragiles car encore peu présents et moins structurés que des tâches éprouvées. Cette volonté d'une recherche interdisciplinaire structurée se trouve au centre des perspectives de recherche dans lesquelles je vais m'inscrire, comme je le développe dans la partie V. Ces perspectives trouvent notamment leur echo dans le projet ANR JCJC DIETS (voir partie V), qui est la suite naturelle de mes orientations scientifiques. Dans cette continuité, le chapitre suivant expose nos travaux entre TAL et réseaux complexes pour la détection de messages abusifs.

Chapitre 7

Structure des échanges pour la détection de messages abusifs

Sommaire

7.1 Introduction

Les réseaux sociaux numériques (RSN), comme nous l'avons vu dans le chapitre 6, constituent un formidable terrain d'étude dans des domaines de recherche très variés, allant de l'informatique aux sciences sociales. Grâce à ces RSN, qui au départ agrégeaient des communautés relativement modestes, des échanges entre personnes sont maintenant devenus possibles au niveau mondial, et dans une proportion de plus en plus importante. Les utilisateurs de ces RSN se réunissent alors au sein de *communautés*, dans le sens où ils se regroupent autour d'un objet d'échange commun, pouvant prendre la forme de messages textuels, de vidéos, d'images... pour, par exemple, informer, débattre, ou simplement discuter autour d'un sujet. Ces communautés en ligne ont acquis une importance que l'on ne peut nier dans la société actuelle, ne serait-ce que par le nombre

d'utilisateurs quotidiens de ces plateformes d'échange et de relais, potentiellement massifs, d'informations (voir partie 6.1). Ils ont un impact social de plus en plus important [Siddiqui et al., 2016], que ce soit dans le cadre de communications publiques (*i.e.* une diffusion sans restriction, accessible au plus grand nombre) ou dans des échanges plus restreints, entre petits groupes de personnes (par exemple, un cadre familial, entre amis, ...).

De par les traces numériques que ces réseaux laissent, le milieu scientifique n'est pas le seul à se pencher sur ces RSN : les gouvernements s'intéressent, entre autres, à la surveillance de ces discours publics, s'intéressant à la légalité des échanges réalisés, comme s'assurer du respect des droits d'auteur, ou encore être certain que les messages respectent la loi, comme pour tout autre forme d'échange entre personnes et groupes (pas de discours haineux, diffamatoires, racistes...). Cette tâche de surveillance des RSN est plus difficile que sur des médias classiques, en particulier du fait de l'anonymat rendu – plus ou moins – possible avec Internet. De même, l'évolution rapide de ces pratiques d'échange rend souvent la loi difficilement applicable avec les textes dont la justice dispose. Il est souvent nécessaire de les faire évoluer (par exemple, la lutte contre les discours haineux est une proposition de loi ne datant que de l'année 2019 [1], promulguée le 24 juin 2020 [2], mais dont une grande partie a été censurée par le Conseil Constitutionnel [3]).

Même sans enfreindre la loi, comme dans toute communauté, des utilisateurs peuvent simplement nuire à son bon fonctionnement. Pour les responsables de ces communautés en ligne, il leur est indispensable d'agir et de traiter ces comportements que l'on pourrait qualifier d'abusifs. En effet, si les responsables ne le font pas, ces *mauvais* utilisateurs fragilisent la cohésion de la communauté, nuisent aux échanges et à la bonne entente du groupe, peuvent *casser* la communauté en faisant fuir certains membres ; et ce, sans parler des problèmes juridiques auxquels les responsables s'exposent [4]. Outre la loi, à laquelle aucune plateforme ne peut déroger, la notion d'*abus* a tendance à varier selon la plateforme d'échange. Il existe cependant presque toujours un noyau commun de règles à respecter (respect entre membres, interdiction d'écrits *violents*, langage correct...). Les responsables veillant à ces échanges sont appelés des *modérateurs*. Ils doivent donc s'assurer que les contenus échangés respectent à la fois les règles édictées par les plateformes et par les gouvernements des différents pays dans lesquels celles-ci sont accessibles. Lorsque les règles sont enfreintes, ils sont alors chargés d'appliquer des sanctions lors d'un processus appelé *modération*.

Dans cette partie, nous nous sommes intéressés aux abus dans les messages textuels sur les RSN. Le travail de modération est pour l'instant principalement réalisé par des modérateurs *humains*. De par la taille des données échangées via certaines plateformes en ligne, il est clair

1. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000038758354
2. https://www.legifrance.gouv.fr/eli/loi/2020/6/24/JUSX1913052L/jo/texte
3. https://www.huffingtonpost.fr/entry/la-loi-avia-un-combat-legitime-avorte-en-raison-de-la-methode_fr_5efca19dc5b612083c55a4fd
4. https://www.service-public.fr/particuliers/vosdroits/F32075

que ce processus manuel est très coûteux (ou quasiment impossible dans certains cas), et qu'il y a un réel intérêt, pour les administrateurs de ces plateformes, a avoir des outils automatiques les aidant dans cette tâche. Les travaux que nous avons menés ici ont suivi deux grands axes de travail, portant au final sur des domaines de recherche différents, mais complémentaires. Ils sont le fruit de la collaboration principale avec un collègue MCF au LIA, Vincent Labatut, dont le domaine d'expertise se concentre sur les réseaux complexes, ainsi que les travaux entrepris par Etienne Papegniès au LIA.

De façon assez naturelle, nous avons tout d'abord proposé de travailler sur la détection d'abus dans les RSN en ne prenant en compte, simplement, que le contenu textuel des documents échangés, ce que nous détaillons dans la partie 7.2. Nous avons alors proposé d'extraire différentes caractéristiques classiques en TAL et de les utiliser dans un processus de classification afin de déterminer, pour chaque message, si celui-ci est abusif ou non [Papegnies et al., 2017b]. Ces travaux préliminaires nous ont amenés à nous questionner sur la pertinence, ou tout du moins, sur les limites d'utiliser le contenu pour cette tâche particulière. Comme nous le détaillons dans la partie 7.3, nous avons ensuite choisi une toute autre voie d'étude, en considérant la modélisation des interactions entre utilisateurs dans une conversation pour y détecter les messages abusifs, ignorant alors complètement le contenu textuel [Papegnies et al., 2017a, 2019]. Nous avons ensuite entrepris d'étudier l'apport de chaque modalité et de leur complémentarité dans le cadre du master, puis de la thèse, de Noé Cécillon [Cecillon et al., 2019] (voir partie 7.4). Enfin, dans la partie 7.5, nous avons proposé un cadre de travail expérimental ainsi que l'enrichissement en conversations d'un corpus existant de messages abusifs sur Wikipedia [Cecillon et al., 2020b]. Finalement, ce travail, qui comprend l'expertise de plusieurs domaines de recherche (TAL et réseaux complexes), suit l'idée que j'avais pu débuter pendant mon année d'ATER pour améliorer la détection de rôles des locuteurs lors de débats télévisés [Dufour et al., 2012d].

7.2 Utilisation du contenu des messages

De façon assez classique, la détection d'abus dans des messages a tout d'abord été vue comme un problème en traitement automatique du langage. Ainsi, de nombreux travaux se sont concentrés sur la proposition de solutions pour traiter le contenu textuel véhiculé. Parmi ces travaux, nous pouvons citer par exemple ceux de [Spertus, 1997, Dinakar et al., 2011, Chen et al., 2012], qui traitent de sujets proches de la détection d'abus (messages offensants, harcèlement, hostilité...) au moyen de caractéristiques extraites automatiquement du texte du message, et/ou de règles manuelles. Des approches ont également proposé de s'intéresser au contexte autour du message pour détecter le harcèlement en ligne [Yin et al., 2009]. Enfin, des approches plus récentes ont proposé d'utiliser des méthodes d'apprentissage par réseaux de neurones [Pavlopoulos et al., 2017, Mishra et al., 2018], mais qui nécessitent souvent de grandes quantités de données

d'apprentissage.

Dans le cadre de nos travaux, nous avons proposé une approche utilisant des caractéristiques extraites automatiquement du texte relativement courantes en TAL. Nous avons ensuite utilisé ces caractéristiques linguistiques en entrée d'un classifieur afin de détecter si un message est abusif ou non. Ce premier travail a constitué notre *baseline* afin d'avoir une première mesure sur la détection de messages abusifs sur des données échangées sur Internet. Nous présentons, dans les sous-parties suivantes, certaines caractéristiques morphologiques extraites, ainsi que d'autres liées à la langue (ici, principalement sur la représentation des mots). Pour plus de détails, ainsi que la proposition d'une caractéristique liée au contexte des messages, le lecteur pourra se référer à [Papegnies et al., 2017b]. Les résultats expérimentaux obtenus seront ensuite présentés, avec les autres approches développées, dans la partie 7.4.

7.2.1 Caractéristiques morphologiques

Il s'agissait ici d'extraire des informations liées à la construction des mots et des messages, en considérant, dans une certaine mesure, que nous étions indépendants de la langue. En effet, nous voulions mesurer une *façon* d'écrire des messages, au lieu de modéliser ici les mots (et leur sémantique associée). Nous avons alors proposé d'utiliser la longueur du message, et la longueur moyenne et maximale des mots, exprimées ici en nombre de caractères. De même, nous avons extrait le nombre de caractères uniques dans le message. Plusieurs classes de caractères ont été identifiées (lettres, chiffres, ponctuation, espaces, majuscules et autres) pour nous permettre de calculer, pour chacune, le nombre d'occurrences et la proportion de caractères dans le message.

En partant du postulat que les messages abusifs contiennent de nombreux *copier/coller*, nous avons appliqué l'algorithme de compression Lempel–Ziv–Welch sur le message, et avons calculé le ratio entre le nombre de caractères initiaux et le nombre compressé. Enfin, nous avons remarqué que, dans les messages abusifs, certains mots avaient tendance à être allongés, surtout pour insister sur un mot en particulier (par exemple, *le groooos naaaaaaaaze*). Nous avons alors choisi de supprimer, dans les mots, les lettres consécutives qui apparaissent plus de deux fois, mas également de calculer la différence entre les longueurs des message bruts et réduits.

7.2.2 Caractéristiques liées à la langue

La deuxième catégorie que nous avions identifiée concernait les caractéristiques directement liées à la langue et à la représentation des mots et des documents. Nous avons ainsi proposé de compter le nombre d'occurrences de mots, de mots uniques et de mots identifiés comme abusifs (liste prédéfinie d'insultes et de symboles considérés comme abusifs). Nous avons également choisi de calculer deux scores globaux avec l'approche TF-IDF, correspondants aux sommes des scores TF-IDF de chaque mot du message. Un score est alors attribué relativement à la classe *Abus* et l'autre à la classe *Non-abus*. Ces scores ont également été calculés sur les messages

réduits (comme vu dans la partie précédente, en retirant les lettres identiques apparaissant successivement dans les mots). Enfin, le texte a été mis en minuscule et nettoyé de sa ponctuation pour en créer un sac-de-mots. Ce sac-de-mots est ensuite utilisé dans un classifieur Naïf Bayésien pour prédire la classe du message (*Abus* ou *Non-abus*) : le score obtenu sera ajouté aux autres caractéristiques en entrée d'un autre classifieur pour prendre la décision finale.

7.3 Modélisation de la structure des conversations

Bien que l'analyse du contenu textuel soit l'approche couramment suivie pour traiter ce type de problème, nous avons rapidement eu l'impression que le contenu seul ne pouvait fournir une réponse complètement efficace, et ce pour les raisons suivantes :

- Les utilisateurs sont maintenant sensibilisés aux possibilités des outils de détection automatique, et peuvent donc masquer leur *attaque*. Les systèmes les plus basiques s'appuyant simplement sur des bases de mots-clés identifiés comme abusifs, remplacer certains caractères suffit à *tromper* les systèmes (par exemple, *ba*ard* ou *samair*), mais restent toujours compréhensibles par des humains.
- Les outils en TAL et RI ont souvent des difficultés à traiter des messages courts sur Internet (manque de contexte, agrammaticalité...).
- Les données d'apprentissage sont généralement insuffisantes pour avoir des modèles assez robustes puisque nous sommes face à des langages et vocabulaires souvent non-standards, incluant de nombreux problèmes grammaticaux et orthographiques, et sur des domaines très variés.

Ainsi, les travaux dans [Hosseini et al., 2017] ont montré qu'il était très facile de contrer des systèmes de détection de messages dits *toxiques*, ici l'API Google Perspective en 2017, en modifiant simplement quelques mots dans la phrase. Les auteurs avaient intentionnellement ajouté des erreurs orthographiques ou des tournures de phrases négatives, menant à rendre le système inopérant.

Les travaux que nous avons débutés avec E. Papegniès et V. Labatut, et que nous continuons toujours avec N. Cécillon, se situent dans le domaine des réseaux complexes, séparant alors l'analyse du contenu textuel de la modélisation de la dynamique des messages échangés. Ils constituent une des principales nouveautés que nous avons proposées pour détecter les abus dans des messages sur Internet. Dans ce que nous présentons dans cette partie, nous avons choisi de complètement ignorer le contenu textuel et de modéliser, au travers de graphes conversationnels, les interactions entre les participants d'une conversation.

L'avantage principal de ne considérer que les interactions entre les utilisateurs est que cette approche n'est pas dépendante d'une langue particulière, et évite tous les problèmes classiques que nous devons gérer en TAL (et que nous avons mis en avant, au travers de solutions propo-

sées, tout au long de ce manuscrit). L'approche que nous avons proposée pour la détection des messages abusifs à partir des interactions des utilisateurs tient en trois étapes :

1. Un graphe conversationnel est extrait pour chaque message au moyen des messages à son voisinage (*i.e.* les messages précédant et/ou suivant le message considéré).
2. Un ensemble de mesures topologiques, pour le message considéré, sont calculées à partir de son graphe conversationnel.
3. Un classifieur, prenant en entrée les différentes mesures topologiques extraites pour chaque message, permet de décider automatiquement si ce message est abusif ou non.

Dans la suite de cette partie, nous présentons tout d'abord brièvement l'approche d'extraction des graphes conversationnels que nous avons proposée (sous-partie 7.3.1). Nous présentons également les mesures topologiques que nous avons choisi d'utiliser, en particulier les catégories auxquelles elles appartiennent ainsi que les informations qu'elles modélisent (sous-partie 7.3.2).

7.3.1 Extraction des graphes conversationnels

La figure 7.1 présente le principe général d'extraction du graphe conversationnel associé à chaque message. À gauche de la figure, nous avons fourni un extrait d'une conversation, transformé ensuite, dans la partie de droite, en un graphe conversationnel représentant les interactions entre les participants. Dans le graphe, un noeud représente un utilisateur, et un lien, une interaction entre deux utilisateurs dans la conversation. Le noeud entouré en rouge correspond au message, et donc son utilisateur/participant, que l'on souhaite identifier comme abusif ou non (les autres noeuds et liens étant les utilisateurs au voisinage du message ciblé). Notons qu'ici, nous ne représentons pas les poids ni les directions des liens (graphe non-orienté), mais ce sont des informations qui ont eu une importance dans notre tâche de détection des messages abusifs [Papegnies et al., 2019].

Dans un cas idéal, nous aurions accès aux échanges *réels*, *i.e.* les conversations seraient structurées afin de savoir qui répond à quel message. Malheureusement, dans le cas de messages échangés sur des plateformes en ligne, cette structuration est souvent soit incomplète, soit inexistante. Dans notre cadre expérimental, qui consistait en des échanges entre utilisateurs sous forme de messages instantanés (*chat messages*), nous n'avions aucune information structurelle des échanges. Cette contrainte a conditionné l'approche que nous avons proposée pour la construction des graphes conversationnels.

La figure 7.2 présente les différents éléments qui nous ont permis de proposer une méthode de construction de graphes conversationnels en partant du message que l'on souhaite identifier comme abusif ou non (en rouge sur la figure). En particulier, deux éléments sont importants et ont une influence directe sur les graphes qui seront, au final, modélisés :

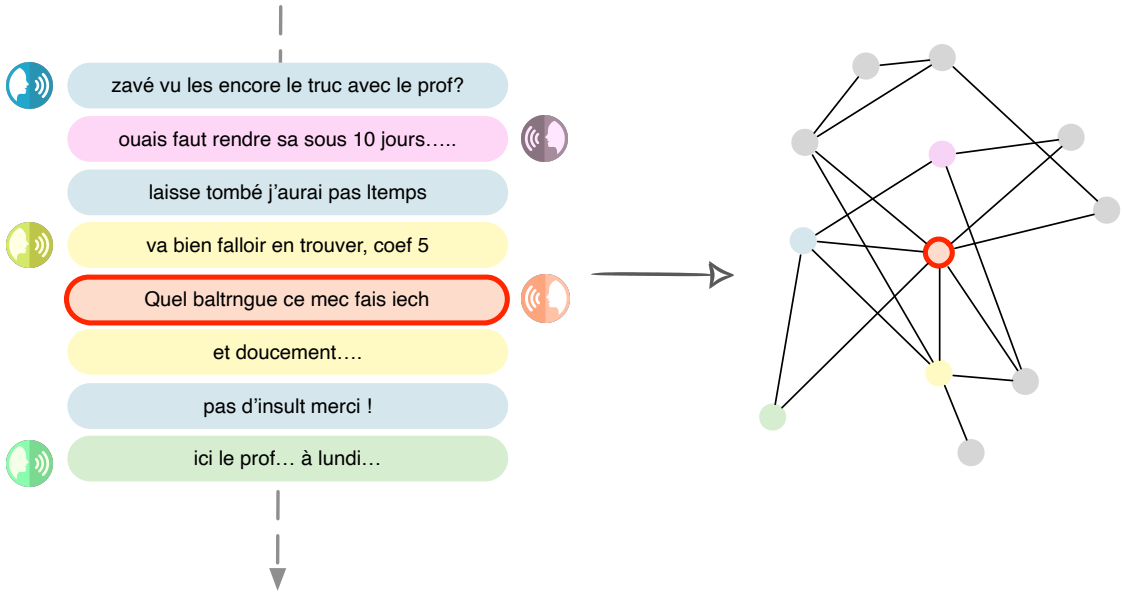

FIGURE 7.1 – Exemple d'extraction du graphe conversationnel d'un message d'une conversation. Le noeud avec cercle rouge représente le message ciblé (l'utilisateur) ; les noeuds en couleur, les utilisateurs participant à la conversation ; les noeuds gris, les autres utilisateurs non-visibles dans l'exemple.

— *Période de contexte.* Les messages pouvant être diffusés sous forme d'un flux continu, nous avons choisi de restreindre la construction d'un graphe à un sous-ensemble de messages dont nous définissions, *a priori*, la taille. Cette période englobe donc le message ciblé, ainsi que son contexte précédant et suivant de façon symétrique.

— *Fenêtre glissante.* Comme nous l'avions évoqué précédemment, il est très difficile de savoir qui discute avec qui dans ce genre de conversations, où les sujets de discussion peuvent s'entre-croiser, sans aucune indentation dans les réponses. Nous avons alors choisi d'utiliser, au sein de cette période de contexte, une fenêtre glissante sur laquelle nous avons un processus de décision permettant de lier des utilisateurs entre eux (*i.e.* les noeuds) et éventuellement d'y associer un poids selon notre confiance dans le lien existant (proximité des échanges, autre utilisateur nommé dans le message...). Il s'agit ici de construire et/ou de mettre à jour les liens du message courant (en bleu dans la figure 7.2) dans chaque fenêtre glissante (un pas de 1 message pour passer à la prochaine fenêtre).

Ces différents éléments sont, pour la plupart, paramétrables (taille du contexte, de la fenêtre glissante, poids sur les liens ou non...). Cela nous a notamment permis d'étudier la modélisation de trois types de graphes conversationnels, à savoir un graphe construit simplement sur l'historique des messages, un graphe construit sur les messages apparaissant après le message ciblé, et, bien entendu, le graphe avec le voisinage complet. De plus amples détails sur la construction des

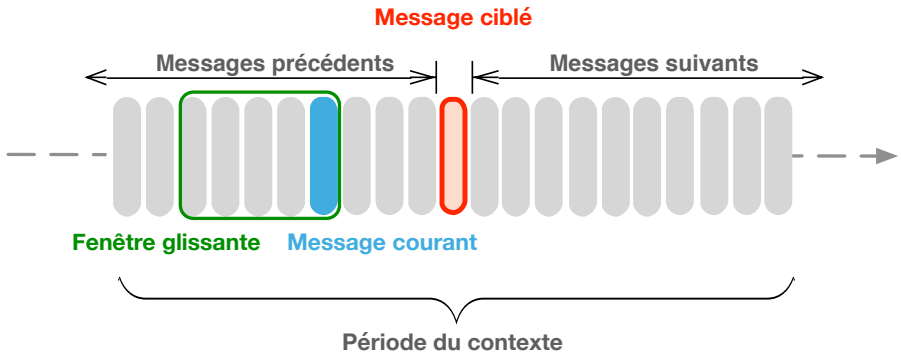

FIGURE 7.2 – Description des éléments constitutifs à la construction du graphe conversationnel d'un message (rectangle en rouge) en fonction des messages dans le contexte (rectangles en gris).

graphes conversationnels, et en particulier sur la création des liens entre utilisateurs et le poids associé, peuvent se trouver dans [Papegnies et al., 2019].

7.3.2 Caractéristiques topologiques

La seconde partie importante de notre approche a concerné l'extraction d'informations à partir de ces graphes conversationnels. Pour ce faire, nous avons défini, manuellement, un ensemble de mesures topologiques standards dans le domaine de l'étude des graphes et qui nous semblaient les plus pertinentes pour décrire, de la façon la plus complète possible, les informations contenues dans un graphe. L'approche par graphe que nous avons proposée étant totalement nouvelle pour traiter le problème de la détection d'abus dans des conversations, nous avons essayé d'être le plus large possible au niveau des caractéristiques extraites (ici, les mesures). Nous nous sommes alors concentrés sur différentes échelles (*scale*) et portées (*scope*).

L'échelle (*scale*) correspond à la nature de l'entité que nous souhaitons caractériser (niveau noeud, sous-graphe, ou graphe complet). Dans notre travail, nous avons proposé de nous concentrer sur des mesures topologiques au niveau du noeud et du graphe complet. À noter que lorsque nous nous sommes intéressés au niveau du noeud, les mesures ont été extraites au niveau de l'auteur ciblé (*i.e.* le message que nous souhaitons identifier comme abusif ou non, en rouge dans les figures 7.1 et 7.2).

La portée (*scope*), quant à elle, correspond à la quantité d'information utilisée pour caractériser l'entité. Trois portées sont possibles au sein d'une échelle :

— *Microscopique* : interconnexion entre un noeud et ses voisins directs.
— *Macroscopique* : structure d'un sous-graphe et de ses voisins directs.
— *Mésoscopique* : structure du graphe complet.

Ainsi, pour chaque échelle (noeud ou graphe complet), nous avons considéré des mesures

intégrant des informations issues de ces trois portées. Le détail de l'ensemble des mesures topologiques extraites pour caractériser le graphe peut se trouver dans [Papegnies et al., 2019].

7.4 Détection des messages abusifs

La détection des messages abusifs a alors consisté en une simple tâche de classification automatique, permettant d'associer, à chaque message, la classe *Abus* ou *Non-abus*. Les différentes expériences ont été menées sur un corpus de conversations instantanées entre utilisateurs dans le jeu en ligne Space Origin. Ce corpus est issu de la collaboration avec l'entreprise Castle Prod, avec qui nous avons travaillé sur les outils de détection d'abus. Le protocole expérimental, incluant une description du corpus Space Origin, est alors présenté dans la prochaine sous-partie 7.4.1. Nous nous intéressons ensuite, dans la sous-partie 7.4.2, aux résultats obtenus par chaque approche (texte et graphe) de façon indépendante afin de pouvoir comparer leurs performances entre elles. Puis nous avons mené différentes expériences afin de montrer la complémentarité des approches, en fusionnant notamment différents niveaux d'informations (sous-partie 7.4.3). Enfin, nous voyons, dans la sous-partie 7.4.4, les résultats d'une étude que nous avons entreprise pour retrouver les caractéristiques des messages ayant un impact important sur la classification en classe d'abus, permettant d'avoir une meilleure compréhension des informations essentielles pour détecter des messages abusifs.

