

AIX-MARSEILLE UNIVERSITÉ
DIRECTION GÉNÉRALE DE L'ARMEMENT
ED 184 MATHÉMATIQUES ET INFORMATIQUE
LABORATOIRE D'INFORMATIQUE ET SYSTÈMES – UMR 7020

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Informatique

Sébastien DELECRAZ

**Approches jointes texte/image pour
la compréhension multimodale de documents**

Soutenue le 10/12/2018 devant le jury composé de :

M.	Yannick ESTÈVE	Professeur des universités	Le Mans Université	Président
M.	Guillaume GRAVIER	Directeur de recherche	CNRS, IRISA	Rapporteur
M.	Benoît SAGOT	Chargé de recherche, HDR	INRIA Paris	Rapporteur
Mme	Lauriane AUFRANT	Ingénieure de l'armement	DGA	Examinatrice
Mme	Meriem BENDRIS	Ingénieure R&D	Dassault Systèmes	Examinatrice
M.	Alexis NASR	Professeur des universités	Aix-Marseille Université	Examinateur
M.	Frédéric BÉCHET	Professeur des universités	Aix-Marseille Université	Directeur
M.	Benoît FAVRE	Maître de conférence	Aix-Marseille Université	Co-directeur



Cette oeuvre est mise à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International](#).

Numéro national de thèse/suffixe local : 2018AIXM0510/019ED184

Remerciements

J'adresse mes premiers remerciements à ceux qui ont assuré la direction de ces travaux par leurs conseils, leur bienveillance et leur patience : Frédéric Béchet et Benoît Favre.

Je voudrais également remercier les rapporteurs de ce document, Guillaume Gravier et Benoît Sagot, ainsi que les autres membres du jury, Yannick Estève, Meriem Bendris, Lauriane Aufrant et Alexis Nasr pour l'intérêt porté à mes travaux.

Je remercie bien évidemment l'ensemble de l'équipe TALEP pour son accueil, avec une pensée particulière pour Alexis ainsi que Leonor Becerra-Bonache pour notre collaboration très enrichissante.

Je remercie également les permanents qui font de ce laboratoire un endroit où il fait bon chercher mais surtout bon vivre, en particulier Arnaud, Basile, Benjamin, Bertrand, Carlos, Jean-Marc et tous les autres.

Il y a aussi tous ceux avec qui j'ai partagé la vie de doctorant. Ceux qui m'ont transmis le flambeau : je pense notamment à Antoine, les Jerems, Makki, Olivier. Ceux qui ont commencé avec moi : Célia, Didier, Éloi, José Luis, Robin, Thibault. Et puis tous ceux qui sont arrivés après : Damien, Florian, Manon avec une mention spéciale à mes colocs de bureau : Cédric, Jeremy et Sébastien.

Une pensée toute particulière à Adrien, Christophe, Cyriac, Quentin, Sorin et Yoann sans qui ces années d'études n'auraient pas eues la même saveur.

Je souhaite aussi remercier chaleureusement ma famille et mes amis pour leur soutien, leurs encouragements, leur relecture et leur amour. Je pense à mes parents et ma soeur, à mes grand-parents, à ceux qui sont aujourd'hui ma belle famille, et aussi à Maxence, Seb, Max, Edith, Antho et Benoît.

Mes derniers remerciements vont naturellement à celle qui est devenue ma femme au cours de cette dernière année de thèse. Je te remercie pour tout le soutien que tu m'a apporté, ton amour et surtout ta bonne humeur.

Résumé

Les mécanismes de compréhension chez l'être humain sont par essence multimodaux : comprendre le monde qui l'entoure revient chez l'être humain à fusionner l'information issue de l'ensemble de ses récepteurs sensoriels. La plupart des documents utilisés en traitement automatique de l'information sont multimodaux. Par exemple du texte et des images dans des documents textuels ou des images et du son dans des documents vidéo. Cependant les traitements qui leurs sont appliqués sont le plus souvent monomodaux. Le but de cette thèse est de proposer des traitements joints s'appliquant principalement au texte et à l'image pour le traitement de documents multimodaux à travers deux études : l'une portant sur la fusion multimodale pour la reconnaissance du rôle du locuteur dans des émissions télévisuelles, l'autre portant sur la complémentarité des modalités pour une tâche d'analyse linguistique sur des corpus d'images avec légendes.

Pour la première étude nous nous intéressons à l'analyse de documents audiovisuels provenant de chaînes d'information télévisuelle. Dans ces documents l'information contenue dans le signal audio et celle visible à l'écran (image et incrustations texte) ne sont pas nécessairement synchronisées. En effet le locuteur n'est pas toujours visible à l'écran et inversement, la personne visible n'est pas forcément en train de parler. Cette non-synchronie des flux audio et vidéo du point de vue de la personne oblige à mettre en œuvre des traitements particuliers visant à « comprendre » le déroulement de l'émission si l'on veut identifier le rôle de chaque intervenant. Dans ce cadre nous nous intéressons à l'utilisation de méthodes permettant de réaliser une compréhension multimodale de documents vidéo pour la tâche d'identification du rôle des locuteurs. Nous proposons une approche utilisant des réseaux de neurones profonds pour la création d'une représentation jointe multimodale pour les représentations et la fusion des modalités.

Dans la seconde partie de cette thèse nous nous intéressons aux approches permettant d'utiliser plusieurs sources d'informations multimodales pour une tâche monomodale de traitement automatique du langage afin d'étudier leur complémentarité. Dans le domaine de l'analyse syntaxique, les rattachements prépositionnels représentent une source d'erreur majeure. Les indices présents dans la phrase ne permettent pas toujours de lever les ambiguïtés de rattachement alors que des indices visuels pourraient y contribuer. Nous étudions ainsi l'impact de l'intégration de tels indices dans un système d'analyse syntaxique. Nous proposons un système complet de correction de rattachements prépositionnels utilisant de l'information visuelle, entraîné sur un corpus multimodal d'images avec légendes.

Mots clés : Compréhension multimodale, Apprentissage automatique multimodal, Réseaux de neurones profonds, Traitement automatique des langues, Traitement automatique de l'image

Abstract

The human faculties of understanding are essentially multimodal: to understand the world around them, human beings fuse the information coming from all of their sensory receptors. Most of the documents used in automatic information processing contain multimodal information, for example text and image in textual documents or image and sound in video documents, however the processings used are most often monomodal. The aim of this thesis is to propose joint processes applying mainly to text and image for the processing of multimodal documents through two studies: one on multimodal fusion for the speaker role recognition in television broadcasts, the other on the complementarity of modalities for a task of linguistic analysis on corpora of images with captions.

In the first part of this study, we interested in audiovisual documents analysis from news television channels. In these documents, the information contained in the audio signal and that visible on the screen (picture and text inlays) are not necessarily synchronized. Indeed, the speaker is not always visible on the screen and conversely, the visible person is not necessarily speaking. This non-synchronization of the audio and video streams from the point of view of the person requires the implementation of special processings to “understand” the course of the show if we want to identify the role of each speaker. In this context, we are interested in the use of methods allowing to achieve a multimodal understanding of video documents for the task of identifying the role of the speakers. We propose an approach that uses in particular deep neural networks for representation and fusion of modalities.

In the second part of this thesis, we are interested in approaches allowing to use several sources of multimodal information for a monomodal task of natural language processing in order to study their complementarity. In the field of parsing, prepositional phrases attachments represent a major source of error. The clues present in the sentence do not always make it possible to remove the ambiguities of attachment whereas visual cues could contribute to it. We study the impact of the integration of such clues in a parsing system. We propose a complete system of correction of prepositional attachments using visual information, trained on a multimodal corpus of images with captions.