7.4.1 Protocole expérimental

Corpus Space Origin

Nous avons eu accès à une base de données d'environ 4 millions de messages instantanés, en français, issus de la communauté d'utilisateurs du jeu de rôle en ligne massivement multijoueur Space Origin. Au sein de ce jeu de données, 779 messages ont été annotés comme des abus, et confirmés comme tels par des modérateurs humains. Ces messages abusifs font donc partie de notre classe *Abus*. Suite à quelques problèmes identifiés dans la base de données (problèmes de dates), notre classe d'abus ne contenait, au final, que 655 messages. Afin d'avoir un corpus de données équilibré, nous avons également constitué un ensemble de données pour la classe *Non-abus* en échantillonnant au hasard le même nombre de messages pour ceux qui n'ont pas été signalés comme abusifs, avec la contrainte qu'un message de cette classe ne doit pas apparaître dans la même conversation qu'un message déjà sélectionné. Au final, notre corpus de données était composé de 1 310 messages, répartis de façon égale entre les classes *Abus* et *Non-abus*.

Paramètres

Les caractéristiques pour l'approche s'appuyant sur le contenu textuel ont été extraites au moyen de la librairie Sklearn [Pedregosa et al., 2011]. La construction des graphes conversationnels a été développée par E. Papegniès, intégrant les différents paramètres évoqués dans la partie 7.3.1. Dans les expériences que nous présentons dans ce manuscrit, les valeurs de ces paramètres ont été choisies selon les meilleures performances obtenues dans [Papegnies et al., 2019] : la taille de la période du contexte a été définie à 1 310 (taille maximum possible), la fenêtre glissante à 10 messages, et les liens sont orientés et ont un poids selon l'importance des interactions entre utilisateurs. Les mesures topologiques ont enfin été extraites des graphes complets (*i.e.* messages avant et après le message considéré) au moyen de la librairie iGraph [Csardi et al., 2006].

Pour la classification, nous avons utilisé le classifieur SVM implémenté dans Sklearn sous le nom SVC (C-Support Vector Classification). En raison de la faible taille des données, nous avons mis en place une validation croisée à 10 plis. Chaque pli est équilibré entre les classes *Abus* et *Non-abus*, 70 % de l'ensemble de données étant utilisés pour l'apprentissage et 30 % pour le test.

Plateforme open-source

Dans le cadre de la thèse de N. Cécillon, nous avons proposé à la communauté scientifique une plateforme *open-source*, disponible en ligne [5], regroupant les méthodes de classification que nous avons développées pour détecter automatiquement les abus dans les conversations. Nous pensons que cette plateforme sera utile pour l'avancée du domaine, dans un souci d'échange et de réplicabilité des expériences.

7.4.2 Évaluation indépendante des approches

Le tableau 7.1 présente les résultats obtenus pour chaque approche proposée, à savoir l'approche utilisant des caractéristiques liées au contenu textuel (*Approche texte*) [Papegnies et al., 2017b] et celles s'intéressant à la modélisation des interactions entre utilisateurs (*Approche graphe*) [Papegnies et al., 2017a, 2019]. Afin de rendre les résultats comparables dans les expériences présentées dans ce manuscrit, les performances sont issues de [Cecillon et al., 2019].

De cette expérience, nous avons pu observer que l'approche par graphes conversationnels surpasse très nettement l'approche s'appuyant sur le contenu textuel. Ce résultat, qui peut paraître assez inattendu, peut s'expliquer par plusieurs facteurs. Tout d'abord, la talle des données étant relativement petite, il semble évident que notre approche TALN supervisée ne possède pas assez d'exemples d'entraînement pour traiter tous les cas d'abus et modéliser, en particulier,

5. https://github.com/CompNet/Alert

Méthode	Précision	Rappel	F-mesure
Approche texte	78,59	83,61	81,02
Approche graphe	**90,21**	**87,63**	**88,90**

Tableau 7.1 – Rappel, précision et F-mesure (%) pour les approches utilisant des caractéristiques issues du contenu textuel (Approche texte) et du graphe conversationnel (Approche graphe).

le langage spécifique associé. De même, l'approche suivie utilise finalement des caractéristiques assez classiques, et qui ont montré leurs limites en TALN, comme nous l'avons démontré dans l'ensemble des travaux de la partie I. Au contraire, l'approche graphe possède l'avantage d'ignorer le contenu, et donc de s'abstraire des difficultés liées au langage : il semble que dans le cas précis de la détection des abus, les interactions entre utilisateurs sont suffisantes pour modéliser les comportements abusifs.

7.4.3 Complémentarité de la nature des informations

De par la nature différente des caractéristiques extraites de l'approche liée au contenu et de celle modélisant le graphe d'interaction des utilisateurs, nous nous sommes ensuite intéressés à la complémentarité potentielle de ces caractéristiques. Nous avons choisi d'étudier trois approches de fusion des caractéristiques :

— *Fusion précoce* : nous proposons de concaténer les caractéristiques issues du contenu du message (partie 7.2) et celles issues du graphe conversationnel (partie 7.3.2) en un seul vecteur fourni en entrée du classifieur SVM. Le classifieur choisi donc les caractéristiques pertinentes à partir de toutes les informations brutes des approches.

— *Fusion tardive* : les deux approches (texte et graphe) sont, comme précédemment, utilisées indépendamment au sein de deux classifieurs SVM, permettant d'obtenir un résultat de classification pour chaque source de données. Puis ces deux scores sont utilisés en entrée d'un troisième classifieur SVM, devant alors prendre la décision finale (message abusif ou non). L'idée de cette fusion était de voir si l'information importante n'était pas extraite en considérant séparément les sources de données, et donc éviter le bruit potentiel contenu en prenant toutes les données brutes (fusion précoce).

— *Fusion hybride* : l'idée ici était de fournir, en entrée d'un classifieur SVM, les données brutes issues de l'ensemble des caractéristiques (fusion précoce) mais également, au sein de ce vecteur d'entrée, les deux scores des deux classifieurs pour chaque source de données. Notre objectif était de voir si des informations différentes pouvaient être extraites des deux classifieurs, par rapport aux données brutes.

Le tableau 7.2 présente les résultats obtenus pour chaque approche de fusion. Pour faciliter la lecture, nous avons remis les résultats présentés dans le tableau 7.1 pour chaque approche évaluée indépendamment l'une de l'autre.

Méthode	Nb carac.	Précision	Rappel	F-mesure
Approche texte	29	78,59	83,61	81,02
Approche graphe	459	90,21	87,63	88,90
Fusion précoce	488	91,25	89,45	90,34
Fusion tardive	2	**94,10**	**92,43**	**93,26**
Fusion hybride	490	91,96	90,48	91,22

Tableau 7.2 – Rappel, précision et F-mesure (%) pour les différentes fusions des approches s'appuyant sur le contenu textuel (Approche texte) et sur les interactions entre utilisateurs (Approche graphe). Le nombre de caractéristiques (*Nb carac.*) pour chaque méthode est également fourni.

Nous avons tout d'abord pu constater que, peu importe la méthode de fusion proposée, les résultats apparaissent meilleurs que ceux obtenus avec l'approche graphe (et par extension l'approche s'appuyant sur le contenu textuel), que ce soit en termes de précision, rappel ou F-mesure. Cela confirme bien le fait que les informations issues du contenu textuel et celles issues de la structure des conversations sont complémentaires.

Ensuite, en comparant les méthodes de fusion, l'approche *fusion tardive*, utilisant 2 caractéristiques, permet d'atteindre les performances les plus élevées. Cette observation est plutôt surprenante, puisqu'*a priori*, nous imaginions l'approche *fusion précoce* comme la plus performante, le classifieur ayant alors la possibilité de choisir les caractéristiques les plus pertinentes lui-même parmi l'ensemble complet (graphe+texte - 488 caractéristiques). Nous avons émis l'hypothèse que les classifieurs séparés, pour le texte et le graphe, permettent une étape de pré-sélection des caractéristiques et évitent le bruit généré par un trop grand nombre de caractéristiques. Cette hypothèse a semblé se confirmer avec l'approche *fusion hybride*, présentant des performances intermédiaires.

De par cette étude, il apparaissait que toutes les caractéristiques n'étaient donc pas pertinentes pour détecter, et donc caractériser, les messages abusifs. Dans la partie suivante, nous présentons notre étude des caractéristiques importantes au travers d'une partie des expériences que nous avons menées.

7.4.4 Étude des caractéristiques importantes

Nous avons alors proposé une approche permettant d'identifier les caractéristiques les plus importantes dans le processus de classification des messages abusifs. Ce sous-ensemble de caractéristiques est nommé *top-caractéristiques* (TC) dans ce manuscrit. Cette sélection des TC s'appuie sur une méthode itérative implémentée dans l'outil Sklearn, nous fournissant, à chaque itération, un classement individuel des caractéristiques d'entrée, reflétant leur pertinence pour la tâche de classification. Nous supprimons ensuite la fonctionnalité la moins importante et entraînons un nouveau modèle en utilisant toutes les caractéristiques restantes. Nous répétons ce

processus jusqu'à ce que les performances de classification atteignent le seuil minimal ciblé de 97 % du score initial de F-mesure.

Le tableau 7.3 présente les performances de la méthode *fusion précoce* initiale, et ses performances avec simplement les top-caractéristiques (*fusion précoce TC*). Le nombre de caractéristiques et les temps de traitement associés (total et en moyenne par message) sont également fournis. Les performances (3 dernières colonnes) sont fournies ici à titre indicatif, puisque nous savons que les performances de *Fusion précoce TC* seront forcément inférieures à celles de *Fusion précoce* de par le processus itératif de sélection des données.

Méthode	Nb carac.	Total tps traitement	Moy. tps traitement	Précision	Rappel	F-mes.
Fusion préc.	488	8:26:41	7,68s	91,25	89,45	90,34
Fusion préc. TC	**4**	**0:11:29**	**0,17s**	89,09	87,12	88,09

Tableau 7.3 – Nombre de caractéristiques et temps de traitement (total et en moyenne par message) pour la méthode *Fusion précoce* (*Texte* + *Graphe*) et le sous-ensemble des top-caractéristiques (*Fusion précoce TC*). Le temps total de traitement est exprimé en h:m:s.

Il a été intéressant de noter la réduction drastique du nombre de caractéristiques extraites par notre approche de sélection des TC. En effet, avec la fusion précoce, nous avions 488 caractéristiques en entrée. Suite à cette sélection, nous n'en avions plus que 4. Dans notre étude, 3 caractéristiques étaient issues de l'approche graphe et une issue de l'approche texte, ce qui prouve encore la complémentarité possible des sources de données.

Enfin, nous avons observé, forcément, une réduction très importante du temps de traitement nécessaire pour la mise en place de l'approche TC. Outre le côté analytique, la sélection des TC peut donc être intéressante à ce niveau, tout en gardant des performances acceptables.

Des études plus détaillées sur les caractéristiques importantes peuvent se trouver dans [Papegnies et al., 2019] pour l'étude de l'approche par graphes conversationnels, et dans [Cecillon et al., 2019] pour les méthodes de fusion des informations (texte et graphe). Des résultats similaires à la méthode *Fusion précoce* ont été observés.

7.5 Le corpus open-source de conversations WAC

Initialement, nous avons mené nos différents travaux sur le corpus Space Origin (partie 7.4.1). Comme nous avons pu le voir dans les expériences présentées dans ce manuscrit, ce corpus est finalement assez limité en taille et nous nous trouvions dans la situation, au début de la thèse de N. Cécillon, où il était nécessaire de passer à une taille supérieure de données.

L'autre problème auquel nous avons également été confrontés concernait l'évaluation des approches d'autres travaux proposés pour la détection des messages abusifs : bien que le sujet soit relativement bien étudié par la communauté, démontré par les *workshops* spécialisés organisés

sur ce thème (le *Workshop on Abusive Language Online* a par exemple eu droit à 4 éditions - 2017, 2018, 2019 et 2020 - au moment de l'écriture du manuscrit), quasiment aucune de ces approches n'était évaluée sur le même corpus, avec des métriques d'évaluation souvent différentes. Son organisation est d'autant plus difficile que ce sujet est relativement nouveau dans la communauté. Cela rejoint les préoccupations que nous avions évoquées dans les chapitres précédents, à savoir le manque d'un cadre expérimental bien défini.

Nous nous sommes alors intéressés aux corpus disponibles sur lesquels nous pourrions réaliser des expériences à plus grande échelle, et qui, en plus, respectaient notre contrainte que les messages soient issus de conversations pour évaluer notre approche par graphes conversationnels. Bien que peu nombreux, les corpus sur cette tâche (ou une tâche proche) contenant des conversations existaient, mais n'étaient pas publics [Yin et al., 2009], incluant bien entendu le corpus Space Origin [Papegnies et al., 2019]. Récemment, une solution était possible au travers du corpus PreTox [Karan and Šnajder, 2019], un grand corpus de discussions issues des pages de commentaires de Wikipédia, disponible publiquement et annoté sous des catégories proches des messages abusifs. Cependant, l'annotation du corpus en classes d'abus avait été réalisée semi-automatiquement, et cela posait problème car les auteurs rapportaient une précision de seulement 51 %. En fait, le corpus PreTox s'appuie sur le corpus WikiConv [Hua et al., 2018] pour les conversations, mais qui n'avait pas les annotations en classes d'abus (ce qu'a donc rajouté le corpus PreTox). Notons enfin que le corpus WikiConv s'appuyait sur les données collectées dans [Wulczyn et al., 2017], ou les auteurs avaient annoté manuellement en classes d'abus les conversations dans Wikipedia, mais ne fournissaient que les messages annotés seuls (*i.e.* aucune information n'était fournie quant à la conversation dans laquelle ces messages se trouvaient).

Au final, nous avions accès à un ensemble très grand de messages issus de Wikipédia en anglais et annotés manuellement, mais sans leurs conversations associées, et de l'autre les conversations associées mais sans leur annotation manuelle. Nous avons donc proposé, dans [Cecillon et al., 2020b], de reconstruire un nouveau corpus de messages à partir de ces deux corpus existants, structuré en conversations complètes et avec des annotations manuelles de haute qualité. Nous avons alors obtenu le corpus *Wikipedia Abusive Conversations* (WAC), contenant au final environ 193 000 conversations et 383 000 messages annotés comme étant abusifs ou non. Pour encourager le développement de nouvelles méthodes sur cette tâche de détection de contenus abusifs, nous avons publié et mis librement à disposition ce corpus, ainsi que le code source développé pour sa reconstruction [6]. De même, sur la page où nous avons fourni le corpus, nous avons proposé une plateforme d'évaluation reprenant toutes les approches que nous avons développées pendant nos travaux.

6. https://github.com/CompNet/WikiSynch

7.6 Conclusion

Il s'agissait, dans ce chapitre, de présenter nos travaux sur la problématique de la détection des messages abusifs dans des conversations sur Internet, et plus particulièrement que l'on retrouve dans le cadre des réseaux sociaux numériques (RSN). Ces travaux ont tout d'abord été initiés avec E. Papegniès et V. Labatut en prenant le problème de façon assez classique, avec une classification automatique prenant en entrée des caractéristiques éprouvées en TAL. Nous avions néanmoins perçu, très tôt, les limites des approches TAL pour ce problème particulier : messages courts, langage atypique, thèmes de discussions vastes, possibilité des utilisateurs de masquer les abus... Nous avons alors rapidement proposé de nous intéresser, non pas au contenu direct échangé, mais aux interactions entre les utilisateurs. Cela a conduit à proposer une modélisation par graphes de ces interactions, que nous avons nommée *graphe conversationnel*, qui prend ses origines dans le domaine des réseaux complexes. Finalement, les expériences menées ont montré de meilleurs résultats au moyen de l'approche graphe, et des mesures topologiques associées, en comparaison de l'approche TAL initiale : ignorer totalement le contenu textuel semble donc une approche alternative intéressante, puisqu'il apparait plus difficile de masquer des comportements que le contenu textuel.

Forts de ces conclusions, nous avons poursuivi le travail en proposant de combiner les approches graphes et textes, dont nous pensions la nature des données complémentaire, car différente. Nous avons pu montrer que les résultats de détection des messages abusifs pouvaient être améliorés au moyen de ces sources de données, tout en proposant une étude qualitative permettant de définir les caractéristiques jouant un rôle important dans le processus de classification.

Enfin, nous avons proposé un nouveau très grand corpus de données conversationnelles annotées en classes d'abus issues de Wikipedia, en nous appuyant sur des corpus existants mais dont les faiblesses ne permettaient pas d'y appliquer les approches que nous proposions. Nous avons diffusé ce corpus librement à la communauté scientifique, ainsi que tous les outils associés, pour la tâche de détection d'abus.

Au final, nous avons, avec succès, combiné deux thématiques de recherche, à savoir le TAL et les réseaux complexes. Avant de proposer les perspectives de ce travail dans la partie V, nous terminons la présentation des travaux interdisciplinaires sur la problématique du doublage vocal dans le chapitre suivant.

Chapitre 8

Doublage vocal et recommandation de voix

Sommaire

8.1 Introduction

Depuis de très nombreuses années, la voix fait partie des objets d'étude qui intéressent de manière assez large le milieu scientifique, que ce soit par exemple en phonétique, pour comprendre la production des sons, ou encore en médecine, avec l'étude de ce phénomène physiologique. Dans le cadre du traitement automatique de la parole (TAP), il s'agit d'étudier la voix en vue de son traitement automatique pour différentes applications (reconnaissance automatique de la parole, reconnaissance et vérification du locuteur, synthèse de la parole...). Outre l'informatique, et tout ce qui concerne le traitement du signal, le TAP regroupe d'autres domaines de recherche tels que la phonétique et la phonologie.

Historiquement, le TAP s'est très tôt organisé autour de tâches clairement identifiées dans lesquelles des cadres expérimentaux ont permis des progrès scientifiques ces dernières décennies. En prenant comme exemple la problématique de la reconnaissance et vérification du locuteur, les campagnes d'évaluation NIST ont permis, depuis le milieu des années 90 [1], de donner une réelle visibilité et une importance internationale à ce domaine de recherche, permettant un grand nombre de financements, et ainsi des avancées certaines. Ce domaine de recherche possède l'avantage d'avoir une tâche clairement identifiée qui est finalement assez simple à évaluer : soit le locuteur est bien reconnu par le système de reconnaissance automatique, soit il s'est trompé, et il convient de l'améliorer. Les recherches dans ce domaine se sont donc moins concentrées sur la définition de la tâche que sur les performances intrinsèques des systèmes et la prise en compte des difficultés liées à l'acoustique, comme la compensation du bruit lié à l'environnement, aux locuteurs... Comme dans de nombreux domaines, les approches par apprentissage profond ont récemment permis de grandes avancées [Snyder et al., 2018].

Ces dernières années ont vu émerger en TAP de nouvelles problématiques de recherche sur des cadres expérimentaux beaucoup plus difficiles à définir, en particulier tout ce qui concerne la para-linguistique. Nous pouvons par exemple citer le domaine de la reconnaissance automatique des émotions dans la voix, qui est finalement une problématique récente en comparaison des nombreux travaux de recherche effectués en sciences humaines avec de nombreux modèles proposés, comme par exemple le modèle catégoriel de Hevner pour la musique [Hevner, 1936], ou encore le modèle Valence-Activation [Russell, 1980]. Ce sont les différentes campagnes d'évaluation et *workshops* spécialisés qui ont permis l'émergence de travaux plus nombreux en vue de leur traitement automatique, comme par exemple dans le cadre de la conférence Interspeech, avec ses *paralinguistic challenges* depuis 2009 [Schuller et al., 2009, 2013, 2020]. Outre le fait d'avoir des cadres expérimentaux sur lesquels s'appuyer, ces problématiques rencontrent des difficultés, et souvent des oppositions d'autres chercheurs en TAP, liées à la subjectivité de ces tâches, en particulier sur la réception et la perception de la voix.

Le projet ANR The Voice vient clairement se positionner sur ces problématiques nouvelles, cherchant à extraire automatiquement de la voix des informations liées à la réception et la perception de celle-ci par des humains. Porté par Nicolas Obin, MCF à l'IRCAM, et en collaboration avec Jean-François Bonastre au LIA, un des objectifs du projet, sur lequel j'ai travaillé, vise à modéliser la *palette vocale* d'une personne (ici, un acteur), *i.e.* être capable de rendre compte des capacités vocales d'un acteur à doubler certains rôles et/ou personnages (voir partie 9.3.2). Le doublage vocal consiste à remplacer une voix originale par la voix d'un nouvel acteur, généralement lorsque l'on doit changer la langue originale, pour diffuser l'oeuvre dans un autre pays, ou, parfois, pour des raisons artistiques (par exemple, doublage de la voix d'Arnold

1. `https://www.nist.gov/itl/iad/mig/speaker-recognition`

Schwarzenegger dans le film *Hercules in New York*[2]). Nous dépassons ici le cadre d'une simple reconnaissance acoustique, comme nous pourrions le retrouver en reconnaissance et vérification du locuteur, l'objectif étant d'extraire des caractéristiques liées aux choix subjectifs effectués par un opérateur humain ainsi qu'à la réception de la voix par le public.

Les travaux présentés dans cette dernière partie sont donc liés au projet The Voice, mais surtout à la thèse d'Adrien Gresse, que j'ai eu l'occasion de co-encadrer, et qui a débuté quelques années avant le début du projet. Plus récemment, Mathias Quillot, que je co-encadre également, continue en thèse sur cette problématique. Dans la partie 8.2, nous présentons le cadre expérimental défini et mis en place pendant la thèse d'A. Gresse pour le doublage et la recommandation de voix, qui était alors inexistant sur cette problématique quasi-inexplorée. Nous voyons ensuite, dans la partie 8.3, la proposition d'un espace de représentation de voix jouées, ici de *personnages*, comme un premier pas vers la caractérisation de cette palette vocale [Gresse et al., 2019, 2020b]. Enfin, dans la partie 8.4, nous détaillons une approche originale que nous avons proposée pour la comparaison et la similarité de voix dans le contexte du doublage vocal [Gresse et al., 2017]. Ce chapitre est l'occasion de rendre compte des difficultés des travaux interdisciplinaires qui y ont été menés, où un cadre expérimental complet a dû être mis en place (de la définition de la tâche à son évaluation), et qui pose encore aujourd'hui de multiples questions scientifiques.

8.2 Définition d'un cadre expérimental

8.2.1 Contexte

Dans l'introduction de ce chapitre, nous avons commencé à définir le doublage vocal et à expliquer le processus permettant de *trouver* la voix de doublage permettant de jouer *vocalement* un personnage cible. De par l'internationalisation des oeuvres multimédias (par exemple les films, séries, jeux vidéos...), et la volonté des producteurs de toucher un public le plus large possible, il s'agit souvent de trouver la voix dans une langue ciblée permettant de remplacer au mieux la voix de la version originale (VO), généralement dans une autre langue. Alors qu'*a priori* nous pourrions nous dire qu'il s'agit de *copier* la VO, il apparaît que ce processus de sélection de voix est bien plus complexe. Il est plutôt question d'une adaptation à une langue, une culture, un pays, des attentes d'un public... qu'à une simple ressemblance vocale. Dans l'industrie audiovisuelle, le doublage vocal d'un média est d'une importance capitale pour l'appropriation de l'oeuvre par son public, lui permettant de s'y immerger. Par exemple, dans le cadre du développement du jeu vidéo *Red Dead Redemption 2*, le producteur Dan Houser a expliqué que[3] :

2. https://fr.wikipedia.org/wiki/Hercule_à_New_York#Autour_du_film
3. https://www.vulture.com/2018/10/the-making-of-rockstar-games-red-dead-redemption-2.html

> We don't bring in name actors anymore because of their egos and, most important of all, because we believe we get a better sense of immersion using talented actors whose voices you don't recognize. *Dan Houser*

Ce processus de sélection des voix, hautement important et stratégique, et appelé communément *casting vocal*, est assuré par des experts humains que l'on retrouve très souvent sous la dénomination de *directeur artistique*. Du fait que ce processus soit actuellement complètement réalisé humainement, plusieurs limites apparaissent :

— Il peut exister une très forte subjectivité dans le choix de l'opérateur humain, la sélection de voix ne se faisant pas sur des critères purement acoustiques (*i.e.* une simple correspondance acoustique entre les voix) mais plus sur des critères liés à des facteurs culturels (par exemple, certains stéréotypes culturels liés à des personnages précis), de la propre expérience passée du directeur artistique, de ses anciennes collaborations avec des acteurs vocaux...

— L'impossibilité de pouvoir auditionner tous les acteurs vocaux, tant les bases de données de voix sont grandes, alors même que les directeurs artistiques sont intéressés par la découverte de nouvelles voix (cf. citation précédente de Dan Houser).

— Des contraintes extérieures, telles que le budget du film (impossibilité de payer un acteur vocal *star*), la disponibilité de l'acteur...

Les directeurs artistiques ont donc un réel intérêt à utiliser des outils automatiques pour les aider dans ce travail de casting vocal, en particulier pour tout ce qui concerne la recommandation automatique de voix, mais également pour comprendre les choix effectués par les systèmes automatiques ou les experts humains eux-mêmes (explication des critères/caractéristiques pour la proposition de voix).

Ce problème de recommandation automatique de voix dans le contexte du doublage vocal a finalement été assez peu étudié [Obin and Roebel, 2016]. Il apparaît surtout que les corpus de données disponibles pour travailler sur cette problématique sont finalement inexistants. La thèse d'A. Gresse, et dans laquelle le travail présenté ici a vu ses origines, a notamment été débutée avant le projet ANR The Voice, qui doit permettre d'avoir de plus grandes ressources pour travailler sur le doublage vocal et la recommandation de voix.

Une des parties de son travail de thèse a alors été consacrée à définir un cadre expérimental complet pour le doublage vocal, allant de sa définition, en particulier sur le contrôle des biais potentiels par rapport à notre compréhension de la tâche et pouvant fausser l'interprétation des résultats, jusqu'à son évaluation. Il s'agit ici d'appairer automatiquement un segment de voix issu de la VO, avec un segment de voix de la version cible. Ce protocole a ensuite été mis en place sur des données vocales issues du jeu *Mass Effect 3*, la VO étant alors l'anglais, et la version cible, la version française (VF).

8.2.2 Importance et gestion des biais

Il s'agissait de lister, pour les contrôler, certains biais liés à l'utilisation des données pour l'apprentissage, mais surtout pour l'évaluation de modèles dans le contexte du doublage vocal. En effet, notre objectif est de mettre en lumière le fait que les approches que nous proposons permettent de modéliser la voix d'un acteur pour un personnage : l'appariement par un système automatique de deux voix ne doit pas être dû à des éléments extérieurs autres que le jeu de l'acteur. Pour ce faire, nous avons listé les biais suivants, en apportant des éléments pour les contrôler [Gresse, 2020] :

— *Déséquilibre du corpus.* Comme dans tout corpus, des données très déséquilibrées ont une influence sur les modèles construits et leur évaluation (ici, certains personnages sont très présents, d'autres beaucoup moins). Afin de ne pas influencer l'appariement de voix, nous avons veillé à avoir des données équilibrées entre les différents personnages.
— *Genre des personnages.* Sachant que nos travaux s'appuient, au départ, sur des représentations largement utilisées dans le cadre de la reconnaissance du locuteur, nous savons qu'il est simple de différencier des personnages masculins de personnages féminins. Nous avons donc veillé à toujours avoir des paires de segments de même genre.
— *Contenu linguistique.* Bien qu'ils soient dans deux langues différentes, deux segments de voix (ici, VO-VF) véhiculent le même message. Nous pensons donc que les systèmes pourraient apprendre une *concordance linguistique* entre ces segments. Pour éviter ce biais, nous avons proposé de mélanger aléatoirement les segments entre les personnages.
— *Segments de durée courte.* Afin de pouvoir représenter correctement un personnage, et donc sa voix sur un extrait vocal, nous avons émis l'hypothèse qu'une durée trop courte pour un segment entraîne une représentation de mauvaise qualité. Ces segments sont alors retirés de l'étude (ici, segments de durée inférieure à 1 seconde).
— *Acteur associé à plusieurs personnages.* Il s'agit ici de s'assurer qu'aucun acteur utilisé pour l'apprentissage des modèles ne soit également présent dans le corpus de test. Si tel était le cas, il serait possible d'imaginer que le système reconnaît le locuteur plutôt que le personnage. Une façon de s'en assurer est de considérer qu'un acteur vocal joue un seul personnage dans le corpus. Cela permet aussi de vérifier que notre système permet de généraliser à des personnages nouveaux (*i.e.* non présents lors de l'apprentissage).

Nous sommes conscients que d'autres biais peuvent exister, mais nous avons essayé de lister ceux dont nous avions, *a priori*, l'intuition qu'ils pouvaient influer sur les modèles construits et leur évaluation, en particulier de par l'expérience dans le domaine de la reconnaissance automatique du locuteur. Bien entendu, ce travail est toujours en cours et ouvert.

8.2.3 Classification binaire

Afin d'évaluer le travail sur le doublage vocal, nous avons proposé, comme précisé précédemment, une classification binaire devant décider si deux segments de voix (donc ici, appartenant à deux acteurs vocaux dans deux langues différentes) appartiennent, ou non, au même personnage. Les paires appartenant au même personnage sont appelées *cible* dans notre étude, celles avec deux personnages différents *non cible*. Toujours dans l'optique d'éviter un déséquilibre dans les données, nous avons choisi de conserver le même nombre pour les paires de segment *cible* et les paires *non cible*. La représentation générale du système est schématisée dans la figure 8.1.

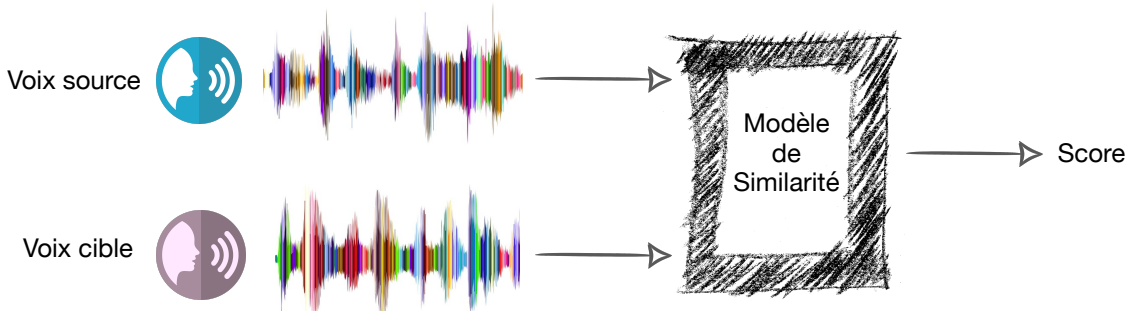

FIGURE 8.1 – Représentation générale du système de similarité de voix.

L'évaluation est, quant à elle, réalisée au moyen de la métrique classique du taux de réussite (*accuracy*) du système.

8.2.4 Corpus Mass Effect 3

Un corpus a été construit durant la thèse d'A. Gresse. Il est issu du jeu vidéo *Mass Effect 3*, dont les interactions vocales ont été extraites. Les dialogues du jeu sont originellement en anglais, mais ont été traduits dans plusieurs langues. Dans le cadre de ce travail, l'anglais a été utilisé comme langue source (VO), et le français comme langue cible (VF). En appliquant les différentes contraintes liées à la gestion des biais (voir partie 8.2.2), 16 personnages sont finalement utilisés, et le nombre de segments par personnage et par langue a été fixé à 90. L'ensemble des fichiers audio représente au final un total de 7,5 heures dans chaque langue (anglais et français), et les segments audio durent en moyenne 3,5 secondes.

De par la faible taille du corpus collecté, le choix a été fait de réaliser une validation croisée pour évaluer nos systèmes. Tour à tour, chaque ensemble de données est ainsi évalué, les autres ensembles de données étant alors utilisés pour l'entraînement des modèles. La figure 8.2 présente le découpage des données en 4 sous-ensembles identiques.

Au niveau du découpage en genre, nous avons ainsi conservé 5 personnages féminins, et donc 11 personnages masculins. Bien que la catégorisation proposée par A. Gresse pendant sa thèse

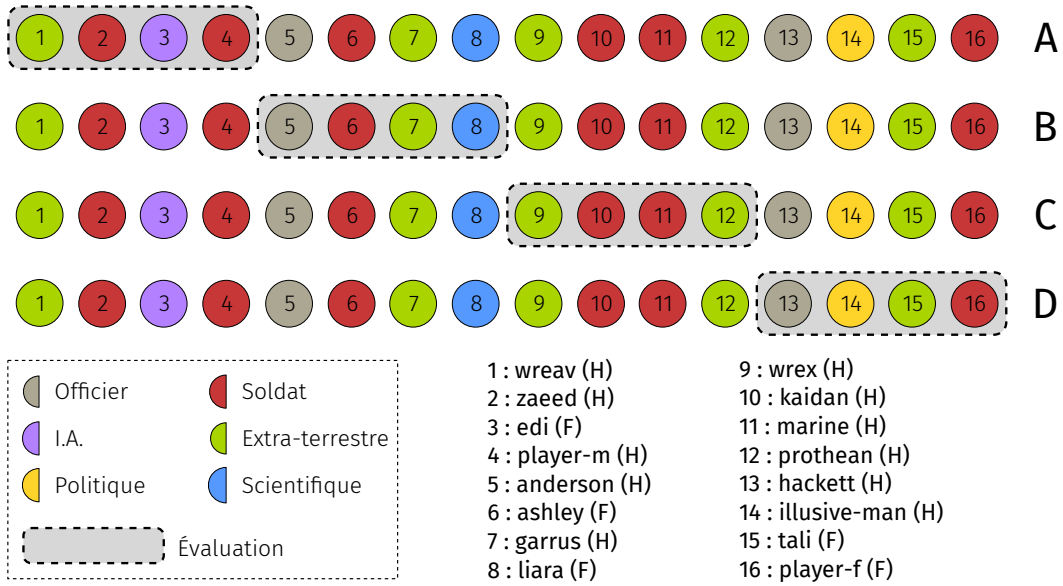

FIGURE 8.2 – Proposition de découpage en 4 ensembles de voix (A, B, C et D) des 16 personnages du jeu Mass Effect 3 (thèse d'Adrien Gresse [Gresse, 2020]). L'information sur la catégorie de chaque personnage (Officier, Soldat...) est purement subjective ici (non fournie dans les métadonnées du jeu).

(et non par les auteurs du jeu eux-mêmes) reste forcément subjective, nous avons essayé de conserver une certaine mixité dans ces catégories pour chaque ensemble de données.

8.3 Représentation de la voix jouée

8.3.1 Paramétrisation acoustique

Une des problématiques développées pour la recommandation de voix dans le cadre du doublage vocal d'oeuvres multimédias concerne la manière de représenter la voix du doubleur. La représentation de la voix d'une personne (*locuteur*), et plus généralement du signal audio, n'est pas un problème récent. En effet, très tôt, les chercheurs dans le domaine de la voix, et en particulier ceux travaillant en traitement du signal, ont relevé qu'il était très difficile de travailler directement sur le signal en vue de son traitement automatique de par sa très grande variabilité (canal de transmission, qualité des enregistrements, bruits environnants, contenu linguistique véhiculé...).

Dans le contexte du TAP, la première étape consiste généralement à paramétrer le signal, en le transformant en un ensemble de vecteurs de paramètres pertinents ayant chacun une dimension fixée au départ. Il est nécessaire de découper le signal audio par trame, en prenant une

taille fixe (souvent définie aux alentours de 25 ms), afin de rendre le signal quasi-stationnaire. Ce découpage est réalisé toutes les 10 ms pour garder un recouvrement entre les différentes trames. Un vecteur de paramètres est ensuite extrait pour chaque trame. Les MFCC (*Mel-Frequency Cepstral Coefficients*) font partie, par exemple, des coefficients cepstraux les plus couramment utilisés [Dave, 2013]. Cette extraction permet alors d'obtenir la séquence d'observations acoustiques X, où $X = x_1x_2...x_n$, *i.e.* un vecteur de paramètres associé à une trame. Ces vecteurs de paramètres sont alors utilisés en entrée de différentes tâches de TAP (vérification du locuteur, RAP...).

À partir de ces paramètres acoustiques, il est possible de construire des modèles permettant de représenter, au mieux, des voix. Les travaux les plus avancés sur ce sujet se trouvent actuellement dans le domaine de la reconnaissance du locuteur. Comme nous l'avons vu, les variabilités contenues dans le signal de parole rendent la modélisation des locuteurs difficile : il s'agit souvent de compenser ces variabilités au maximum pour avoir une représentation robuste du locuteur.

8.3.2 Représentations classiques du locuteur

Représentation i-vecteur

Comme nous l'avons décrit succinctement dans la partie 2.5.2, l'approche i-vecteur [Dehak et al., 2010] est une manière de réduire la grande dimension des données fournies en entrée en un vecteur de caractéristiques de petite taille, dont l'objectif est de retenir le maximum d'informations pertinentes liées au locuteur.

Dans le cadre de notre étude, le fait que l'approche i-vecteur ait longtemps été à l'état de l'art dans le cadre de la vérification et reconnaissance automatique du locuteur prouve son efficacité pour représenter des locuteurs et leurs segments de parole associés. Cela permet de travailler avec une représentation robuste en entrée, mais également de nous affranchir du problème des durées variables des segments de parole, puisque la représentation i-vecteur permet de représenter des séquences de durées variables par un vecteur de taille fixe. Bien entendu, le fait que l'approche i-vecteur soit conçue pour représenter un locuteur peut poser question, sachant que notre objectif n'est pas ici de représenter l'identité du locuteur, mais plutôt le personnage que celui-ci joue. Ce questionnement est à l'origine de l'approche p-vecteur que nous présentons dans la partie 8.3.3.

Représentation x-vecteur

Récemment, de nouvelles approches par apprentissage profond ont permis de surpasser les performances obtenues par la représentation i-vecteur. Dans [Snyder et al., 2018], les auteurs proposent d'apprendre des caractéristiques *haut-niveau* du locuteur au moyen de réseaux de neurones profonds (DNN) grâce à une tâche d'identification du locuteur, *i.e.* au moyen d'une tâche de classification automatique cherchant à classifier des segments de parole parmi n locuteurs.

Dans ce contexte, les différentes couches du DNN sont entraînées pour extraire des informations pertinentes permettant de discriminer les différents locuteurs. L'idée principale est d'extraire un vecteur de caractéristiques du segment de parole donné en entrée en prenant une des couches cachées du DNN (dans [Snyder et al., 2018], la couche cachée 6). Ce vecteur de caractéristiques est alors appelé x-vecteur, ou plongement de locuteur (*speaker embedding*).

8.3.3 Représentation p-vecteur pour le personnage

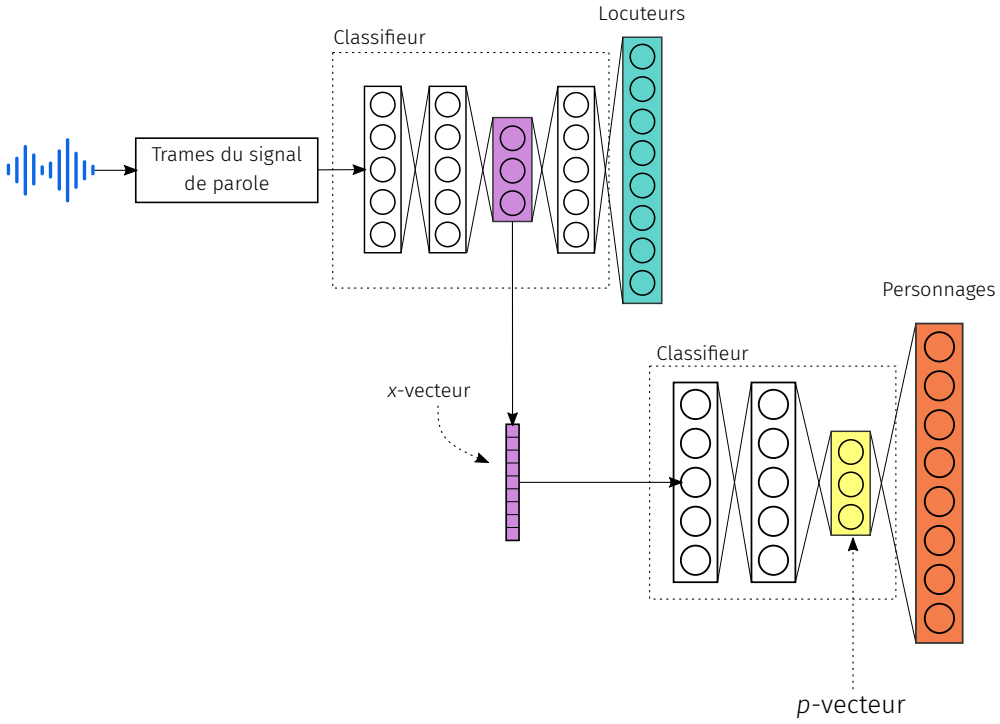

FIGURE 8.3 – Schéma pour l'apprentissage de p-vecteurs [Gresse, 2020].

Nous partons de la problématique que les caractéristiques permettant de décrire la voix jouée, ici dans le contexte d'un *personnage* ou *rôle* identifié, ne sont pas connues : aucune étude, en particulier en sciences humaines, n'existe pour décrire précisément ce qui caractérise une voix pour un personnage identifié. Nous avons alors proposé et étudié une nouvelle représentation des voix, ici pour la représentation des personnages, que nous avons nommée p-vecteur [Gresse et al., 2020b]. L'approche p-vecteur s'appuie sur l'idée développée dans les travaux sur les x-vecteurs en proposant d'utiliser des réseaux de neurones pour apprendre des plongements de personnages. Une des contraintes que nous avions était que peu de données d'apprentissage étaient disponibles, alors que ce type d'approche nécessite de gros volumes de données. Nous avons donc proposé

d'apprendre les p-vecteurs en deux étapes : 1) une première étape consiste à apprendre des représentations x-vecteurs entraînées sur un grand nombre de locuteurs extérieurs à notre tâche ; 2) la seconde étape consiste à projeter les segments de voix dans l'espace des x-vecteurs, pour les utiliser ensuite en entrée d'un nouveau réseau de neurones entraîné à reconnaître, cette fois-ci, des personnes. La dernière couche cachée de ce dernier réseau correspond au p-vecteur. La figure 8.3 illustre le processus d'apprentissage des p-vecteurs.

8.4 Comparaison et similarité de voix

8.4.1 Notion de similarité

Dans le domaine de la reconnaissance et de la vérification du locuteur, la similarité recherchée est ici purement acoustique, la tâche elle-même ayant évacué toute subjectivité, en particulier dans l'annotation (et donc l'évaluation) des données : soit les extraits vocaux appartiennent à la même personne, soit à des personnes différentes, seules variant les conditions acoustiques. Cependant, de manière plus générale, la façon dont les humains perçoivent des voix comme *similaires* reste une question complètement ouverte. Dans nos travaux sur la recommandation de voix dans le contexte du doublage vocal, la similarité entre une voix originale et la voix de doublage choisie dépasse le simple cadre de similarité acoustique (voir partie 8.2.1). La thèse d'A. Gresse s'est intéressée à la notion de similarité entre deux voix, dépassant le cadre du TAP pour s'aventurer sur des notions définies en sciences humaines.

Dans les années 1980, des travaux ont notamment débuté dans le domaine de la phonétique. Nous pouvons, par exemple, citer les travaux de [Laver, 1980] qui a proposé d'étudier certaines caractéristiques de la voix, au travers de la description d'une terminologie, pour comprendre leur influence dans ce qu'il nomme la *qualité vocale*. L'évaluation du niveau de similarité des voix a également été explorée dans de nombreux travaux tels que [Phil, 1999, McDougall, 2013]. Ils montrent l'existence de corrélations entre des caractéristiques acoustiques particulières et la façon dont nous comprenons que deux voix sont perçues comme similaires. Cependant, il n'existe pas de méthodes bien établies pour estimer automatiquement cette similitude.

Comme nous l'avons énoncé, les travaux de [Obin and Roebel, 2016] sont un premier pas sur la problématique de la similarité de voix dans le cadre du doublage vocal. Nos travaux préliminaires sur la similarité de voix s'intéressent à l'application d'une approche classique en reconnaissance du locuteur, associant une représentation i-vecteur avec une Analyse Discriminante Linéaire Probabiliste (PLDA) [Gresse et al., 2017]. Cette première étude a servi à montrer la faisabilité d'un système automatique d'appariement de voix originales et de voix de doublage dans deux langues différentes. Cela nous a notamment permis de mettre en place le cadre expérimental décrit dans la partie 8.2. L'approche PLDA se focalisant plutôt sur l'identité du locuteur, nous avons ensuite proposé une autre approche s'appuyant sur les réseaux de neurones

siamois adaptés à la similarité de voix [Gresse et al., 2019] dont nous présentons quelques détails dans la partie 8.4.2, ainsi que les expériences menées en conjonction avec la représentation p-vecteur [Gresse et al., 2020b] (partie 8.4.3).

8.4.2 Réseaux de neurones siamois

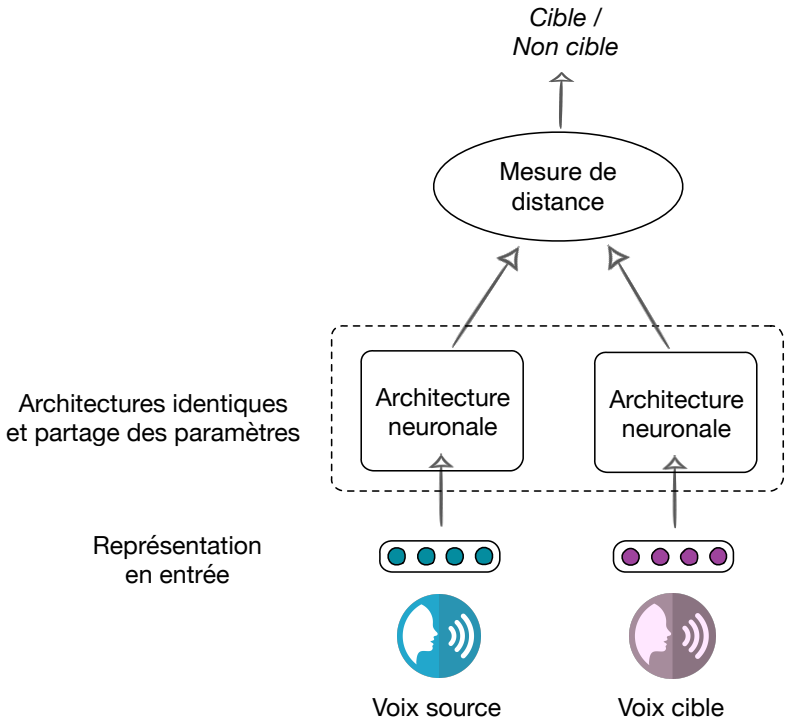

FIGURE 8.4 – Schéma du réseau siamois utilisé pour déterminer si deux voix appartiennent (cible) ou n'appartiennent pas (non cible) au même personnage.

Les réseaux de neurones siamois [Bromley et al., 1994] peuvent être vus comme une architecture capable d'apprendre une mesure de similarité au moyen de deux entrées indépendantes, chacune de ces entrées partageant cependant une relation abstraite de similarité. Ces deux entrées sont projetées dans deux réseaux de neurones identiques, partageant les mêmes paramètres, et se rejoignant au moyen d'une fonction de pénalité. Cette fonction s'appuie sur une métrique calculée à partir des dernières couches de chacun des réseaux. Une description précise de l'architecture et de son application dans le cadre de ce travail est disponible dans [Gresse, 2020]. En particulier, dans les expériences présentées ensuite, le réseau de neurones siamois s'appuie sur deux perceptrons multicouches avec deux couches cachées de 256 neurones et fonction d'activation tangente hyperbolique.

Dans le cadre de ce travail, la figure 8.4 présente un schéma du réseau de neurones siamois

utilisé. L'objectif est ici de partir d'une représentation de chaque voix (source, pour la voix originale ; cible, pour la voix du doubleur) en entrée de chaque réseau de neurones. Comme expliqué, une distance est ensuite utilisée pour déterminer si les deux voix en entrée correspondent au même personnage (*cible*) ou non (*non cible*). L'apport des réseaux de neurones siamois a été démontré dans [Gresse, 2020], en comparaison avec l'utilisation d'un réseau de neurones classique avec pour entrées les deux représentations directes. Dans les expériences présentées dans la partie suivante, il s'agit de présenter l'apport à la tâche de similarité de voix au moyen de l'architecture par réseaux de neurones siamois sur la représentation p-vecteur, conçue pour représenter des voix jouées, par rapport à des représentations classiques en reconnaissance du locuteur.

8.4.3 Expériences

Protocole expérimental

Nous décrivons une partie des expériences que nous avons menées sur le doublage vocal. Nous comparons ici l'utilisation des représentations p-vecteurs, avec les représentations classiques en locuteurs (i-vecteur et x-vecteur). L'évaluation présentée se focalise sur la tâche d'appariement des voix de doublage au moyen de l'architecture par réseaux de neurones siamois. Pour rappel, il s'agit de retrouver les paires de segments appartenant au même personnage (*cible*), et donc ceux appartenant à des personnages différents (*non cible*), en donnant en entrée un segment en anglais (VO) et un segment en français (VF), comme décrit dans la partie 8.2.4.

Comparaison des représentations

Les résultats obtenus sur les données de test (taux de réussite) avec les 3 représentations de voix considérées (i-vecteur, x-vecteur et p-vecteur) sur la tâche d'appariement des voix de doublage sont présentés dans le tableau 8.1.

Les résultats obtenus dans cette expérience sont relativement contrastés. Même si une amélioration semble visible avec les p-vecteurs, en comparaison des représentations *locuteur* classiques, la seule performance sur la tâche d'appariement de voix n'apparaît pas suffisante pour démontrer clairement que cette représentation intègre une information supplémentaire au locuteur et permet de mieux appairer des voix de doublage VO-VF. Suspectant une limite au niveau de nos données, de nouvelles approches, en particulier sur l'apport de nouvelles données par distillation des connaissances, ont montré des améliorations quant aux résultats [Gresse et al., 2020b].

Représentation	Sous-ensemble	Taux de réussite
i-vecteur	A	0,60
	B	0,52
	C	0,54
	D	0,49
	moyenne	0,54
x-vecteur	A	0,60
	B	0,54
	C	0,52
	D	0,49
	moyenne	0,54
p-vecteur	A	0,58
	B	0,54
	C	**0,57**
	D	**0,54**
	moyenne	**0,55**

Tableau 8.1 – Comparaison des performances (taux de réussite) obtenues sur la tâche d'appariement de voix par la représentation p-vecteur, orientée personnage, et les deux représentations classiques du locuteur (i-vecteur et x-vecteur).

8.5 Conclusion

Ce chapitre conclut les travaux que nous avons menés autour du thème de l'interdisciplinarité et du traitement du langage. J'ai introduit une partie des travaux entrepris dans le contexte du doublage vocal et de la recommandation de voix. Ces travaux prennent leur origine dans ceux que j'ai pu mener dans le cadre de la campagne d'évaluation MediaEval 2013 sur la tâche *MusiClef tracks* et publiés dans la conférence ISMIR [Morchid et al., 2014g], ainsi que le stage de Master d'A. Gresse (voir partie 10.2.3). Une grande majorité des travaux exposés ici sont également issus de sa thèse. Cette problématique de recherche étant nouvelle, et quasiment inexplorée, une partie des travaux s'est concentrée à proposer un cadre expérimental permettant d'évaluer les approches proposées pour la similarité de voix de doublage, qui était alors inexistant. De ce cadre, nous avons pu proposer différentes approches originales, avec tout d'abord une représentation des voix de doublage multilingues, à savoir ici la représentation p-vecteur, permettant de dépasser les représentations actuelles orientées *locuteur* (i-vecteur et x-vecteur) pour représenter le niveau personnage/rôle de la voix jouée. Nous avons également mis en avant l'architecture des réseaux de neurones siamois, mieux adaptée ici pour la tâche d'appariement des voix de doublage (VO-VF) que des architectures neuronales classiques. Ces travaux préliminaires ont permis de mettre en avant le fait que des caractéristiques liées aux personnages/rôles sont présentes dans les voix de doublage, alors même que le processus de sélection de ces voix apparaît fortement subjectif.

Bien entendu, ce travail est une amorce vers une meilleure compréhension de la voix jouée, et en particulier sur ce problème de définition de la *palette vocale*. Mathias Quillot poursuit actuellement ce travail, en se focalisant sur l'extraction de caractéristiques dans la voix et la mise en avant d'informations autres que liées au locuteur. Les perspectives sont très nombreuses et l'interdisciplinarité indispensable pour permettre des avancées dans cette problématique. Il s'agira notamment de travailler sur la réception et la perception humaine, en menant, par exemple, des tests perceptifs, que ce soit sur les choix initiaux manuels faits par les directeurs artistiques ou par les systèmes automatiques proposés, ce que s'emploient à faire les chercheurs en sciences humaines impliqués dans le projet ANR The Voice.

Ce chapitre conclut mon investissement ces dernières années dans des travaux consacrés au traitement du langage, à la frontière d'autres domaines de recherche. Cela nécessite un investissement différent, et il apparaît clairement, en conclusion de cette partie, que des avancées significatives ne pourront se faire sans l'apport d'autres domaines de recherche : la conception d'un système automatique ne peut s'appuyer uniquement sur des approches statistiques. Le bilan global des différents travaux menés, mais également les perspectives de recherche envisagées, sont exposés dans le chapitre V. Juste avant, je présente, dans la partie IV, un résumé de mes activités d'encadrement et d'administration de la recherche.

Quatrième partie

Administration de la recherche et encadrement

Chapitre 9

Thématiques développées et Projets de recherche

Sommaire

Ce chapitre est dédié à mon implication dans différents projets de recherche. Les premières parties seront consacrées à la synthèse de ces projets de recherche et leur intégration dans mes problématiques et thématiques de recherche. La dernière partie s'intéressera à un résumé de mon implication dans des collaborations industrielles.

9.1 Reconnaissance automatique de la parole et extraction d'information

Mes travaux de recherche ont débuté en reconnaissance automatique de la parole (RAP), en particulier sur le traitement de la parole spontanée. Cette thématique se retrouve dans le projet ANR EPAC (sous-partie 9.1.1), où mes travaux m'ont amené à m'intéresser à l'adaptation des modèles de RAP. J'ai également pu travailler sur l'extraction d'information à travers les systèmes de RAP, que l'on retrouve à la fois dans le projet ANR EPAC, pour la détection de la parole

spontanée, mais également le projet ANR PERCOL (sous-partie 9.1.2), pour la reconnaissance et correction de noms propres.

9.1.1 Participation au projet ANR EPAC (2007-2010)

Le projet ANR EPAC est le premier projet financé auquel j'ai eu l'occasion de participer. Bien que je n'étais pas financé par ce projet, il intervient dans la cadre de ma thèse sur le traitement de la parole spontanée. Le projet visait l'exploration de masses de documents audio pour l'extraction et le traitement de la parole conversationnelle, et impliquait 4 partenaires académiques : Le LIUM (Le Mans Université - coordinateur du projet), le LI (Université de Tours), le LIA (Avignon Université) et l'IRIT (Université de Toulouse).

L'objectif principal de mon travail de thèse, prenant part dans le projet, a été de proposer des solutions pour améliorer les performances des systèmes de RAP sur la parole spontanée, caractérisée par de nombreuses spécificités (disfluences, agrammaticalité, baisse de la fluidité de la parole...). Ainsi, dans un premier temps, un outil de détection automatique de la parole spontanée a été proposé [Dufour et al., 2009]. Cet outil est très important puisqu'il a permis, dans un deuxième temps, de définir une approche non-supervisée d'adaptation des modèles acoustiques et des modèles de langage du système de RAP à la parole spontanée, en sélectionnant automatiquement les segments contenant ce type de parole [Dufour et al., 2010].

9.1.2 Participation au projet ANR PERCOL (2012-2014)

Dès mon arrivée au LIA en tant que maître de conférences en septembre 2012, j'ai pu m'intégrer dans le projet ANR PERCOL [Favre et al., 2013], supervisé au LIA par Corinne Fredouille, regroupant alors 2 autres laboratoires de recherche académique (LIF - Aix Marseille Université - coordinateur scientifique, et LIFL - Université de Lille) ainsi qu'un partenaire industriel (Orange Labs). Débuté en septembre 2010, l'objectif du projet PERCOL était d'avoir un consortium pour la participation au défi REPERE. De manière générale, le but était de proposer un système de détection et de reconnaissance de personnes dans des documents audiovisuels.

Ce projet m'a permis de continuer sur des thématiques initiées pendant ma thèse, mon année d'ATER et mon post-doctorat, en me familiarisant à l'outil de RAP du LIA, nommé Speeral [Linares et al., 2007], me permettant par la suite de m'intégrer dans des collaborations industrielles (voir partie 9.4). Cela m'a également permis d'obtenir des premières collaborations au LIA sur la problématique de recherche de noms de personnes dans le flux parlé, que ce soit avec des approches acoustiques et par espaces de thèmes [Senay et al., 2013], ou bien avec des approches hybrides avec des modèles de contextes continus [Bigot et al., 2013b,a].

9.2 Robustesse des représentations de documents

Le spectre de mes thématiques de recherche s'est ensuite élargi pour travailler à la fois sur des documents écrits et parlés, avec comme point central la problématique de la représentation des mots contenus dans les documents en vue de leur utilisation pour d'autres tâches. Mes travaux m'ont alors amené à travailler sur des représentations de documents de plus haut niveau que le simple niveau *mot*, que l'on retrouve en particulier dans le projet ANR SuMACC (sous-partie 9.2.1). La prise en compte de l'aspect temporel des mots et des documents a également fait partie de mes orientations scientifiques, que l'on retrouve dans le projet ANR ContNomina (sous-partie 9.2.2).

9.2.1 Participation au projet ANR SuMACC (2013-2014)

Tout comme le projet PERCOL, j'ai pu rapidement m'intégrer dans le projet SuMACC [Morchid et al., 2014i], débuté en 2010, porté par le LIA, et coordonné par Georges Linarès. Ce projet avait pour partenaires le laboratoire EURECOM, ainsi que les entreprises Wikio et Syllabs. La problématique développée concernait la recherche d'entités sur Internet, prenant alors en considération l'hétérogénéité des documents traités sur ce média, ainsi que l'explosion constante du nombre de documents nouveaux mis en ligne. L'intérêt ici était de porter l'étude sur le côté multimédia, avec des travaux sur la vidéo, le texte et l'audio.

Ce projet a été ma première expérience de co-encadrement, avec G. Linarès, d'un étudiant en thèse (Mohamed Morchid). Mes travaux ont évolué vers le traitement de documents issus du web, alors que l'ensemble de mes travaux se concentrait sur des émissions télévisuelles ou radiophoniques. La structure des documents ainsi que les problématiques différentes développées m'ont, en particulier, permis d'évoluer sur ma manière d'appréhender le contenu textuel et audio, beaucoup plus bruité dans ce contexte d'étude, et de travailler sur des approches nouvelles, comme par exemple les travaux pour la représentation de documents par espaces de thèmes latents [Morchid et al., 2013b] ou l'application de l'approche i-vecteur, issue de la reconnaissance du locuteur, pour les documents textuels [Morchid et al., 2014a,b].

9.2.2 Participation au projet ANR ContNomina (2013-2017)

Le projet ANR ContNomina est le premier projet auquel j'ai pu participer dès son lancement en septembre 2013. Le sujet concernait la reconnaissance des noms propres dans les documents diachroniques audio. Le caractère *diachronique* de ces noms propres, qui évoluent donc dans le temps, constituent l'originalité du projet, puisque des nouveaux noms apparaissent continuellement dans les médias, ce qui oblige les systèmes de RAP, qui ont un vocabulaire fermé, à gérer dynamiquement les dictionnaires de mots ainsi que les modèles de langage associés. Il s'agissait d'exploiter au mieux le contexte d'apparition des mots pour fournir des stratégies de recherche,

de détection, et de modélisation de nouveaux mots pour permettre une meilleure transcription automatique. En plus du LIA, dont le responsable scientifique était G. Linarès, le projet était coordonné par le LORIA (Université de Lorraine).

Les travaux auxquels j'ai pu participer se sont concentrés sur la représentation du contexte des mots en sortie des systèmes de transcription, et plus particulièrement sur la modélisation du contexte au moyen de représentations par des modèles contextuels continus [Bigot et al., 2013b] ou encore par plongement lexical (*word embeddings*), que l'on retrouve dans les travaux de thèse de Killian Janod [Janod et al., 2015, 2016a]. J'ai également supervisé la mise en place de l'outil de visualisation des noms propres dans les transcriptions automatiques, réalisé par Mathias Quillot.

9.3 Traitement automatique du langage et interdisciplinarité

Les problématiques en traitement du langage, intégrant une forte interdisciplinarité, ont pris une place de plus en plus importante dans les travaux de recherche que je mène depuis plusieurs années. Ceux-ci se concrétisent, notamment, dans trois projets de recherche ANR, à savoir les projets GaFes (sous-partie 9.3.1), TheVoice (sous-partie 9.3.2) et RePoGa (sous-partie 9.3.3). En particulier, ceux-ci ont fait émerger des problématiques d'évaluation originales, nécessitant la mise en place de cadres expérimentaux souvent inexistants. Nous avons donc oeuvré à la proposition de nouveaux cadres de travail dont la dimension *sciences humaines* a été un enjeu important, nous permettant également de proposer des approches originales comme par exemple pour l'exploration des réseaux sociaux numériques (projets GaFes et RePoGa) ou l'étude de la voix jouée (projet TheVoice).

9.3.1 Participation au projet ANR GaFes (2015-2018)

Le projet ANR GaFes a eu pour but d'analyser l'activité significative sur le web pour mieux comprendre les pratiques festivalières et développer des approches permettant la visualisation et l'accès aux contenus générés sur Internet dans et autour des festivals. Les problèmes scientifiques soulevés relèvent à la fois de l'informatique, mais également des sciences humaines et sociales, en particulier de la sociologie des publics. Le projet a été coordonné par le LIA, par G. Linarès, en collaboration avec le laboratoire EURECOM, les entreprises Syllabs et GECE, ainsi que les sociologues du CNE (Centre Norbert Elias).

Le projet GaFes a constitué une évolution dans mes activités de recherche dont les problématiques sont encore, à l'heure actuelle, au coeur de mon projet scientifique. Il est le premier projet au volet pluridisciplinaire auquel j'ai pu participer. De par son aspect original, le projet a permis la réalisation de travaux scientifiques autour de l'exploration des données, notamment par rapport à leur temporalité [Quillot et al., 2017], ou encore sur la détection d'opinions dans

les tweets [Dufour et al., 2018]. Il a également servi de plateforme de démonstration, réalisée pendant l'alternance de M. Quillot, à destination du grand public (voir partie 11.4), et utilisée dans le cadre des travaux de recherche et enquêtes des sociologues.

9.3.2 Participation au projet ANR TheVoice (2018-_)

L'innovation du projet ANR TheVoice tient dans sa manière de s'intéresser à la création de voix pour la production de contenu pour le domaine artistique audiovisuel (jeux vidéos, séries, films...). En particulier, il s'agit ici de pouvoir modéliser la *palette vocale* d'un acteur afin de pouvoir recommander de nouvelles voix à partir de celle-ci, pour le doublage vocal par exemple. Il est nécessaire de dépasser le cadre d'une simple reconnaissance acoustique, telle que celle-ci peut être entreprise en reconnaissance et vérification du locuteur par exemple, pour extraire des caractéristiques liées aux choix subjectifs effectués par un opérateur humain ainsi qu'à la réception de la voix par le public. L'IRCAM est ici le coordinateur du projet, et l'entreprise Dubbing Brothers s'occupe de l'industrialisation des outils ainsi que de la mise à disposition des données. Au LIA, la responsable scientifique est Jean-François Bonastre.

Dans la continuité des travaux initiés avec le projet GaFes, j'ai pu travailler, dans le cadre du projet The Voice, sur une problématique mêlant, par essence, le TAL et les sciences humaines pour tout ce qui concerne la perception et la réception de la voix. Ces travaux sont issus de la thèse d'Adrien Gresse sur le doublage vocal et la représentation *personnage* [Gresse et al., 2017, 2019], ainsi que ceux de la thèse de Mathias Quillot.

9.3.3 Responsable scientifique Informatique du projet RePoGa (2020)

Le projet RePoGa a été proposé dans le cadre des élections municipales 2020, avec pour objectif principal d'utiliser des sources d'information disponibles sur Internet pour comprendre les réseaux d'interaction des acteurs politiques d'un territoire ciblé (dans notre contexte, le Grand Avignon) et analyser les contenus qu'ils partagent sur les réseaux sociaux numériques (RSN) tels que Twitter et Facebook. Le projet est ici local à Avignon Université, qui le finance via la Fédération de Recherche (FR) Agorantic ayant comme tutelle le CNRS et l'université, porté par des politologues via leur laboratoire LBNC. De mon côté, je suis le responsable scientifique pour la partie informatique du projet.

Les travaux qui y sont menés sont proches de ceux développés pendant le projet ANR GaFes. L'objectif est d'explorer les possibilités de fouille et d'extraction automatique d'informations en ligne à partir des listes des candidats déclarés aux élections, afin d'enrichir les données biographiques et contextuelles. En particulier, il s'agira d'être capable de construire des réseaux d'interaction à partir des différentes personnalités politiques étudiées et de leur présence médiatique sur Internet. L'analyse de ces réseaux d'interaction s'appuie sur les techniques de représentation des réseaux complexes, telles que des graphes d'interaction [Papegnies et al., 2019]. Le projet a

permis de financer 3 mois Malek Hajjem, docteure en informatique et spécialiste en fouille de données, en tant qu'ingénieure de recherche.

9.4 Collaborations industrielles

Sous la direction de G. Linarès, et ce dès mon arrivée, j'ai pu m'impliquer dans des collaborations industrielles avec les entreprises Orkis et EDD, qui se sont principalement concrétisées par les thèses CIFRE respectives de Killian Janod et de Mohamed Bouaziz, toutes deux soutenues en 2017 (voir parties 10.1.2 et 10.1.3). Durant le contrat qui lie le LIA à l'entreprise EDD, toujours d'actualité pendant l'écriture de ce document, j'ai pu travailler à leur fournir le système de RAP du LIA (Speeral), en particulier sur l'aspect modélisation du langage et mise à jour des modèles. La collaboration m'a amené à les former sur les outils permettant de créer et d'adapter des modèles de langage, mais également de leur fournir les connaissances essentielles à la compréhension théorique de ces modèles.

J'ai enfin eu l'opportunité d'être responsable scientifique pour le projet *Topping automatique pour la synchronisation répliques/sous-titres dans le cadre du spectacle vivant*, porté par Jean-François Bonastre, en collaboration et financé par l'entreprise Atelier 144. Ce projet, visant à fournir un outil de synchronisation automatique en temps réel de sous-titres lors d'un spectacle vivant (ici une pièce de théâtre), a permis de financer un stagiaire Master 2 (Anthony Poujade) qui a travaillé sur l'application de méthodes issues du traitement automatique du locuteur.

9.5 Conclusion

Que ce soit depuis mes premiers travaux scientifiques, en thèse, et depuis mon recrutement en tant que maître de conférences au sein du LIA, j'ai eu la chance et l'opportunité de participer à une multitude de projets de recherche qui m'ont conduit, en partie, à l'orientation scientifique de mes travaux actuels. Grâce à ces collaborations, j'ai pu participer au montage de certains projets, comme les projets ANR GaFes et The Voice. Cette expérience m'a poussé à proposer de nouveaux projets de recherche et de nouvelles orientations scientifiques, comme le projet Agorantic RePoGa, ou encore le projet ANR DIETS, accepté en 2020 sur le diagnostic automatique des erreurs des systèmes de transcription de parole bout-en-bout à partir de leur réception par les utilisateurs, qui constitue le point de départ de mes perspectives de recherche, et dont je donne de plus amples informations dans le chapitre V. Enfin, au niveau européen, je suis impliqué dans le projet SELMA (accepté en 2020), dont je vais être responsable du workpackage 3 : *Joint Multilingual and User-Feedback Transfer Learning.*

Chapitre 10

Encadrement scientifique

Sommaire

Dans ce chapitre, je détaille mes contributions et mon implication dans l'encadrement de jeunes chercheurs. Dans un premier temps, je présente les thèses que j'ai eu l'opportunité de co-encadrer ainsi que celles que je co-encadre actuellement, puis je m'intéresse aux stages de Master Recherche pour lesquels j'ai participé à l'encadrement, ainsi qu'au suivi d'un étudiant en alternance à vocation recherche.

10.1 Thèses

10.1.1 Thèse de Mohamed Morchid (2012-2014)

La thèse de Mohamed Morchid, financée sur le projet ANR SuMACC (voir partie 9.2.1), a débuté en octobre 2011 avant mon arrivée au LIA. Initialement dirigée par Georges Linarès seul, j'ai rejoint l'encadrement à mon arrivée en tant que maître de conférences. Les travaux de thèse de M. Morchid se sont principalement concentrés sur la classification automatique de messages textuels bruités, que l'on retrouve dans des documents de différentes natures (transcriptions automatiques, messages courts sur les réseaux sociaux...), les bruits apparaissant sous différentes

formes (erreurs de transcription, agrammaticalité, fautes d'orthographe, vocabulaire spécifique et non standard...). Ces travaux font echo à ce que j'ai pu rencontrer lors de ma thèse concernant la transcription automatique de la parole spontanée. M. Morchid a alors proposé des approches originales pour la représentation de ces messages bruités, pour en compenser ou en atténuer le bruit, en proposant de dépasser le simple niveau *mot* au travers de représentations abstraites. Parmi les approches proposées, nous pouvons par exemple citer l'application de la représentation par espace de thèmes *author-topic* [Morchid et al., 2015d], ou l'approche multi-vue, s'appuyant sur des représentations par espace de thèmes multiples et leur fusion au moyen de l'analyse factorielle [Morchid et al., 2014a]. La qualité de ces représentations a pu être évaluée sur des tâches de classification de documents textuels et audio.

M. Morchid a soutenu sa thèse intitulée *Représentations robustes de documents bruités dans des espaces homogènes* le 25 novembre 2014. Pendant sa thèse, il a effectué 4 mois à Microsoft Research Cambridge. Après une année en tant qu'ATER, il exerce maintenant en tant que maître de conférences au LIA depuis septembre 2015. Les travaux menés pendant sa thèse ont donné lieu à de très nombreuses publications nationales et internationales (23 au total) de premier plan, dont les principales sont :

- revue internationale Computer Speech & Language [Morchid et al., 2016b] ;
- revue internationale IEEE/ACM Transactions on Audio, Speech, and Language Processing [Morchid et al., 2015a] ;
- article dans la conférence internationale EMNLP [Morchid et al., 2014a] ;
- articles dans la conférence internationale ISCA Interspeech [Morchid et al., 2014d,b, 2015c, 2016a] ;
- article dans la conférence internationale IEEE ICASSP [Morchid et al., 2014e] ;
- article dans la conférence internationale IEEE ASRU [Morchid et al., 2015b] ;
- article dans la conférence internationale IEEE SLT [Morchid et al., 2014c] ;
- article dans la conférence internationale ISMIR [Morchid et al., 2014g].

10.1.2 Thèse de Killian Janod (2013-2017)

Killian Janod a effectué sa thèse sous la direction de G. Linarès, ainsi que mon co-encadrement et celui de M. Morchid. La thèse a été financée par une bourse CIFRE, en collaboration avec l'entreprise Orkis. Les travaux menés par Kilian pendant sa thèse ont concerné la thématique de la compréhension de la parole. En considérant toujours le fait que les documents que nous avions à traiter étaient bruités, nous avons poursuivi avec K. Janod les travaux sur la robustesse de documents textuels. Ces travaux interviennent au début des avancées majeures que nous connaissons actuellement en apprentissage automatique, et il a notamment contribué à proposer des méthodes d'abstraction et de débruitage de documents textuels s'appuyant sur les réseaux de neurones, ici des approches de type auto-encodeurs débruitants, permettant d'améliorer les

performances de systèmes de compréhension de la parole.

K. Janod a soutenu sa thèse, intitulée *La représentation des documents par réseaux de neurones pour la compréhension de documents parlés* le 27 novembre 2017. Suite à sa thèse, il a travaillé plusieurs années comme expert en science des données, et est actuellement embauché à ce titre dans l'entreprise Alten. Des publications nationales et internationales ont été publiées dans le cadre de sa thèse :

- revue internationale IEEE/ACM Transactions on Audio, Speech, and Language Processing [Janod et al., 2017] ;
- article dans la conférence internationale IEEE SLT [Janod et al., 2016b] ;
- article dans la conférence internationale ISCA Interspeech [Janod et al., 2016d] ;
- article dans la conférence nationale TALN [Janod et al., 2015] ;
- article dans la conférence nationale JEP [Janod et al., 2016c] ;
- article dans la conférence nationale CORIA [Janod et al., 2016a].

10.1.3 Thèse de Mohamed Bouaziz (2014-2017)

Débutée en janvier 2014, j'ai co-encadré la thèse de Mohamed Bouaziz avec M. Morchid, sous la direction de G. Linarès. Cette thèse CIFRE a constitué le point de départ de la collaboration avec l'entreprise EDD. Ce travail s'inscrit dans la volonté de l'entreprise de proposer des services innovants sur l'audio, l'entreprise collectant et analysant des centaines de flux audio en continu et en parallèle provenant de différentes sources d'information (chaînes télévisées, radios, plateformes de partage de vidéos sur Internet...). L'originalité du travail de M. Bouaziz s'est alors située sur la combinaison de deux problématiques, à savoir la prise en considération de la séquentialité de flux audiovisuels dans leur traitement, mais également le fait de pouvoir tirer profit de flux se déroulant en parallèle (dans notre contexte d'étude, la diffusion de programmes de plusieurs chaînes de télévision). Nous pouvons par exemple citer une partie de ses travaux, où il a proposé une approche neuronale, nommée PLSTM (Parallel Long-Short Term Memory), étendant le concept des LSTM, traitant tout d'abord chaque séquence dans une couche récurrente indépendante, puis en sommant ces différentes sorties, permettant d'obtenir la sortie finale [Bouaziz et al., 2016d]. Cette représentation multi-flux a montré une amélioration sur une tâche de classification, en comparaison de l'utilisation de flux séparés.

M. Bouaziz a soutenu sa thèse le 6 décembre 2017, avec pour titre *Réseaux de neurones récurrents pour la classification de séquences dans des flux audiovisuels parallèles*. Après une année passée dans l'entreprise Airbus Defence and Space suite à sa thèse en tant que chercheur R&D, il est actuellement chercheur R&D spécialisé en traitement du langage pour l'entreprise Aquila Data Enabler et consultant pour Engie. Nous avons pu publier plusieurs de ses travaux de thèse dans des conférences nationales et internationales :

- articles dans la conférence internationale IEEE SLT [Bouaziz et al., 2016b,d] ;

— articles dans la conférence nationale JEP [Bouaziz et al., 2016a,c].

10.1.4 Thèse d'Adrien Gresse (2015-2020)

Adrien Gresse a débuté sa thèse en octobre 2015 sous la direction de Jean-François Bonastre, mon co-encadrement, ainsi que celui de Vincent Labatut. Au regard du sujet particulier, à la coloration multidisciplinaire, la thèse a été financée par la Fondation Pierre Bergé d'Avignon Université, sensible à ces sujets. Comme expliqué dans les projets de recherche auxquels j'ai pu participer (voir par exemple la partie 9.3.2), les travaux menés dans le cadre de cette thèse constituent une partie de l'orientation de mon projet scientifique, avec une évolution vers des collaborations interdisciplinaires. Au cours de sa thèse, A. Gresse a travaillé sur la problématique du doublage vocal dans les productions audiovisuelles. Plus particulièrement, il s'est intéressé à la représentation des voix de *personnages*, dépassant largement le cadre classique d'une simple comparaison acoustique entre deux voix. Il a proposé un protocole et un cadre expérimental totalement nouveaux, lui permettant de mener ses expériences, prenant en considération les différents biais pouvant exister dans ce problème de doublage (contenu linguistique, genre...). Il a ensuite cherché à appliquer des méthodes classiques de reconnaissance du locuteur, qui ont rapidement montré leurs limites. La proposition d'une représentation abstraite des personnages, au moyen d'approches neuronales, a constitué ici un premier pas sur ce problème difficile de représentation de voix.

La thèse d'A. Gresse a été soutenue le 6 février 2020 sous le titre *L'art de la voix : caractériser l'information vocale dans un choix artistique*. Au cours de sa thèse, et malgré les difficultés rencontrées sur un sujet qui trouve difficilement sa place dans les problématiques habituelles de nos conférences, nous avons pu publier ses travaux dans plusieurs conférences :

— articles dans la conférence internationale ISCA Interspeech [Gresse et al., 2017, 2020b] ;
— article dans la conférence internationale IEEE ICASSP [Gresse et al., 2019] ;
— articles dans la conférence nationale JEP [Gresse et al., 2018, 2020a].

10.1.5 Thèse de Mathias Quillot (2018-_)

Depuis janvier 2018, Mathias Quillot a débuté sa thèse sous la direction de J.-F. Bonastre, mon co-encadrement ainsi que celle de Nicolas Obin (maître de conférences à l'IRCAM). La direction réunit donc deux laboratoires, s'expliquant par le fait que M. Quillot soit financé par le projet ANR The Voice (voir partie 9.3.2). Finalement, les travaux de M. Quillot sont dans la continuité de ceux initiés par A. Gresse sur le doublage vocal, et plus précisément la représentation de la voix. Sans aller jusqu'à un niveau de définition et d'explicabilité qui pourrait conduire à une représentation précise de la palette vocale d'un acteur, ses travaux de thèse cherchent à fournir une information quant à la description d'une voix de personnage, avec des caractéristiques fines. Il s'agit aussi de montrer clairement si l'on retrouve des informations

dans ces représentations qui rendent compte de la dimension *personnage*, et non simplement d'éléments constitutifs au locuteur.

M. Quillot devrait soutenir sa thèse en 2021. Ses travaux actuels ont été soumis dans des conférences internationales, et il a participé à deux articles acceptés par A. Gresse à ICASSP [Gresse et al., 2019] et Interspeech [Gresse et al., 2020b]. Il a également publié un article dans la conférence nationale JEP [Quillot et al., 2020].

10.1.6 Thèse de Noé Cécillon (2019-_)

Dirigée par G. Linarès, co-encadrée par V. Labatut et moi-même, la thèse de Noé Cécillon a débuté en octobre 2019. La thèse est financée par une bourse ministérielle, et est en lien avec le stage de Master 2 effectué par N. Cécillon (voir partie 10.2.4). Elle fait suite à des premiers travaux de recherche entrepris avec V. Labatut sur une problématique à la frontière entre le domaine du traitement du langage, dans lequel j'ai apporté mon expertise, et le domaine des réseaux complexes, dans lequel V. Labatut effectue sa recherche. Nous avons ainsi eu des premiers résultats très encourageants sur la détection des messages abusifs dans des conversations en modélisant les interactions entre les participants, au moyen de graphes d'interaction [Papegnies et al., 2019]. Il s'agit dans cette thèse de tirer profit de sources d'informations multiples liées au document, qui peuvent prendre la forme de contenus linguistiques, acoustiques, structurels... L'intérêt étant de ne plus considérer indépendamment chacune de ces sources, mais de les combiner en une représentation unique pouvant prendre la forme de plongements (*embeddings*), intégrant au plus tôt toutes ces informations. Cela suit les dernières avancées dans le domaine des graphes, avec les *graphes embeddings*, dont de nombreux pans sont encore à étudier.

Les travaux de N. Cécillon ont déjà bien débuté, avec un article accepté pendant son stage de Master, dans le workshop international Soc2net [Cecillon et al., 2019], un article accepté à la conférence internationale LREC [Cecillon et al., 2020b], où il a proposé un nouveau corpus diffusé librement de messages abusifs dans les conversations Wikipedia, et un résumé étendu avec présentation orale dans la conférence nationale MARAMI [Cecillon et al., 2020a].

10.2 Stages et Alternance

10.2.1 Stage de Licence 2 et Licence 3 de Mathias Quillot (2014)

Les stages de quelques semaines de Mathias Quillot en Licence 2 et Licence 3 ont été orientés sur le développement d'une application dans le cadre du projet ANR ContNomina (voir partie 9.2.2) pour valoriser le projet. Il s'agissait ici d'être capable de visualiser des transcriptions automatiques pendant la diffusion d'une vidéo, en mettant en lumière les avancées scientifiques réalisées pendant le projet, à savoir la récupération et la correction de noms propres apparaissant de façon diachronique.

10.2.2 Alternance de Master de Mathias Quillot (2015-2017)

Mathias a effectué deux années d'alternance, pendant la préparation de son Master, au sein du LIA. Il a alors eu la responsabilité de la valorisation du projet ANR GaFes (voir partie 9.3.1). Il a participé à la mise en place du site Internet du projet [1], et surtout à la création d'une plateforme permettant la visualisation d'entités et leurs relations dans les réseaux sociaux.

Dans les derniers mois de son alternance, M. Quillot a travaillé sur des problématiques orientées recherche, toujours dans le cadre du projet GaFes. Un article a notamment été publié dans la conférence SLSP [Quillot et al., 2017], où nous avons proposé des analyses statistiques et visuelles sur deux représentations différentes de données issues de messages publiés sur Twitter : des plongements temporels de mots (*temporal embeddings*) et l'approche Word2Vec [Mikolov et al., 2013a].

10.2.3 Stage de Master Recherche d'Adrien Gresse (2015)

Le stage de Master 2 d'Adrien Gresse, co-encadré avec G. Linarès, concernait la recommandation de musiques de films. Ce stage fait suite aux travaux entrepris sur une thématique proche pendant la campagne d'évaluation MediaEval 2013 sur la tâche MusiClef (recommandation de musiques pour des publicités) [Morchid et al., 2013a]. Durant ce stage, déjà orienté interdisciplinaire, A. Gresse a proposé d'étudier certaines caractéristiques musicales pouvant conduire à l'émotion dans le choix d'une musique. Des premières expériences ont montré que des propriétés musicales, comme le rythme et le timbre, pourraient jouer un rôle dans la classification d'émotions véhiculées dans une musique (ici, *joyeux* et *triste*).

10.2.4 Stage de Master Recherche de Noé Cecillon (2019)

Avant de débuter sa thèse sous mon co-encadrement, j'ai co-encadré, avec V. Labatut, le stage de Master 2 Recherche de Noé Cécillon. Au cours de son stage, N. Cécillon a pu se former à la fois au domaine du traitement du langage mais également à celui des réseaux complexes. Il a repris la problématique initiée quelques années plus tôt sur la détection d'abus dans les conversations textuelles, en proposant un travail original sur la fusion d'informations provenant à la fois de caractéristiques issues du contenu textuel [Papegnies et al., 2017b], mais également des caractéristiques calculées à partir des graphes conversationnels [Papegnies et al., 2019]. Ce travail a conduit, pendant son stage, à publier un article dans le workshop Soc2Net [Cecillon et al., 2019], où nous avons proposé différentes approches de fusion, mais également une analyse originale des caractéristiques les plus pertinentes pour la détection des messages abusifs.

1. https ://anr-gafes.univ-avignon.fr

10.3 Conclusion

Depuis mon recrutement en tant que maître de conférences au sein du LIA en 2012, j'ai eu l'opportunité de co-encadrer 4 thèses soutenues en 2014, 2017 (x2) et 2019. Parmi ces thèses soutenues, la dernière montre ma volonté de travailler sur des problématiques interdisciplinaires.

Je co-encadre actuellement deux thèses, une sur le traitement de la parole avec des enjeux sociologiques, et une entre le traitement du langage et les réseaux complexes.

Je me suis également investi dans l'encadrement d'étudiants en Master Recherche, qui ont tous poursuivi en thèses financées ensuite, et que j'ai eu l'occasion d'encadrer, ou que je co-encadre actuellement.

Je continue bien entendu mon engagement dans l'encadrement scientifique. Cela me conduit à rechercher des financements, que ce soit au niveau ministériel, dans la demande de financement de bourses de thèses ou de projets ANR, ou bien de contrats CIFRE, dont le contrat avec l'entreprise EDD devrait mener à une nouvelle bourse de thèse. Dans même, le projet ANR JCJC DIETS, accepté en septembre 2020 et dont je suis le responsable scientifique, me permet de financer un nouvel étudiant en thèse qui débutera en 2021 (voir partie V).

Chapitre 11

Rayonnement et vulgarisation

Sommaire

Ce chapitre fournit un panorama de mes actions en faveur de la dissémination de la recherche, qu'elles soient à destination de la communauté scientifique francophone et internationale, mais également pour le grand public. Je présente donc, tout d'abord dans cette partie, mon implication dans la relecture d'articles, dans les sociétés savantes, ainsi que dans des commissions. Je détaille ensuite les campagnes d'évaluation auxquelles j'ai participé ainsi que leur impact. Je rends ensuite compte de la réalisation de présentations scientifiques à destination d'un public varié, avant de terminer par un résumé de mes responsabilités scientifiques et académiques.

11.1 Relectures et sociétés savantes

Au cours de ces dernières années, mes contributions scientifiques m'ont amené à évaluer les travaux d'autres chercheurs en traitement du langage, en particulier en tant que relecteur pour différentes revues et conférences internationales reconnues dans le domaine. Ces relectures peuvent se résumer ainsi :

— **Revues internationales**

1. IEEE Signal Processing Letters - 2013.
2. IEEE/ACM Transactions on Audio, Speech, and Language Processing - 2015.
3. ACM Transactions on Knowledge Discovery from Data - 2018.
4. IEEE Access - 2019.

5. Elsevier Information and Software Technology - 2019.

— **Conférences internationales et workshops internationaux spécialisés**

1. ISCA Interspeech - tous les ans depuis 2010.
2. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - tous les ans depuis 2010.
3. IEEE EUSIPCO - tous les ans depuis 2013.
4. International Conference on Machine Learning and Signal Processing (MALSIP) - 2015.
5. IEEE Symposium on Computers and Communications - 2015.
6. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) - 2017, 2019.
7. IEEE Global Conference on Signal and Information Processing (GlobalSIP) - 2017, 2018.
8. IEEE Spoken Language Technology Workshop (SLT) - 2020.

— **Revue et conférence francophones**

1. Journées d'Étude de la Parole (JEP) - 2010, 2012, 2014, 2016, 2018.
2. Revue Traitement Automatique des Langues (TAL) - Numéro spécial *Apprentissage profond pour le traitement automatique des langues* - 2018.

Enfin, depuis le début de mon parcours scientifique, je suis membre des sociétés savantes internationales ISCA, IEEE, et de la société savante francophone Association Francophone de la Communication parlée (AFCP), au sein de laquelle j'ai eu l'occasion de prendre des responsabilités (voir partie 11.6).

11.2 Commissions d'évaluation et expertises

Depuis mon arrivée au sein du LIA, mes prises de responsabilité m'ont amené à participer à différentes commissions d'évaluation :

— Participation au comité de programme et scientifique de la conférence francophone JEP 2014.

— Membre du comité scientifique de la revue TAL pour le numéro spécial *Apprentissage profond pour le traitement automatique des langues* (2017).

— Participation aux comités de suivi de thèse de Salima Mdhaffar (2019), Amira Barhoumi (2019), Thibault Magallon (2018), Luis Adrian Cabrera Diego (2015), et Ilaria Brunetti (2014).

— Vice-président du comité de sélection pour le poste MCF 27 (profil Intelligence Artificielle) à Avignon Université (2020).

11.3 Campagnes d'évaluation

Dans le cadre des travaux scientifiques que j'ai pu mener, j'ai eu l'opportunité de participer à plusieurs campagnes d'évaluation, qu'elles soient nationales ou internationales. Ces campagnes ont en particulier été l'occasion de collaborer avec des chercheurs au sein de mon laboratoire d'accueil actuel, le LIA, mais également d'autres laboratoires :

— **Campagnes d'évaluation internationales**

1. **MediaEval 2013**. Collaboration à la proposition de systèmes pour trois tâches : *Crowdsourcing*, *MusiClef tracks* et *Spoken Web Search*. Le système proposé pour la tâche *MusiClef tracks* a été publié dans la conférence internationale de rang A ISMIR 2014. Un article collaboratif entre tous les participants à la tâche *Crowdsourcing* (4 laboratoires de recherche internationaux) a été réalisé et publié dans un workshop international avec comité de relecture (CrowdMM 2014).
2. **RepLab 2013**. Collaboration à la proposition d'un système sur la tâche de détection de l'impact (positif / négatif / neutre) sur la réputation d'une marque. Le système est arrivé deuxième sur 5 participants internationaux.
3. **IWSLT 2014**. Collaboration à la proposition d'un système de traduction automatique de textes de l'anglais vers le polonais et de l'anglais vers le slovène. Ce travail m'a permis de travailler sur une nouvelle problématique en collaboration avec Stéphane Huet, MCF au LIA, spécialiste en traduction automatique.
4. **DSTC5 2016**. Collaboration à la proposition d'un système pour la détection de l'état du dialogue à chaque tour de parole. Le système proposé, s'appuyant sur les modèles *author-topic*, a donné lieu à un article au workshop spécialisé de rang A IEEE SLT 2016.

— **Campagnes d'évaluation nationales**

1. **DEFT 2013**. Dans le cadre de cette conférence, qui revient annuellement et très connue dans la communauté scientifique francophone du TALN, j'ai pu participer à la mise en place d'un système de classification de recettes de cuisine selon le niveau de difficulté (3^{e} sur 6 participants) et le type de plat (1er sur 5 participants).
2. **Défi ANR REPERE 2014**. Participation à ce défi au sein du consortium PERCOL (3 laboratoires de recherche français et 1 industriel). L'objectif de ce projet était de reconnaître et nommer les personnes présentes dans des documents télévisés au moyen des différents média disponibles (audio, vidéo, sous-titres...). J'ai pu y participer au sein de problématiques sur l'audio, dans les tâches de transcription automatique et de reconnaissance des noms de personne prononcés. Le système proposé par PERCOL est arrivé en tête dans les tâches principales du défi (sur trois consortiums), et second sur

les tâches orientées audio. Ces travaux ont donné lieu à 4 articles dans des conférences internationales de rang A (Interspeech et ICASSP).

11.4 Dissémination et vulgarisation dans des événements

Grâce à l'expertise scientifique acquise durant ces années, je me suis investi dans différents événements scientifiques (sous-partie 11.4.1), souvent à l'initiative d'invitation de la communauté qu'elle soit scientifique, industrielle, ou à destination du grand public. Dans le cadre de la dissémination de mes activités de recherche, j'ai également pu participer à la mise en place d'enseignements scientifiques, avec création de cours et suivi de projets sous forme de travaux pratiques (sous-partie 11.4.2). Cette activité précise de dissémination de la science, destinée à différents publics (experts et grand public), a eu pour vocation de rendre accessible des notions scientifiques, sur des sujets parfois complexes, et de donner l'envie à des étudiants, professionnels ou simplement esprits curieux de découvrir des thématiques de recherche, et éventuellement de les approfondir ensuite. Ces activités se résument dans les sous-parties suivantes.

11.4.1 Invitations dans des événements scientifiques

Cette sous-partie liste les différents événements auxquels j'ai pu participer et qui sont directement reliés à mes activités scientifiques. Ces événements ont été organisés selon leur destination, à savoir la communauté scientifique, qu'elle soit nationale ou internationale, la communauté industrielle, et enfin le grand public.

Événements à destination de la communauté scientifique

Communauté scientifique internationale

— Invitation à présenter mes travaux de thèse et de post-doctorat à la Carnegie Mellon University (CMU) à Pittsburgh (États-Unis) (mai 2012), sur le sujet *Enriching automatic transcriptions with acoustic and linguistic features.*

— Invitation à proposer un tutoriel *Playing around with CLEF collections, Elastic Search and Kibana*, avec mes collègues Vincent Labatut et Liana Ermakova (MCF Université de Bretagne Occidentale), sur l'utilisation de la suite Elastic pour la collecte, l'indexation, et la visualisation de très grandes bases de données. Ce tutoriel a eu lieu pendant la conférence CLEF 2018.

Communauté scientifique nationale

— Invitation à présenter au LIA mes travaux de recherche sur le sujet *Enrichissement des transcriptions automatiques au moyen de caractéristiques acoustiques et linguistiques* (février 2013).

— Invitation de l'Institut National de l'Audiovisuel (INA) à présenter un début de projet interdisciplinaire sur le débat de l'entre-deux tours de l'élection présidentielle 2017 (juin 2018). L'objectif était de pouvoir réfléchir à la mise en place d'outils pour l'analyse de cet événement sur la plateforme Twitter.
— Invitation à présenter mes travaux de recherche actuels au LIUM, en particulier sur les problématiques de représentation des documents et d'interdisciplinarité en traitement du langage (avril 2018).

Événements à destination des industriels

— Présentations et formations dans le cadre du contrat entre le LIA et l'entreprise EDD depuis 2013. En particulier, j'ai assuré une formation aux outils de modélisation du langage pour la reconnaissance automatique de la parole, et j'ai animé une présentation, pour un public d'informaticiens mais non spécialistes en traitement du langage, sur l'adaptation des modèles de langage avec Georges Linarès (janvier 2020).
— Invitation à présenter, avec mon collègue Mickaël Rouvier, durant la conférence-débat *Intelligence Artificielle : efficiente et responsable ?* (novembre 2019), sur les thématiques des biais en apprentissage automatique et de l'explicabilité dans les réseaux de neurones profonds. Cet événement spécial, organisé pendant les *Journées IA en Région Sud*, était en premier lieu à destination de professionnels, non experts en intelligence artificielle, mais également ouvert à tout public. La présentation est accessible sur le site https://www.canal-u.tv.

Événements à destination du grand public

— Lorsque mes activités me le permettent, je participe à la Fête de la science. J'ai, par exemple, pu présenter la thématique du traitement du langage vers le grand public (2015). J'ai également présenté, avec G. Linarès, une synthèse sur le traitement automatique du langage naturel à l'invitation d'une ville du Vaucluse (2015 - Caumont-sur-Durance). Enfin, j'ai participé à la présentation du projet ANR GaFes ainsi que de l'outil *Observatoire des festivals* développé dans le cadre du projet (2016 et 2017).
— Invitation à une présentation *grand public* lors d'un *Café des Sciences* (octobre 2018), dont l'objectif est de fournir une vulgarisation scientifique, en permettant au public de débattre et de discuter sur un sujet identifié. Avec mes collègues M. Rouvier, MCF en informatique au LIA, et Guillaume Champy, MCF en droit privé et sciences criminelles, nous avons présenté et débattu sur la thématique *Fake news, comment les détecter, comment s'en préserver ?*.
— Participation à divers forums étudiants. Les présentations, parfois scientifiques, sont *grand public* car elles concernent principalement les lycéens et collégiens. Je réalise environ 2

présentations de ce type tous les ans depuis 2012.

— Responsabilité de l'organisation de la Journée Portes Ouvertes du département informatique d'Avignon Université (tous les ans depuis 2013), avec notamment présentation au grand public (étudiants et parents) de certains travaux de recherche réalisés au laboratoire.

11.4.2 Enseignements reliés à mon expertise scientifique

— Invitation à proposer et concevoir un cours sous forme d'un séminaire de 18 heures à destination des étudiants en Master 2 Publics de la Culture et Communication d'Avignon Université (2017). Ce séminaire avait pour vocation de présenter, de manière assez générale, le traitement automatique du langage et son intérêt pour des étudiants avec un profil de *sociologue*. Outre cette thématique, j'ai pu présenter des travaux sur les réseaux sociaux ainsi que sur la représentation des données. Enfin, pendant ce séminaire, en lien avec le projet ANR GaFes, j'ai pu faire travailler les étudiants sur une extraction de données Twitter du festival des Transmusicales, en leur proposant de manipuler des outils d'analyse automatique de ces données.

— Dans le cadre de mes enseignements, j'ai eu l'occasion de participer, dès sa mise en place en 2018, à l'Unité d'Enseignement *Application d'Innovation*, dont l'objectif est de fournir une première entrée dans le monde de la recherche aux étudiants de Master 2 en Informatique d'Avignon Université. Dans ce cadre, j'ai pu participer à la conception du cours, orienté traitement du langage et extraction d'informations, ainsi qu'au projet à réaliser par les étudiants, sur la problématique de la détection automatique de polarité de tweets sur la plateforme Twitter.

11.5 Diffusion de corpus et plateforme d'évaluation

Dans le cadre de la thèse de Noé Cécillon (voir partie 10.1.6), nous avons proposé d'enrichir un corpus existant de messages abusifs dans des commentaires Wikipedia [Cecillon et al., 2020b]. La contribution majeure sur ce corpus concerne la reconstruction des conversations dans ces commentaires, actuellement inexistante. Au final, nous avons pu enrichir environ 380 000 messages. Nous avons également proposé une plateforme d'évaluation complète [1] sur ce corpus, en mettant à disposition toutes les approches que nous allons mettre en place dans la thèse de N. Cécillon, afin de stimuler la communauté scientifique autour de cette problématique.

1. `https://github.com/CompNet/WikiSynch`

11.6 Responsabilités scientifiques et académiques

Depuis mon arrivée au LIA, et en complément de mes activités scientifiques présentées dans les parties précédentes, j'ai eu l'opportunité de prendre diverses responsabilités scientifiques et académiques :

Responsabilités académiques

— Responsable de la Communication du Centre d'Enseignement et de Recherche en Informatique (CERI) d'Avignon Université depuis 2012. L'implication principale concerne la gestion de la participation à une douzaine de forums étudiants, impliquant la coordination des enseignants du CERI (relais des informations, recherche d'intervenants, contact direct avec les responsables des forums...). L'autre volet concerne la communication ponctuelle, vers l'extérieur, d'informations à destination du grand public, en particulier au travers des canaux numériques (site Internet et réseaux sociaux), en lien avec la Cellule Communication d'Avignon Université. De même, j'assure la mise à jour ponctuelle du site Internet du CERI selon les modifications nécessaires (1 à 2 fois par an). Je participe aux réunions sur la qualité de l'établissement, celui-ci étant certifié ISO 9001 (environ 2 réunions par an). Enfin, cette fonction m'amène à m'impliquer au Conseil Pédagogique du CERI (3 à 4 réunions par an), avec un compte-rendu annuel aux enseignants du CERI concernant les actions de communication effectuées dans l'année.

— Responsable du Master Informatique parcours-type *Ingénierie du Logiciel de la Société Numérique* (ILSEN) d'Avignon Université depuis janvier 2020. Cela constitue la gestion d'une des trois spécialités du Master au CERI. Il s'agit de prendre en charge les deux années du Master ILSEN, pour un effectif global d'environ 80 étudiants. Le travail le plus important concerne la mise en place, chaque semestre, de l'emploi du temps. Ce travail nécessite la récupération des contraintes des enseignants du CERI (environ 15) mais également des intervenants extérieurs (une dizaine environ) afin de travailler en coordination avec le bureau des études et des emplois du temps (BEE) pour la mise en place des emplois du temps. Durant les périodes d'enseignement, ce travail nécessite une implication quotidienne sur des tâches ponctuelles (modification de l'emploi du temps, suivi et réponse aux étudiants, gestion des problèmes...). Un compte-rendu annuel sur le déroulé de l'année est également réalisé (questionnaire aux étudiants, statistiques de réussite...). Enfin, il s'agit de réfléchir aux évolutions pédagogiques du Master, que ce soit pour la mise à jour des enseignements (1 fois par an) ou des maquettes complètes (1 fois tous les 4 ans) avec l'ensemble de l'équipe pédagogique du CERI.

Responsabilités scientifiques

— Membre élu suppléant du conseil d'administration de l'Association Francophone de la Communication Parlée (AFCP) pour les mandats 2013-2015 et 2015-2017. Durant ces années, j'ai été amené à participer à quelques réunions, mais surtout à m'impliquer dans

le comité scientifiques de la conférence francophone JEP 2014 (voir partie 11.2).

— Membre élu du conseil scientifique du LIA depuis mars 2014. Cette fonction m'a amené à participer régulièrement aux réunions ainsi qu'aux décisions du laboratoire (environ 5 par an). Il s'agit ici de discuter et voter les orientations scientifiques du laboratoire.

— Coordinateur de l'axe scientifique *Langage & Cognition* de l'Institut Carnot Cognition depuis avril 2020, s'agissant d'un des 4 axes scientifiques structurant l'institut composé de 22 laboratoires. En particulier, cet axe implique 7 laboratoires de recherche, pour environ 130 chercheurs. Le travail le plus important consiste, dans un premier temps, à structurer cet axe en réalisant une cartographie des problématiques scientifiques qui y sont développées. Il s'agira ensuite de proposer des actions pour faire *vivre* cet axe, en proposant par exemple des journées thématiques, afin de permettre aux chercheurs de laboratoires différents d'échanger ensemble, mais également avec des industriels invités. De façon régulière, je participe à environ 2 réunions par mois.

11.7 Conclusion

Depuis mon arrivée au LIA en 2012, mon investissement dans mes activités scientifiques, en particulier sur le volet rayonnement et vulgarisation, a été croissant. De façon constante, je participe à des relectures d'articles, dans des revues et conférences majeures du domaine du traitement du langage, et je m'investis dans les différentes commissions dans lesquelles je suis sollicité. Mes travaux en recherche et mon expertise, mais également mon investissement sur la partie enseignement, m'ont amené à proposer, concevoir et présenter des cours, mais également divers séminaires scientifiques à destination d'un public large, auquel il est nécessaire de s'adapter.

Mes dernières responsabilités, que ce soit au niveau de la responsabilité du Master ILSEN ou de l'animation de l'axe *Langage & Cognition* de l'Institut Carnot Cognition, montrent mon engagement et ma volonté dans les prochaines années de prendre en charge de nouveaux engagements, toujours dans l'objectif d'animer et de diffuser la science au plus grand nombre.

BIBLIOGRAPHIE PERSONNELLE

Sommaire

ACLI : Revues internationales avec comité de lecture (7)

Revues majeures du domaine (6)

R. Dufour, Y. Esteve, and P. Deléglise. Characterizing and detecting spontaneous speech : Application to speaker role recognition. *Speech communication*, 56 :1–18, 2014.

K. Janod, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Denoised bottleneck features from deep autoencoders for telephone conversation analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9) :1809–1820, 2017.

M. Morchid, R. Dufour, P.-M. Bousquet, G. Linarès, and J.-M. Torres-Moreno. Feature selection using principal component analysis for massive retweet detection. *Pattern Recognition Letters*, 49 :33–39, 2014.

M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. Compact multiview representation of documents based on the total variability space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8) :1295–1308, 2015.

M. Morchid, R. Dufour, and G. Linarès. Impact of word error rate on theme identification task of highly imperfect human–human conversations. *Computer Speech & Language*, 38 :68–85, 2016.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*, 6(1) :38–55, 2019.

Autres revues (1)

M. Morchid, J.-M. Torres-Moreno, R. Dufour, J. Ramírez-Rodríguez, and G. Linarès. Automatic text summarization approaches to speed up topic model learning process. *International Journal of Computational Linguistics and Applications*, 2016. ISSN 0976-0962.

ACTI : Communications avec actes dans un congrès international (57)

Congrès internationaux majeurs du domaine (33)

F. Béchet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille, et al. Multimodal understanding for person recognition in video broadcasts. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014.

B. Bigot, G. Senay, G. Linarès, C. Fredouille, and R. Dufour. Combining acoustic name spotting and continuous context models to improve spoken person name recognition in speech. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1584–1588. ISCA, 2013a.

B. Bigot, G. Senay, G. Linarès, C. Fredouille, and R. Dufour. Person name recognition in asr outputs using continuous context models. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 8470–8474. IEEE, 2013b.

M. Bouallegue, M. Morchid, R. Dufour, D. Matrouf, G. Linarès, and R. De Mori. Factor analysis based semantic variability compensation for automatic conversation representation. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014a.

M. Bouallegue, M. Morchid, R. Dufour, D. Matrouf, G. Linarès, and R. De Mori. Subspace gaussian mixture models for dialogues classification. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014b.

M. Bouaziz, M. Morchid, R. Dufour, and G. Linarès. Improving multi-stream classification by mapping sequence-embedding in a high dimensional space. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 224–231. IEEE, 2016a.

M. Bouaziz, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Parallel long short-term memory for multi-stream classification. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 218–223. IEEE, 2016b.

R. Dufour and Y. Estève. Correcting asr outputs : specific solutions to specific errors in french. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 213–216. IEEE, 2008.

R. Dufour and B. Favre. Semi-supervised part-of-speech tagging in speech applications. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2010.

R. Dufour, Y. Estève, P. Deléglise, and F. Béchet. Local and global models for spontaneous speech segment detection and characterization. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, pages 558–561. IEEE, 2009.

R. Dufour, F. Bougares, Y. Estève, and P. Deléglise. Unsupervised model adaptation on targeted speech segments for lvcsr system combination. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2010.

R. Dufour, Y. Estève, and P. Deléglise. Investigation of spontaneous speech characterization applied to speaker role recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2011.

R. Dufour, G. Damnati, and D. Charlet. Automatic error region detection and characterization in lvcsr transcriptions of tv news shows. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4445–4448. IEEE, 2012a.

R. Dufour, G. Damnati, D. Charlet, and F. Béchet. Automatic transcription error recovery for person name recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2012b.

R. Dufour, M. Morchid, and T. Parcollet. Tracking dialog states using an author-topic based representation. In *IEEE Workshop Spoken Language Technology (SLT)*. IEEE, 2016.

A. Gresse, M. Rouvier, R. Dufour, V. Labatut, and J.-F. Bonastre. Acoustic pairing of original and dubbed voices in the context of video game localization. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre. Similarity metric based on siamese neural networks for voice casting. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6585–6589. IEEE, 2019.

A. Gresse, M. Quillot, R. Dufour, and J.-F. Bonastre. Learning voice representation using knowledge distillation for automatic voice casting. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2020.

K. Janod, M. Morchid, R. Dufour, and G. Linarès. A log-linear weighting approach in the word2vec space for spoken language understanding. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 356–361. IEEE, 2016a.

K. Janod, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Deep stacked autoencoders for spoken language understanding. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 720–724. ISCA, 2016b.

S. Mdhaffar, Y. Estève, N. Hernandez, A. Laurent, R. Dufour, and S. Quiniou. Qualitative evaluation of asr adaptation in a lecture context : Application to the pastel corpus. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 569–573. ISCA, 2019.

M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. An i-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 443–454, 2014a.

M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. I-vector based representation of highly imperfect automatic transcriptions. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014b.

M. Morchid, R. Dufour, M. Bouallegue, and G. Linarès. Author-topic based representation of call-center conversations. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 218–223. IEEE, 2014c.

M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and R. De Mori. Theme identification in human-human conversations with features from specific speaker type hidden spaces. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014d.

M. Morchid, R. Dufour, P.-M. Bousquet, M. Bouallegue, G. Linarès, and R. De Mori. Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 126–130. IEEE, 2014e.

M. Morchid, R. Dufour, and G. Linarès. A combined thematic and acoustic approach for a music recommendation service in tv commercials. In *International Conference on Music Information Retrieval Conference (ISMIR)*, pages 465–470, 2014f.

M. Morchid, R. Dufour, and G. Linarès. Topic-space based setup of a neural network for theme identification of highly imperfect transcriptions. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, pages 346–352. IEEE, 2015a.

M. Morchid, R. Dufour, and D. Matrouf. A comparison of normalization techniques applied to latent space representations for speech analytics. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2015b.

M. Morchid, M. Bouaziz, W. Kheder, K. Janod, P.-M. Bousquet, R. Dufour, and G. Linarès. Spoken language understanding in a latent topic-based subspace. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 710–714. ISCA, 2016.

T. Parcollet, M. Morchid, P.-M. Bousquet, R. Dufour, G. Linarès, and R. De Mori. Quaternion neural networks for spoken language understanding. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 362–368. IEEE, 2016.

M. Rouvier, R. Dufour, G. Linares, and Y. Estève. A language-identification inspired method for spontaneous speech detection. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2010.

G. Senay, B. Bigot, R. Dufour, G. Linarès, and C. Fredouille. Person name spotting by combining acoustic matching and lda topic models. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1584–1588. ISCA, 2013.

Autres congrès internationaux du domaine (24)

N. Cecillon, V. Labatut, R. Dufour, and G. Linarès. Abusive language detection in online conversations by combining content-and graph-based features. In *International Workshop on Modeling and Mining Social-media-driven Complex Networks (Soc2Net)*. Frontiers, 2019.

N. Cecillon, V. Labatut, R. Dufour, and G. Linarès. Wac : A corpus of wikipedia conversations for online abuse detection. In *International Conference on Language Resources and Evaluation (LREC)*, 2020.

R. Dufour. From prepared speech to spontaneous speech recognition system : a comparative study applied to french language. In *International Conference on Soft Computing as Transdisciplinary Science and Technology (CSTST)*, pages 595–599, 2008.

R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linares. Spontaneous speech characterization and detection in large audio database. In *International Conference on Speech and Computer (SPECOM)*, 2009.

R. Dufour, Y. Estève, and P. Deléglise. Automatic indexing of speech segments with spontaneity levels on large audio database. In *IEEE International Workshop on Searching spontaneous conversational speech (SSCS)*, pages 39–44, 2010.

R. Dufour, M. Rouvier, A. Delorme, and D. Malinas. Mining events opinion argumentation from raw unlabeled twitter data using convolutional neural network. In *Conference and Labs of the Evaluation Forum (CLEF)*, 2018.

Y. Estève, P. Deléglise, S. Meignier, S. Petitrenaud, H. Schwenk, L. Barrault, F. Bougares, R. Dufour, V. Jousse, A. Laurent, et al. Some recent research work at lium based on the use of cmu sphinx. In *CMU SPUD Workshop*, 2010.

Y. Estève, M. Bouallegue, C. Lailler, M. Morchid, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. Integration of word and semantic features for theme identification in telephone conversations. In *International Workshop Series on Spoken Dialogue Systems Technology (IWSDS)*, pages 223–231. Springer, 2015.

B. Favre, G. Damnati, F. Béchet, M. Bendris, D. Charlet, R. Auguste, S. Ayache, B. Bigot, A. Deltei, R. Dufour, et al. Percoli : a person identification system for the 2013 repere challenge. In *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

B. Loni, J. Hare, M. Georgescu, M. Riegler, X. Zhu, M. Morchid, R. Dufour, and M. Larson. Getting by with a little help from the crowd : Practical approaches to social image labeling. In *International ACM Workshop on Crowdsourcing for Multimedia*, pages 69–74, 2014.

S. Mdhaffar, Y. Estève, A. Laurent, N. Hernandez, R. Dufour, D. Charlet, G. Damnati, S. Quiniou, and N. Camelin. A multimodal educational corpus of oral courses : Annotation, analysis and case study. In *International Conference on Language Resources and Evaluation (LREC)*, 2020.

M. Morchid, R. Dufour, and G. Linarès. Event detection from image hosting services by slightly-supervised multi-span context models. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 103–107. IEEE, 2013a.

M. Morchid, R. Dufour, and G. Linarès. Thematic representation of short text messages with latent topics : Application in the twitter context. In *International Conference of the Pacific Association for Computational Linguistics (PACLING)*, 2013b.

M. Morchid, R. Dufour, and G. Linarès. A lda-based topic classification approach from highly imperfect automatic transcriptions. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1309–1314, 2014a.

M. Morchid, R. Dufour, U. Niaz, F. Bouvier, C. de Groc, C. de Loupy, G. Linarès, B. Merialdo, and B. Peralta. Sumacc project's corpus : A topic-based query extention approach to retrieve multimedia documents. In *International Conference on Text, Speech and Dialogue (TSD)*, 2014b.

M. Morchid, G. Linarès, and R. Dufour. Characterizing and predicting bursty events : The buzz case study on twitter. In *International Conference on Language Resources and Evaluation (LREC)*, pages 2766–2771, 2014c.

M. Morchid, R. Dufour, G. Linarès, and Y. Hamadi. Latent topic model based representations for a robust theme identification of highly imperfect automatic transcriptions. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pages 596–605. Springer, 2015a.

M. Morchid, D. Josselin, Y. Portilla, R. Dufour, E. Altman, and G. Linarès. A topic modeling based representation to detect tweet locations. In *ISPRS Geospatial Week*, 2015b.

M. Morchid, Y. Portilla, D. Josselin, R. Dufour, E. Altman, M. El-Beze, J.-V. Cossu, G. Linarès, and A. Reiffers-Masson. An author-topic based approach to cluster tweets and mine their location. *Procedia Environmental Sciences*, 27 :26–29, 2015c.

M. Morchid, J.-M. Torres-Moreno, R. Dufour, J. Ramírez-Rodríguez, and G. Linarès. Automatic text summarization approaches to speed up topic model learning process. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pages 87–109. Springer, 2017.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Graph-based features for automatic online abuse detection. In *International Conference on Statistical Language and Speech Processing (SLSP)*, pages 70–81. Springer, 2017a.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Impact of content features for automatic online abuse detection. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pages 404–419. Springer, 2017b.

M. Quillot, C. Ollivier, R. Dufour, and V. Labatut. Exploring temporal analysis of tweet content from cultural events. In *International Conference on Statistical Language and Speech Processing (SLSP)*, pages 82–93. Springer, 2017.

M. Rouvier, R. Dufour, and P.-M. Bouquet. Review of different robust x-vector extractors for speaker verification. In *European Signal Processing Conference (EUSIPCO)*, 2020.

ACTN : Communications avec actes dans un congrès national (19)

F. Béchet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille, et al. Identification de personnes dans des flux multimédia. In *Conférence en Recherche d'Information et Applications (CORIA)*, pages 239–251, 2015.

B. Bigot, G. Senay, G. Linares, C. Fredouille, and R. Dufour. Modèles contextuels continus pour la reconnaissance des noms de personnes dans des transcriptions automatiques. In *Journées d'Étude sur le Parole (JEP)*, 2014.

M. Bouaziz, M. Morchid, P.-M. Bousquet, R. Dufour, K. Janod, W. B. Kheder, and G. Linarès. Un sous-espace thématique latent pour la compréhension du langage parlé. In *Journées d'Étude sur le Parole (JEP)*, 2016a.

M. Bouaziz, M. Morchid, R. Dufour, G. Linarès, and P. Correa. Un corpus de flux tv annotés pour la prédiction de genres. In *Journées d'Étude sur le Parole (JEP)*, 2016b.

R. Dufour, Y. Estève, and P. Deléglise. Corrections spécifiques du français sur les systèmes de reconnaissance automatique de la parole. In *Rencontre des Jeunes Chercheurs en Parole (RJCP)*, 2009.

R. Dufour, Y. Esteve, P. Deléglise, and F. Béchet. Utilisation conjointe de modeles locaux et globaux pour la caractérisation et la détection de segments de parole spontanée. In *Journées d'Étude sur le Parole (JEP)*, page 113, 2010.

R. Dufour, G. Damnati, and D. Charlet. Détection et caractérisation des régions d'erreurs dans des transcriptions de contenus multimédia : application à la recherche des noms de personnes. In *Journées d'Étude sur le Parole (JEP)*, 2012a.

R. Dufour, A. Laurent, and Y. Estève. Combinaison d'approches pour la reconnaissance du rôle des locuteurs. In *Journées d'Étude sur le Parole (JEP)*, 2012b.

A. Gresse, R. Dufour, V. Labatut, M. Rouvier, and J.-F. Bonastre. Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix. In *Journées d'Étude sur le Parole (JEP)*, 2018.

A. Gresse, M. Quillot, R. Dufour, and J.-F. Bonastre. Apprentissage automatique de représentation de voix à l'aide d'une distillation de la connaissance pour le casting vocal. In *Journées d'Étude sur le Parole (JEP)*, 2020.

K. Janod, M. Morchid, R. Dufour, and G. Linares. Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, 2015.

K. Janod, M. Morchid, R. Dufour, and G. Linarès. Réseaux de neurones pour la représentation des contextes continus des mots. In *Conférence en Recherche d'Information et Applications (CORIA)*, pages 656–668, 2016a.

K. Janod, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Auto-encodeurs pour la compréhension de documents parlés. In *Journées d'Étude sur le Parole (JEP)*, 2016b.

M. Morchid, R. Dufour, and G. Linarès. Combinaison de thèmes latents pour la contextualisation de tweets. In *Conférence sur l'Extraction et la Gestion des Connaissances (EGC)*, 2013.

M. Morchid, R. Dufour, G. Linarès, and R. de Mori. Classification de transcriptions automatiques imparfaites : Doit-on adapter le calcul du taux d'erreur-mot ? In *Journées d'Étude sur le Parole (JEP)*, 2014.

M. Morchid, R. Dufour, and G. Linarès. Initialisation de réseaux de neurones à l'aide d'un espace thématique. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, 2015.

E. Papegnies, R. Dufour, V. Labatut, and G. Linarès. Détection de messages abusifs au moyen de réseaux conversationnels. In *Conférence sur les Modèles et l'Analyse des Réseaux (MARAMI)*, 2017a.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Detection of abusive messages in an online community. In *Conférence en Recherche d'Information et Applications (CORIA)*, 2017b.

M. Quillot, L. Guillou, A. Gresse, R. Ferro, R. Roth, D. Malinas, , R. Dufour, A. Roebel, N. Obin, J.-F. Bonastre, and E. Ethis. La voix actée : pratiques, enjeux, et applications. In *Journées d'Étude sur le Parole (JEP)*, 2020.

Campagnes d'évaluation avec actes (6)

X. Bost, I. Brunetti, L. A. Cabrera-Diego, J.-V. Cossu, A. Linhares, M. Morchid, J.-M. Torres-Moreno, M. El-Bèze, and R. Dufour. Systemes du lia a deft'13. In *Défi Fouille de Textes (DEFT)*, 2017.

M. Bouallegue, G. Senay, M. Morchid, D. Matrouf, G. Linarès, and R. Dufour. Lia@ mediaeval 2013 spoken web search task : An i-vector based approach. In *Benchmarking Initiative for Multimedia Evaluation (MediaEval)*, 2013.

J. V. Cossu, B. Bigot, L. Bonnefoy, M. Morchid, X. Bost, G. Senay, R. Dufour, V. Bouvier, J.-M. Torres-Moreno, and M. El-Bèze. Lia@replab 2013. In *Evaluation Campaign for Online Reputation Management systems (RepLab)*, 2013.

M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and D. Matrouf. Lia@ mediaeval 2013 crowdsourcing task : Metadata or not metadata ? that is a fashion question. In *Benchmarking Initiative for Multimedia Evaluation (MediaEval)*, 2013a.

M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and D. Matrouf. Lia@ mediaeval 2013 musiclef task : A combined thematic and acoustic approach. In *Benchmarking Initiative for Multimedia Evaluation (MediaEval)*, 2013b.

M. Morchid, S. Huet, and R. Dufour. A topic-based approach for post-processing correction of automatic translations. In *International Conference on Spoken Language Translation (IWSLT)*, 2014.

Thèses (2)

R. Dufour. Représentation des connaissances sémantiques dans le cadre du dialogue homme-machine finalisé. Rapport de Master Recherche, Le Mans Université, 2007.

R. Dufour. *Transcription automatique de la parole spontanée.* Thèse de doctorat en informatique, Le Mans Université, 2010.

Cinquième partie

Conclusion et Perspectives de recherche

Conclusions et Perspectives

Bilan personnel

Au travers de ce manuscrit, j'ai présenté mes principales activités en traitement automatique du langage (TAL) ces dernières années, en continuité de mes travaux de thèse en reconnaissance automatique de la parole (RAP) spontanée. Les différents projets auxquels j'ai pu participer ont grandement enrichi les thèmes et domaines sur lesquels je travaille actuellement. En particulier, ceux-ci ont en commun trois axes qui se retrouvent dans la plupart de mes travaux de recherche, et qui guideront mes recherches futures : la représentation robuste de documents, l'étude des systèmes automatiques, et enfin leur évaluation.

J'ai tout d'abord présenté, dans la partie I, une partie des travaux dans lesquels je me suis investi concernant la robustesse des représentations de documents textuels. Nous avons tout d'abord proposé une approche par réseaux de neurones permettant de tirer profit d'informations issues de documents parallèles, *i.e.* plusieurs flux de données diffusés pendant une même période de temps. Nous avons également proposé d'améliorer la prise en compte du contexte dans le cadre de représentations par plongements de mots (*word embeddings*) en pondérant les mots du contexte en fonction de leur position par rapport au mot ciblé. Nous nous sommes ensuite intéressés à la représentation de transcriptions automatiques fortement bruitées (langage atypique, forte agrammaticalité, très grand nombre d'erreurs de transcription...). Cela nous a permis de comparer deux représentations de document classiques, s'appuyant soit sur les mots directement (TF-IDF), soit sur un niveau de représentation plus élevé au moyen d'espaces de thèmes avec l'approche LDA. Comme attendu, l'approche par espace de thèmes a permis d'atteindre de meilleures performances, mais a également révélé une instabilité des résultats liée aux choix des hyper-paramètres des modèles. Une approche, adaptant l'approche i-vecteur, issue de la reconnaissance du locuteur, au domaine de la représentation de documents, a alors été proposée afin d'obtenir une représentation compacte à partir d'un très grand nombre de vues (espaces de thèmes) d'un même document.

Puis j'ai mis en avant, dans la partie II, la problématique de l'évaluation en traitement du langage, et ce, en particulier, concernant la RAP. Une première étude a été proposée sur une tâche de classification de documents mettant en parallèle le taux d'erreur-mot (WER) des transcriptions automatiques et la performance de la classification automatique s'appuyant sur celles-ci. De ce travail, nous avons mis en avant le fait que le WER ne reflète que peu la *qualité* de la transcription. J'ai ensuite présenté des travaux menés sur la détection et la caractérisation

de régions d'erreurs dans les transcriptions automatiques, avant de proposer une correction *a posteriori* des noms de personne. Bien qu'ayant un impact limité sur le WER, ces erreurs spécifiques se doivent d'être traitées, de par leur importance dans de nombreuses applications. Enfin, j'ai exposé une continuité de réflexion sur l'évaluation des systèmes de RAP sur des travaux récents de transcription de cours en ligne.

Enfin, la partie III a été consacrée aux travaux collaboratifs se trouvant à la frontière entre le domaine du traitement automatique langage, qui constitue le point d'ancrage de mes recherches, et d'autres domaines de recherche, que ce soit en informatique, avec les réseaux complexes, ou en sciences humaines, avec la sociologie. J'ai mis en avant plusieurs études que nous avons réalisées sur l'exploitation des réseaux sociaux pour l'analyse d'événements, en particulier sur la plateforme Twitter. Outre une meilleure compréhension des tâches étudiées (relais massif des messages, plongements lexicaux pour l'analyse d'événements culturels, argumentation d'opinions...), il en est ressorti une difficulté réelle pour les évaluer, manquant très souvent de cadres expérimentaux bien définis. J'ai ensuite décrit les travaux que nous avons menés sur la détection des messages abusifs, avec une approche intégrant la structure des conversations, ignorant le contenu textuel. Ces travaux ont notamment mené à la mise en place d'une plateforme d'expérimentation et la diffusion libre d'un corpus de conversations. Enfin, les derniers travaux présentés se sont focalisés sur le doublage vocal et la recommandation de voix, problématique de recherche particulièrement difficile car faisant intervenir des éléments liés à la réception des utilisateurs, hautement subjective, et difficilement mesurable par une machine.

Ces différents travaux ont pu avoir lieu grâce à la dynamique existante au LIA dès mon arrivée en tant que maître de conférences. J'ai tout de suite pu m'impliquer dans le projet ANR PERCOL, qui faisait suite aux travaux que j'ai menés pendant mon année de post-doctorat à Orange Labs. S'en est suivie une implication dans différents projets ANR, tels que SuMACC, ContNomina, GaFes, ou encore le projet The Voice. Je m'implique activement de plus en plus dans le montage de projets. J'ai également eu la chance de participer au co-encadrement de différentes thèses, dont 4 déjà soutenues (Mohamed Morchid, Killian Janod, Mohamed Bouaziz et Adrien Gresse) et 2 en cours (Mathis Quillot et Noé Cécillon). Enfin, j'essaie de participer régulièrement à des campagnes d'évaluation (MediaEval, DEFT, DSTC...), qui me semblent importantes à la fois pour maintenir des systèmes au niveau international et pour aider les jeunes chercheurs à valoriser leur travail. J'ai enfin pris la responsabilité de coordonner et d'animer l'axe *Langage & Cognition* de l'Institut Carnot, impliquant 7 laboratoires de recherche.

Perspectives générales

Le bilan personnel a présenté une partie des travaux ainsi que des activités que j'ai pu mener après mes travaux de thèse réalisés au sein du LIUM. Ceux-ci ont principalement été menés au

LIA depuis mon recrutement. Plusieurs perspectives à ces travaux sont déjà en cours, alors que d'autres constitueront les axes de recherche que je souhaite développer. Les prochaines sous-parties résument ces axes, incluant des perspectives à plus long terme. Je conclue ce manuscrit en présentant le projet ANR JCJC DIETS, accepté en 2020 et devant démarrer au cours de l'année 2021, et qui constituera le coeur de mes travaux dans les prochaines années.

Représentation de documents par plongements (*embeddings*) multi-sources

Un des axes principaux que je souhaite continuer à développer concerne la représentation des documents, thème récurrent dans mes recherches. Plus particulièrement, je vais m'investir dans le développement d'approches par plongements (*embeddings*), dont la performance a été démontrée dans de nombreux domaines ces dernières années. Nous l'avons notamment vu dans le chapitre 1 pour le TAL avec les plongements de mots. En ce moment, nous travaillons, avec Vincent Labatut et Noé Cécillon, à appliquer des approches de plongements de graphes (*graph embeddings*) à notre problème de détection des messages abusifs. En effet, l'approche par graphes conversationnels que nous avons proposée comporte plusieurs limites liées à :

— *Définition des caractéristiques.* Les mesures topologiques utilisées pour représenter le graphe ont été choisies manuellement et le plus large possible pour couvrir le maximum d'informations utiles que l'on pourrait extraire. Nous avons vu, notamment par l'analyse des top-caractéristiques, que seule une partie des celles-ci sont utiles. Nous pouvons cependant nous poser la question de savoir si certaines mesures ne sont pas manquantes, ou si certaines d'entre elles n'existent tout simplement pas.

— *Temps de traitement.* Nous avons également vu que ce travail d'extraction des caractéristiques par mesure topologique est coûteux, à la fois pour les choisir initialement, mais également en termes de temps de traitement machine.

Les plongements de graphes résolvent en partie ces problèmes en proposant d'extraire automatiquement, de façon non supervisée, les informations contenues dans un graphe et de le représenter dans un vecteur de dimension réduite devant conserver ses propriétés topologiques. Nos premières études montrent que ces vecteurs peuvent rivaliser avec nos précédents résultats, et donc nos caractéristiques choisies manuellement. Un article sur la détection de messages abusifs au moyen de plongements de graphes a été soumis à la revue internationale Springer Nature : Computer Science, et est actuellement en cours de révision.

Une des limites des approches par plongement actuelles concerne le fait qu'elles sont souvent proposées pour représenter des objets et/ou médias particuliers, indépendamment les uns des autres (audio, vidéo, graphe, texte...). Or, dans le chapitre 7, nous avons vu que la combinaison d'informations issues du texte et de la structure des conversations étaient complémentaires. Des approches ont commencé à représenter des mots au sein de graphes pour ensuite en créer des plongements, mais cela reste finalement une représentation liée aux mots. Je souhaite donc

continuer sur cette idée en permettant la construction de graphes de représentation à partir de contenus multi-sources.

Dans le cadre d'un document multimédia, cela pourrait revenir, par exemple, à construire des représentations par *embeddings* à deux niveaux : une première reviendrait à extraire des représentations *embeddings* de chaque source de données dans un même espace de représentation, puis de les combiner sous la forme d'un graphe, dont les noeuds seraient ces *embeddings*, et les liens des relations entre ces noeuds selon leur proximité vectorielle. Il s'agirait ensuite ici d'avoir un *embedding* de ce graphe pour obtenir une représentation multi-sources. Nous devrions avoir des premiers résultats sur ce type d'approche hybride dans le cadre de la thèse de Noé Cécillon.

Évaluation des systèmes de TAL par le prisme de l'utilisateur final

Comme nous l'avons vu en filigrane de ce manuscrit, la notion de performance n'est pas anodine et conditionne, pour beaucoup, l'évolution d'un domaine de recherche. Les environnements expérimentaux complets (*benchmark*) ont été, et sont toujours, une réponse au problème d'évaluation et de comparaison entre les systèmes développés.

Nous construisons donc des machines qui minimisent les erreurs sur une tâche précise selon une référence considérée, sans essayer de connaître, de comprendre et d'évaluer l'impact de cette erreur du point-de-vue de l'utilisateur final. Or, même les humains font des erreurs : lorsqu'ils s'en aperçoivent, ceux-ci vont chercher à les compenser, soit en les corrigeant, soit en les ignorant. Ce processus est bien évidemment le même lorsque des utilisateurs vont prendre en main un système automatique : ils vont devoir *gérer* ces erreurs. L'impact de ces erreurs faites par les systèmes automatiques sur les humains, et la manière dont ils les perçoivent, n'est alors jamais évalué par les métriques d'évaluation existantes. L'analyse, et les métriques d'évaluation, sont pour l'instant orientées *système* et non orientées *humain*, alors même que ces systèmes leur sont destinés et sont conçus pour modéliser le langage humain.

Je souhaite développer cet axe concernant l'évaluation des systèmes en m'intéressant à l'impact cognitif des erreurs. Une des premières possibilités est de travailler sur les erreurs de transcription, domaine dans lequel je travaille depuis de nombreuses années et dans lequel j'ai eu l'occasion de réfléchir à leur évaluation. Cela pourrait être possible à travers des tests perceptifs mis en place pour analyser les réactions d'utilisateurs face aux erreurs de transcription. Un tel corpus de données couplé à des tests perceptifs sur les erreurs de systèmes de RAP n'existe pas actuellement, et constituerait une réelle avancée dans le domaine, et plus généralement, en TAL.

Étude des erreurs des systèmes d'apprentissage profond

En lien étroit avec l'axe précédent, je souhaiterais également étudier les erreurs dans les systèmes d'apprentissage profond, au travers des systèmes de RAP dits de bout-en-bout (*end-to-end*). Dans ces approches, plusieurs niveaux d'abstraction (acoustique, phonétique, lexical,

syntaxique...) sont intégrés dans un modèle unique par réseaux de neurones, alors que les systèmes de transcription classiques intégraient, jusqu'alors, chaque module séparément (modèles acoustiques, de langage...). L'étude des erreurs des systèmes de RAP classiques, comme nous l'avons vu dans ce manuscrit, n'est pas nouvelle en soit et a été assez largement menée. Les systèmes de bout-en-bout font l'objet d'un intérêt actuel très fort de la part de la communauté scientifique, ces systèmes réussissant maintenant à rivaliser avec ceux plus classiques et bien établis. Cependant, la compréhension de ces systèmes, et la manière dont les informations sont traitées à l'intérieur de ces réseaux profonds, n'en est qu'à ses prémisses.

Je souhaite débuter plusieurs travaux sur les systèmes de RAP de bout-en-bout. Au niveau analytique, il serait intéressant de pouvoir, dans un premier temps, comparer les erreurs réalisées de ces systèmes avec les architectures de RAP classiques : types d'erreurs, leur nature grammaticale et linguistique, qualité du signal... Il s'agirait surtout de dresser un état des lieux des erreurs en RAP. Notre intérêt est ici d'intégrer dans cette étude les tests perceptifs mis en place dans l'axe précédent. De même, j'aimerais intégrer une analyse fine de ces erreurs au niveau linguistique, travail qui devrait être réalisé en collaboration avec le LIUM, et en particulier Jane Wottawa, maîtresse de conférences en linguistique, qui partage ces centres d'intérêts.

Enfin, suite à ces études, je souhaiterais travailler sur la problématique de visualisation des erreurs de transcription dans ces systèmes d'apprentissage profond. En effet, de nombreuses techniques de visualisation des réseaux de neurones existent, mais celles-ci sont quasiment exclusivement réservées à l'analyse d'images, de par la nature de ce média. Ces approches permettent notamment de mettre en valeur certaines régions, dans l'image, qui ont été utilisées par les réseaux profonds pour la définir. Par exemple, des algorithmes de visualisation ont été proposés pour les réseaux de neurones convolutifs (CNN), mettant en évidence ce que chaque couche de convolution utilise pour l'analyse d'images. Les techniques de visualisation du signal acoustique, et en particulier en RAP, sont pour l'instant inexistantes. Sachant la nature de ce que nous souhaitons visualiser, nous pouvons imaginer adapter les approches déjà proposées pour l'image, pour, par exemple, mettre en avant les régions où les erreurs se produisent, et ce, selon leur nature linguistique et/ou leur impact au niveau de l'utilisateur final. Cette perspective de recherche est en lien direct avec projet ANR DIETS que je présente dans la dernière partie (voir partie V), intégrant notamment les idées développées précédemment concernant l'évaluation des systèmes de TAL par le prisme de l'utilisateur final.

Perspectives à long terme

Les axes de recherche présentés précédemment, et que je développerai dans les prochaines années, ont tous un point commun : ceux-ci requièrent une ouverture du TAL à d'autres thématiques et domaines de recherche, ce que j'essaie d'initier, à mon niveau, depuis plusieurs dernières années. En effet, j'ai eu l'occasion de voir, sur une décennie, l'évolution du domaine du TAL :

au départ cantonné dans les laboratoires de recherche, ce domaine s'est ouvert progressivement au milieu industriel, pour s'imposer aujourd'hui au grand public. De nombreuses technologies liées au langage font maintenant partie intégrante de la vie des utilisateurs (assistants vocaux, recherche documentaire, traduction automatique, orientation automatique de patients atteints du coronavirus suite à un appel téléphonique de malades...).

Il y a donc des attentes sociétales de plus en plus fortes concernant les technologies qui s'immiscent dans leur vie quotidienne. Ceci est particulièrement important depuis l'avénement des approches par apprentissage profond, systèmes souvent entraînés sur des très grands corpus de données. Le danger se situe dans le fait d'être incapable de décrire ce que la machine apprend et sur quelles données celle-ci s'appuie pour prendre une décision. La communauté scientifique a donc un rôle très important à jouer, précisément sur le côté explicatif des données, outils et algorithmes mis à leur disposition, mais aussi sur le contrôle des données utilisées (présence de biais ? données personnelles ?...). Il semble indispensable d'entreprendre une approche pluridisciplinaire, puisque les problématiques de TAL ont tendance, maintenant, à toucher de nombreux autres domaines scientifiques. Par exemple, si nous prenons la problématique des *fake news*, nous pouvons le voir du point-de-vue informatique (détection automatique), mais également épidémiologique (propagation d'une *fake news*), politique (réseaux d'influence), juridique (réponse pénale), sociologique (impact sur la société)... Mon engagement dans mes travaux de recherche à long terme suivront donc cette voie pluridisciplinaire, qui me semble indispensable pour faire progresser tout domaine de recherche. Remettre et comprendre l'humain face à des systèmes automatiques me paraît indispensable et indissociable.

Il semble enfin important, pour mener à bien cet objectif de transparence et compréhension des algorithmes et données, de continuer l'effort débuté dans la thèse de Noé Cécillon sur la diffusion d'un cadre expérimental complet, incluant les outils développés et les données pour les évaluer. Je m'efforcerai, dès que possible, à diffuser les outils et données produits, mais aussi les résultats scientifiques obtenus que ce soit dans des conférences scientifiques ou des événements grands publics, ce que j'essaie déjà de réaliser dès que l'occasion se présente.

Le projet ANR JCJC DIETS

Le projet ANR JCJC (Jeunes Chercheuses - Jeunes Chercheurs) DIETS [2], accepté en septembre 2020 suite à l'appel à projets générique 2020 [3], est dans la continuité de mes perspectives de recherche exposées précédemment. D'un point de vue général, le projet DIETS propose de se focaliser sur la problématique du diagnostic et de l'évaluation des systèmes de RAP de bout-en-bout en intégrant la réception humaine des erreurs de transcription. Le défi est double : 1)

2. DIETS : Diagnostic automatique des erreurs des systèmes de transcription de parole end-to-end à partir de leur réception par les utilisateurs

3. https://anr.fr/fr/detail/call/appel-a-projets-generique-2020/

analyser finement les erreurs de RAP par rapport à une réception humaine ; et 2) comprendre et détecter comment ces erreurs se manifestent dans un cadre de RAP de bout-en-bout, dont les travaux s'inspirent du cerveau humain.

La contribution majeure du projet se situera également au niveau du corpus produit. Ainsi, nous proposons de fournir un ensemble de conversations audio transcrites et annotées manuellement, ainsi que les transcriptions automatiques de ces conversations. En nous concentrant sur les erreurs de transcription, nous enrichirons le corpus avec les résultats de différents tests perceptifs concernant la réception des transcriptions et de leurs erreurs par les utilisateurs finaux, en prenant alors en considérant leur impact cognitif. Nous fournirons également une annotation détaillée de ces erreurs de transcription en tenant compte d'une analyse linguistique fine. Nous diffuserons à la communauté l'ensemble du protocole expérimental lié à la création de ce corpus, afin de pouvoir créer facilement d'autres corpus similaires plus rapidement, éventuellement dans de nouveaux langages (dans le cadre du projet, seul le français sera considéré). À notre connaissance, il n'existe pas actuellement de base de données permettant de travailler sur la réception humaine des erreurs de systèmes de TAL.

Bien que le projet soit financé en tant que *jeune chercheur*, celui-ci est le reflet d'une volonté commune entre plusieurs thématiques de collaborer : outre l'informatique et le traitement automatique du langage, nous retrouvons les domaines de la linguistique, avec Jane Wottawa, maîtresse de conférences au LIUM, de la psychologie cognitive, avec Arnaud Rey, chercheur CNRS au LPC, et la collecte et annotation de corpus, avec Thierry Bazillon, ingénieur spécialiste des données dans l'entreprise AlloMedia. Avec un budget de 180k euros, le projet me permettra de financer en particulier un étudiant en thèse, qui travaillera sur l'étude et le diagnostic automatique des erreurs de transcription à partir de la réception finale dans le cadre de systèmes de RAP de bout-en-bout. Jane Wottawa participera à l'encadrement scientifique et apportera son expertise en linguistique et sur les études perceptives. Le projet financera également un chercheur post-doctoral pendant 8 mois, spécialisé en sciences cognitives, afin d'apporter des compétences supplémentaires au TAL. Le projet ANR DIETS me permet, finalement, une indépendance financière et scientifique dans mes travaux de recherche. Le projet durant 42 mois, celui-ci me permet de travailler sereinement sur une nouvelle branche scientifique au sein du LIA et d'assumer de nouvelles responsabilités. Ce financement sera, à terme, un tremplin pour des demandes plus importantes, que ce soit dans d'autres instituts de recherche multidisciplinaires (par exemple, l'*Institute of Language, Communication and the Brain* - ILCB), dont lequel le LIA est impliqué, ou pour des financements européens.

Ce projet est l'aboutissement des travaux et réflexions auxquels j'ai pu m'investir en TAL. Il reprend des problématiques historiques du domaine (évaluation, analyse des erreurs, compréhension des systèmes...) tout en espérant insuffler une manière nouvelle de les traiter, dont nous pensons que la solution se trouve non pas exclusivement dans l'environnement technique et

technologique (données, puissance de calcul, apprentissage automatique...), mais dans la manière dont l'humain utilise et s'approprie l'outil automatique. Il me semble évident qu'une meilleure connaissance de l'utilisateur final conduira à une meilleure compréhension des systèmes automatiques, et donc à une voie d'amélioration non plus guidée par la performance brute d'un système automatique, mais par des critères qualitatifs fondés sur l'humain. Il m'apparaît évident que ces travaux ne pourront se faire sans une étroite collaboration entre chercheurs issus de thématiques et de cultures scientifiques différentes, tout en ayant bien conscience, pour avoir participé à plusieurs projets de ce type, que de tels travaux nécessitent du temps : le projet DIETS devrait poser des premières avancées dans ces réflexions entre systèmes automatiques et réception humaine.

BIBLIOGRAPHIE GÉNÉRALE

A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet. Feel : a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3) :833–855, 2017.

J. Authier and A. Meunier. Norme, grammaticalité et niveaux de langue. *Langue française*, (16) :49–62, 1972.

P. Basile, A. Caputo, and G. Semeraro. Analysing word meaning over time by exploiting temporal random indexing. In *Italian Conference on Computational Linguistics CLiC-it*, 2014.

T. Bazillon, V. Jousse, F. Béchet, Y. Estève, G. Linarès, and D. Luzzati. La parole spontanée : transcription et traitement. *Revue Traitement Automatique des Langues (TAL)*, 49(3), 2008.

F. Béchet and E. Charton. Unsupervised knowledge acquisition for extracting named entities from speech. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5338–5341. IEEE, 2010.

F. Béchet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot. Decoda : a call-centre human-human spoken conversation corpus. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1343–1347, 2012.

B. Bigot, G. Senay, G. Linarès, C. Fredouille, and R. Dufour. Combining acoustic name spotting and continuous context models to improve spoken person name recognition in speech. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1584–1588. ISCA, 2013a.

B. Bigot, G. Senay, G. Linarès, C. Fredouille, and R. Dufour. Person name recognition in asr outputs using continuous context models. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 8470–8474. IEEE, 2013b.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022, 2003.

V. Bolotaeva and T. Cata. Marketing opportunities with social networks. *Journal of Internet Social Networking and Virtual Communities*, 2010 :1–8, 2010.

M. Bouaziz. *Réseaux de neurones récurrents pour la classification de séquences dans des flux audiovisuels parallèles*. Thèse de doctorat en informatique, Avignon Université, 2017.

M. Bouaziz, M. Morchid, P.-M. Bousquet, R. Dufour, K. Janod, W. B. Kheder, and G. Linarès. Un sous-espace thématique latent pour la compréhension du langage parlé. In *Journées d'Étude sur le Parole (JEP)*, 2016a.

M. Bouaziz, M. Morchid, R. Dufour, and G. Linarès. Improving multi-stream classification by mapping sequence-embedding in a high dimensional space. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 224–231. IEEE, 2016b.

M. Bouaziz, M. Morchid, R. Dufour, G. Linarès, and P. Correa. Un corpus de flux tv annotés pour la prédiction de genres. In *Journées d'Étude sur le Parole (JEP)*, 2016c.

M. Bouaziz, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Parallel long short-term memory for multi-stream classification. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 218–223. IEEE, 2016d.

A. Bouchekif, G. Damnati, and D. Charlet. Intra-content term weighting for topic segmentation. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 7113–7117. IEEE, 2014.

P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

N. Cecillon, V. Labatut, R. Dufour, and G. Linarès. Abusive language detection in online conversations by combining content-and graph-based features. In *International Workshop on Modeling and Mining Social-media-driven Complex Networks (Soc2Net)*. Frontiers, 2019.

N. Cecillon, R. Dufour, V. Labatut, and G. Linarès. Tuning graph2vec with node labels for abuse detection in online conversations. In *Conférence sur les Modèles et l'Analyse des Réseaux (MARAMI)*, 2020a.

N. Cecillon, V. Labatut, R. Dufour, and G. Linarès. WAC : A Corpus of Wikipedia Conversations for Online Abuse Detection. In *International Conference on Language Resources and Evaluation (LREC)*, 2020b.

S. Chabi. De l'importance des réseaux sociaux en marketing. *Reflets et perspectives de la vie économique*, 47(2) :95–102, 2008.

C.-C. Chang and C.-J. Lin. LIBSVM : A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3) :1–27, 2011.

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4) :359–394, 1999.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*, pages 71–80. IEEE, 2012.

G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *ACM Conference on Hypertext and Social Media*, pages 123–132, 2012.

G. Csardi, T. Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5) :1–9, 2006.

N. Dave. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International journal for advance research in engineering and technology*, 1(6) :1–4, 2013.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4) : 788–798, 2010.

E. Delcroix, S. Proulx, and J. Denouël. *Les réseaux sociaux sont-ils nos amis ? : Un débat sur l'impact de leur utilisation.* Editions Le Muscadier, 2016.

K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *International AAAI Conference on Weblogs and Social Media*, 2011.

C. Dubois and D. Charlet. Using textual information from lvcsr transcripts for phonetic-based spoken term detection. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4961–4964. IEEE, 2008.

R. Dufour. From prepared speech to spontaneous speech recognition system : a comparative study applied to french language. In *International Conference on Soft Computing as Transdisciplinary Science and Technology (CSTST)*, pages 595–599, 2008.

R. Dufour. *Transcription automatique de la parole spontanée.* Thèse de doctorat en informatique, Le Mans Université, 2010.

R. Dufour and Y. Estève. Correcting asr outputs : specific solutions to specific errors in french. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 213–216. IEEE, 2008.

R. Dufour, Y. Estève, P. Deléglise, and F. Béchet. Local and global models for spontaneous speech segment detection and characterization. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, pages 558–561. IEEE, 2009.

R. Dufour, F. Bougares, Y. Estève, and P. Deléglise. Unsupervised model adaptation on targeted speech segments for lvcsr system combination. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2010.

R. Dufour, G. Damnati, and D. Charlet. Automatic error region detection and characterization in lvcsr transcriptions of tv news shows. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4445–4448. IEEE, 2012a.

R. Dufour, G. Damnati, and D. Charlet. Détection et caractérisation des régions d'erreurs dans des transcriptions de contenus multimédia : application à la recherche des noms de personnes. In *Journées d'Étude sur le Parole (JEP)*, 2012b.

R. Dufour, G. Damnati, D. Charlet, and F. Béchet. Automatic transcription error recovery for person name recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2012c.

R. Dufour, A. Laurent, and Y. Estève. Combinaison d'approches pour la reconnaissance du rôle des locuteurs. In *Journées d'Étude sur le Parole (JEP)*, 2012d.

R. Dufour, M. Rouvier, A. Delorme, and D. Malinas. Mining events opinion argumentation from raw unlabeled twitter data using convolutional neural network. In *Conference and Labs of the Evaluation Forum (CLEF)*, 2018.

N. Duta, R. Schwartz, and J. Makhoul. Analysis of the errors produced by the 2004 bbn speech recognition system in the darpa ears evaluations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 14(5) :1745–1753, 2006.

L. Ermakova, L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, and E. Sanjuan. Cultural microblog contextualization 2016 workshop overview : data and pilot tasks. In *Conference and Labs of the Evaluation Forum (CLEF)*, 2016.

L. Ermakova, L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, and E. SanJuan. Clef 2017 microblog cultural contextualization lab overview. In *Conference and Labs of the Evaluation Forum (CLEF)*, pages 304–314. Springer, 2017.

B. Favre, G. Damnati, F. Béchet, M. Bendris, D. Charlet, R. Auguste, S. Ayache, B. Bigot, A. Deltei, R. Dufour, et al. Percoli : a person identification system for the 2013 repere challenge. In *First Workshop on Speech, Language and Audio in Multimedia*, 2013.

M. Federico and R. De Mori. Language modelling. *Spoken Dialogues with Computers*, pages 199–230, 1998.

T. D. Gautheir. Detecting trends using spearman's rank correlation coefficient. *Environmental forensics*, 2(4) :359–362, 2001.

J.-L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech communication*, 37(1-2) :89–108, 2002.

S. Ghannay, B. Favre, Y. Esteve, and N. Camelin. Word embedding evaluation and combination. In *International Conference on Language Resources and Evaluation (LREC)*, pages 300–305, 2016.

S. Goldwater, D. Jurafsky, and C. D. Manning. Which words are hard to recognize ? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3) :181–200, 2010.

A. Gresse. *L'art de la voix : caractériser l'information vocale dans un choix artistique*. Thèse de doctorat en informatique, Avignon Université, 2020.

A. Gresse, M. Rouvier, R. Dufour, V. Labatut, and J.-F. Bonastre. Acoustic pairing of original and dubbed voices in the context of video game localization. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

A. Gresse, R. Dufour, V. Labatut, M. Rouvier, and J.-F. Bonastre. Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix. In *Journées d'Étude sur le Parole (JEP)*, 2018.

A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre. Similarity metric based on siamese neural networks for voice casting. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 6585–6589. IEEE, 2019.

A. Gresse, M. Quillot, R. Dufour, and J.-F. Bonastre. Apprentissage automatique de représentation de voix à l'aide d'une distillation de la connaissance pour le casting vocal. In *Journées d'Étude sur le Parole (JEP)*, 2020a.

A. Gresse, M. Quillot, R. Dufour, and J.-F. Bonastre. Learning voice representation using knowledge distillation for automatic voice casting. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2020b.

C. Guinaudeau, G. Gravier, and P. Sébillot. Improving asr-based topic segmentation of tv programs with confidence measures and semantic relations. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2010.

M. Hajjem, J. V. Cossu, C. Latiri, and E. SanJuan. Clef mc2 2018 lab overview. In *Conference and Labs of the Evaluation Forum (CLEF)*, pages 302–308. Springer, 2018.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv :1605.09096*, 2016.

Z. S. Harris. Distributional structure. *Word*, 10(2-3) :146–162, 1954.

M. A. Hearst. Texttiling : Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1) :33–64, 1997.

K. Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2) :246–268, 1936.

T. B. N. Hoang and J. Mothe. Predicting information diffusion on twitter – analysis of predictive features. *Journal of Computational Science*, 28 :257–264, 2018.

S. Hochreiter and J. Schmidhuber. LSTM can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997.

L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *International Conference on World Wide Web (WWW)*, pages 57–58, 2011.

H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv :1702.08138*, 2017.

Y. Hua, C. Danescu-Niculescu-Mizil, D. Taraborelli, N. Thain, J. Sorensen, and L. Dixon. Wikiconv : A corpus of the complete conversational history of a large online collaborative community. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2818–2823, 2018.

K. Janod. *La représentation des documents par réseaux de neurones pour la compréhension de documents parlés*. Thèse de doctorat en informatique, Avignon Université, 2017.

K. Janod, M. Morchid, R. Dufour, and G. Linares. Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, 2015.

K. Janod, M. Morchid, R. Dufour, and G. Linarès. Réseaux de neurones pour la représentation des contextes continus des mots. In *Conférence en Recherche d'Information et Applications (CORIA)*, pages 656–668, 2016a.

K. Janod, M. Morchid, R. Dufour, and G. Linarès. A log-linear weighting approach in the word2vec space for spoken language understanding. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 356–361. IEEE, 2016b.

K. Janod, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Auto-encodeurs pour la compréhension de documents parlés. In *Journées d'Étude sur le Parole (JEP)*, 2016c.

K. Janod, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Deep stacked autoencoders for spoken language understanding. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 720–724. ISCA, 2016d.

K. Janod, M. Morchid, R. Dufour, G. Linarès, and R. De Mori. Denoised bottleneck features from deep autoencoders for telephone conversation analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 25(9) :1809–1820, 2017.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter : understanding microblogging usage and communities. In *WebKDD and SNA-KDD Workshop on Web mining and social network analysis*, pages 56–65, 2007.

K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

K. S. Jones and J. R. Galliers. *Evaluating natural language processing systems : An analysis and review*, volume 1083. Springer Science & Business Media, 1995.

M. Karan and J. Šnajder. Preemptive toxic language detection in wikipedia comments using thread-level context. In *Workshop on Abusive Language Online*, pages 129–134, 2019.

F. H. Khan, S. Bashir, and U. Qamar. Tom : Twitter opinion mining framework using hybrid classification scheme. *Decision support systems*, 57 :245–257, 2014.

J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. 2001.

C. Lailler, A. Landeau, F. Béchet, Y. Estève, and P. Deléglise. Enhancing the ratp-decoda corpus with linguistic annotations for performing a large range of nlp tasks. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1047–1050, 2016.

J. Laver. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31 :1–186, 1980.

O. Le Deuff. Autorité et pertinence vs popularité et influence : réseaux sociaux sur internet et mutations institutionnelles. 2006.

K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *IEEE International Conference on Data Mining Workshops*, pages 251–258. IEEE, 2011.

J. Lilleberg, Y. Zhu, and Y. Zhang. Support vector machines and word2vec for text classification with semantic features. In *IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pages 136–140. IEEE, 2015.

G. Linares, P. Nocéra, D. Massonie, and D. Matrouf. The lia speech recognition system : from 10xrt to 1xrt. In *International Conference on Text, Speech and Dialogue (TSD)*, pages 302–308. Springer, 2007.

Z. Luo, M. Osborne, J. Tang, and T. Wang. Who will retweet me ? finding retweeters in twitter. In *ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 869–872, 2013.

J. Makhoul, F. Kubala, R. Schwartz, R. Weischedel, et al. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252. Herndon, VA, 1999.

J. Mauclair, Y. Esteve, S. Petit-Renaud, and P. Deléglise. Automatic detection of well recognized words in automatic speech transcriptions. In *LREC*, pages 793–798. Citeseer, 2006.

A. K. McCallum. Mallet : A machine learning for language toolkit. *http ://mallet.cs.umass.edu*, 2002.

K. McDougall. Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. *International Journal of Speech, Language & the Law*, 20(2), 2013.

S. Mdhaffar. *Reconnaissance de la parole dans le contexte de cours magistraux : évaluation, avancées et enrichissement.* Thèse de doctorat en informatique, Le Mans Université, 2020.

S. Mdhaffar, Y. Estève, N. Hernandez, A. Laurent, R. Dufour, and S. Quiniou. Qualitative evaluation of asr adaptation in a lecture context : Application to the pastel corpus. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 569–573. ISCA, 2019.

S. Mdhaffar, Y. Estève, A. Laurent, N. Hernandez, R. Dufour, D. Charlet, G. Damnati, S. Quiniou, and N. Camelin. A multimodal educational corpus of oral courses : Annotation, analysis and case study. In *International Conference on Language Resources and Evaluation (LREC)*, 2020.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013a.

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

P. Mishra, H. Yannakoudakis, and E. Shutova. Neural character-based composition models for abuse detection. *Workshop on Abusive Language Online*, 2018.

M. Morchid. *Représentations robustes de documents bruités dans des espaces homogènes*. Thèse de doctorat en informatique, Avignon Université, 2014.

M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and D. Matrouf. Lia@ mediaeval 2013 musiclef task : A combined thematic and acoustic approach. In *Benchmarking Initiative for Multimedia Evaluation (MediaEval)*, 2013a.

M. Morchid, R. Dufour, and G. Linarès. Thematic representation of short text messages with latent topics : Application in the twitter context. In *International Conference of the Pacific Association for Computational Linguistics (PACLING)*, 2013b.

M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. An i-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 443–454, 2014a.

M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. I-vector based representation of highly imperfect automatic transcriptions. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014b.

M. Morchid, R. Dufour, M. Bouallegue, and G. Linarès. Author-topic based representation of call-center conversations. In *IEEE Workshop Spoken Language Technology (SLT)*, pages 218–223. IEEE, 2014c.

M. Morchid, R. Dufour, M. Bouallegue, G. Linarès, and R. De Mori. Theme identification in human-human conversations with features from specific speaker type hidden spaces. In

Conference of the International Speech Communication Association (INTERSPEECH). ISCA, 2014d.

M. Morchid, R. Dufour, P.-M. Bousquet, M. Bouallegue, G. Linarès, and R. De Mori. Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 126–130. IEEE, 2014e.

M. Morchid, R. Dufour, P.-M. Bousquet, G. Linarès, and J.-M. Torres-Moreno. Feature selection using principal component analysis for massive retweet detection. *Pattern Recognition Letters*, 49 :33–39, 2014f.

M. Morchid, R. Dufour, and G. Linarès. A combined thematic and acoustic approach for a music recommendation service in tv commercials. In *International Conference on Music Information Retrieval Conference (ISMIR)*, pages 465–470, 2014g.

M. Morchid, R. Dufour, and G. Linarès. A lda-based topic classification approach from highly imperfect automatic transcriptions. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1309–1314, 2014h.

M. Morchid, R. Dufour, U. Niaz, F. Bouvier, C. de Groc, C. de Loupy, G. Linarès, B. Merialdo, and B. Peralta. Sumacc project's corpus : A topic-based query extention approach to retrieve multimedia documents. In *International Conference on Text, Speech and Dialogue (TSD)*, 2014i.

M. Morchid, M. Bouallegue, R. Dufour, G. Linarès, D. Matrouf, and R. De Mori. Compact multiview representation of documents based on the total variability space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(8) :1295–1308, 2015a.

M. Morchid, R. Dufour, and G. Linarès. Topic-space based setup of a neural network for theme identification of highly imperfect transcriptions. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, pages 346–352. IEEE, 2015b.

M. Morchid, R. Dufour, and D. Matrouf. A comparison of normalization techniques applied to latent space representations for speech analytics. In *Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2015c.

M. Morchid, Y. Portilla, D. Josselin, R. Dufour, E. Altman, M. El-Beze, J.-V. Cossu, G. Linarès, and A. Reiffers-Masson. An author-topic based approach to cluster tweets and mine their location. *Procedia Environmental Sciences*, 27 :26–29, 2015d.

M. Morchid, M. Bouaziz, W. Kheder, K. Janod, P.-M. Bousquet, R. Dufour, and G. Linarès. Spoken language understanding in a latent topic-based subspace. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 710–714. ISCA, 2016a.

M. Morchid, R. Dufour, and G. Linarès. Impact of word error rate on theme identification task of highly imperfect human–human conversations. *Computer Speech & Language*, 38 :68–85, 2016b.

D. Murthy. Towards a sociological understanding of social media : Theorizing twitter. *Sociology*, 46(6) :1059–1073, 2012.

M. Naili, A. H. Chaibi, and H. H. B. Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112 :340–349, 2017.

N. Obin and A. Roebel. Similarity search of acted voices for automatic voice casting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 24(9) :1642–1651, 2016.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Graph-based features for automatic online abuse detection. In *International Conference on Statistical Language and Speech Processing (SLSP)*, pages 70–81. Springer, 2017a.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Impact of content features for automatic online abuse detection. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, pages 404–419. Springer, 2017b.

E. Papegnies, V. Labatut, R. Dufour, and G. Linarès. Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*, 6(1) :38–55, 2019.

C. Parada, M. Dredze, D. Filimonov, and F. Jelinek. Contextual information improves oov detection in speech. In *Human Language Technology conference (HLT-NAACL)*, pages 216–224. Association for Computational Linguistics, 2010.

J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. Deep learning for user comment moderation. In *Workshop on Abusive Language Online*, pages 25–35. ACL, 2017. URL `http://www.aclweb.org/anthology/W/W17/W17-30.pdf`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn : Machine learning in Python. *Journal of machine Learning research (JMLR)*, 12 :2825–2830, 2011.

M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados : user classification in twitter. In *ACM SIGKDD International Conference on Knowledge discovery and data mining (SIGKDD)*, pages 430–438, 2011.

J. Pennington, R. Socher, and C. D. Manning. Glove : Global vectors for word representation. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

R. Phil. Differences and distinguishability in the acoustic characteristics of hello in voices of similar-sounding speakers. *Australian Review of Applied Linguistics*, 22(1) :1–42, 1999.

J.-P. Poli. *Structuration automatique de flux télévisuels.* Thèse de doctorat en informatique, Université Paul Cézanne – Aix-Marseille III, 2007.

J.-P. Poli. An automatic television stream structuring system for television archives holders. *Multimedia systems*, 14(5) :255–275, 2008.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011.

M. Quillot, C. Ollivier, R. Dufour, and V. Labatut. Exploring temporal analysis of tweet content from cultural events. In *International Conference on Statistical Language and Speech Processing (SLSP)*, pages 82–93. Springer, 2017.

M. Quillot, L. Guillou, A. Gresse, R. Ferro, R. Roth, D. Malinas, , R. Dufour, A. Roebel, N. Obin, J.-F. Bonastre, and E. Ethis. La voix actée : pratiques, enjeux, et applications. In *Journées d'Étude sur le Parole (JEP)*, 2020.

D. A. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009.

A. Rousseau, G. Boulianne, P. Deléglise, Y. Estève, V. Gupta, and S. Meignier. Lium and crim asr system combination for the repere evaluation campaign. In *International Conference on Text, Speech and Dialogue (TSD)*, pages 441–448. Springer, 2014.

M. Rouvier and B. Favre. Sensei-lif at semeval-2016 task 4 : Polarity embedding fusion for robust sentiment analysis. In *International Workshop on Semantic Evaluation (SemEval)*, pages 202–208, 2016.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088) :533–536, 1986.

J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6) :1161, 1980.

B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2009.

B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge : Social signals, conflict, emotion, autism. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2013.

B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, et al. The interspeech 2020 computational paralinguistics challenge : Elderly emotion, breathing & masks. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2020.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11) :2673–2681, 1997.

G. Senay, B. Bigot, R. Dufour, G. Linarès, and C. Fredouille. Person name spotting by combining acoustic matching and lda topic models. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1584–1588. ISCA, 2013.

M. Seok, H.-J. Song, C.-Y. Park, J.-D. Kim, and Y. Kin. Comparison of ner performance using word embeddings. In *International conference on artificial intelligence and application*, pages 754–88, 2015.

M. Seok, H.-J. Song, C.-Y. Park, J.-D. Kim, and Y.-s. Kim. Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and Its Applications*, 10(2) :93–104, 2016.

I. Sheikh, I. Illina, D. Fohr, and G. Linares. Oov proper name retrieval using topic and lexical context models. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5291–5295. IEEE, 2015.

S. Siddiqui, T. Singh, et al. Social media its impact with positive and negative aspects. *International Journal of Computer Applications Technology and Research*, 5(2) :71–75, 2016.

S. R. Singh, H. A. Murthy, and T. A. Gonsalves. Feature selection for text classification based on gini coefficient of inequality. *Workshop on Feature Selection in Data Mining (FSDM)*, 10 : 76–85, 2010.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors : Robust dnn embeddings for speaker recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

E. Spertus. Smokey : Automatic recognition of hostile messages. In *National Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence*, pages 1058–1065. AAAI, 1997. URL `http://dl.acm.org/citation.cfm?id=1867616`.

B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 841–842, 2010.

L. J. Strahilevitz. A social networks theory of privacy. *The University of Chicago Law Review*, pages 919–988, 2005.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted ? large scale analytics on factors impacting retweet in twitter network. In *IEEE Second International Conference on Social Computing*, pages 177–184. IEEE, 2010.

D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou. Coooolll : A deep learning system for twitter sentiment classification. In *International Workshop on Semantic Evaluation (SemEval)*, pages 208–212, 2014.

R. Troncy. Etude du manuel d'indexation commun à tous les documentalistes. *Rapport de recherche, Institut National de l'Audiovisuel*, 2001.

V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5) :988–999, 1999.

P. C. Woodland, S. E. Johnson, P. Jourlin, and K. S. Jones. Effects of out of vocabulary words in spoken document retrieval. In *ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 372–374, 2000.

E. Wulczyn, N. Thain, and L. Dixon. Ex machina : Personal attacks seen at scale. In *International Conference on World Wide Web (WWW)*, pages 1391–1399, 2017.

D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2 :1–7, 2009.

T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in twitter. *Workshop on Computational Social Science and the Wisdom of Crowds*, 104(45) : 17599–601, 2010.

AVIGNON UNIVERSITÉ

LABORATOIRE INFORMATIQUE D'AVIGNON

Titre : Traitement Automatique du Langage : Études et apports aux frontières de l'interdisciplinarité

Mot clés : Traitement du langage et de la parole, Classification automatique, Evaluation, Réseaux sociaux

Résumé : Le traitement automatique du langage naturel (TALN) est un vaste domaine de recherche intégrant de nombreuses thématiques scientifiques (reconnaissance automatique de la parole, indexation automatique de documents, traduction automatique, synthèse vocale, etc.). Ce manuscrit propose un panorama des différents travaux de recherche auxquels j'ai pu participer ces dernières années, mettant alors en perspective l'évolution de mes travaux, qui m'ont conduit à travailler en collaboration avec d'autres disciplines scientifiques pour l'avancée du domaine du TALN. La première partie du manuscrit est consacrée à une des problématiques historiques, à savoir la représentation du contenu écrit et parlé. Nous voyons ensuite, dans la deuxième partie, certains des travaux que nous avons menés sur la performance et l'évaluation en traitement du langage, allant de l'analyse et caractérisation des erreurs de reconnaissance automatique de la parole, à leur correction. La troisième partie montre l'évolution de mes activités de recherche, qui se sont alors orientées vers des problématiques interdisciplinaires pour le traitement du langage, avec nos travaux sur l'exploration des réseaux sociaux pour l'analyse d'événements, la détection de messages abusifs, et enfin le doublage vocal et la recommandation de voix. Ces derniers travaux ont notamment permis des collaborations avec des chercheurs en sociologie des publics, ainsi qu'en réseaux complexes.

Title: Natual Language Processing: Studies and contributions at the frontiers of interdisciplinarity

Keywords: Speech and language processing, Automatic classification, Evaluation, Social networks

Abstract: Natural language processing (NLP) is a vast field of research integrating many scientific themes (automatic speech recognition, automatic document indexing, machine translation, speech synthesis, etc.). This manuscript offers an overview of the various research works in which I have been able to participate in recent years, putting into perspective the evolution of my work, which has led me to work in collaboration with other scientific disciplines for the advancement of the NLP domain. The first part of the manuscript is devoted to one of the historical issues, namely the representation of written and spoken content. We then see, in the second part, some of the works we have carried out on performance and evaluation in language processing, ranging from the analysis and characterization of automatic speech recognition errors, to their correction. The third part shows the evolution of my research activities, which then turned towards interdisciplinary issues for language processing, with our work on the exploration of social networks for the analysis of events, the detection of abusive messages, and finally voice dubbing and voice recommendation. This last work has notably enabled collaborations with researchers in sociology, as well as in complex networks.