Keywords: Multimodal understanding, Multimodal machine learning, Deep neural networks, Natural language processing, Image processing

Table des matières

Table des matières	9
Table des figures	13
Liste des tableaux	15
Liste des acronymes	17
Glossaire	19
1 Introduction générale	23
1.1 Introduction	23
1.2 Problématique	25
1.3 Organisation du document	27
2 Compréhension multimodale	29
2.1 Définition	29
2.2 Méthodes neuronales pour la représentation des modalités	31
2.2.1 Représentation de l'image	31
2.2.2 Représentation du texte	36
2.3 Récupération d'informations complémentaires dans les différentes modalités	41
2.3.1 Les différents niveaux de fusion	42
2.3.2 Apprentissage par noyaux multiples	43
2.3.3 Approches neuronales	44
2.4 Conclusion	47
3 Modèles de classification multimodaux	49
3.1 Introduction	49
3.2 La tâche d'identification du rôle du locuteur	50
3.2.1 Prise de décision sur l'ensemble des tours de parole d'un locuteur	52
3.2.2 Prise de décision à chaque tour de parole	53
3.2.3 Utilisation de la modalité image	53
3.3 Approche neuronale multimodale pour la tâche de SRR	54
3.4 Cadre expérimental	57
3.4.1 Corpus REPERE	57

3.4.2	Segmentation du document	58
3.4.3	Apprentissage des <i>embeddings</i>	58
3.4.4	Métriques d'évaluation	61
3.5	Résultats	62
3.5.1	Évaluation des systèmes monomodaux	62
3.5.2	Évaluation de l'approche multimodale	62
3.5.3	Évaluation de la capacité de généralisation	63
3.6	Conclusion	65
4	Alignement dans les corpus texte/image	67
4.1	Introduction	67
4.2	Corpus multimodaux texte/image	68
4.2.1	Flickr8k et Flickr30k	68
4.2.2	Microsoft Common Objects in Context	69
4.2.3	Flickr30k Entities	69
4.3	Alignement automatique entre les groupes nominaux et l'image	71
4.3.1	Reconnaissance des objets de l'image	72
4.3.2	Identification des groupes nominaux	75
4.3.3	Alignement des modalités	75
4.4	Cadre expérimental	76
4.4.1	Traitement de l'image	76
4.4.2	Traitement du texte	77
4.4.3	Appariement	77
4.5	Conclusion	78
5	Compréhension multimodale pour le TAL	81
5.1	Introduction	81
5.2	Adaptation d'un corpus multimodal pour la tâche de désambiguïsation	83
5.2.1	Traitement linguistique automatique des descriptions	83
5.2.2	Sélection des rattachements ambigus	84
5.2.3	Annotation manuelle de rattachements prépositionnels	84
5.3	Désambiguïsation de rattachements prépositionnels	86
5.3.1	Bases de connaissances sémantiques	86
5.3.2	Utilisation de corpus	88
5.3.3	Approches multimodales	88
5.4	Stratégie de correction de rattachements prépositionnels	89
5.4.1	Détection des erreurs de rattachement	90
5.4.2	Stratégie de correction	91
5.5	Expériences	93
5.5.1	Résultats	93
5.5.2	Analyse d'erreurs	96
5.6	Conclusion	99
6	Conclusion générale	101

6.1	Conclusion	101
6.2	Perspectives	102
	Bibliographie	105
	Annexes	115
A	Statistiques complètes sur le corpus annoté manuellement	116
B	Performances du système de correction de rattachements prépositionnels	119

Table des figures

2.1	Représentation des 96 premiers filtres du réseau AlexNet	33
2.2	Architecture du CNN AlexNet	35
2.3	Comparaison entre les modèles CBOW et <i>skip-gram</i>	38
2.4	Diagramme d'une cellule LSTM.	40
2.5	Diagramme d'une cellule GRU.	40
2.6	Illustration du processus de récupération d'informations complémentaires dans les différentes modalités	41
2.7	Exemples de légendes générées par le modèle NIC	46
3.1	Images extraites du corpus REPERE mettant en scène les différents rôles utilisés dans la tâche de SRR.	51
3.2	Exemple d'annotation du corpus REPERE	55
3.3	Illustration de l'approche de fusion des espaces de représentations	59
3.4	Illustration de la sélection de l'image pour chaque tour de parole	60
4.1	Exemple de l'annotation du corpus Flickr30k Entities	70
4.2	Illustration d'un document texte/image	72
4.3	Chaîne de traitement automatique pour l'appariement texte/image	73
5.1	Informations linguistiques produites automatiquement par l'outil Macaon	84
5.2	Exemple de l'utilisation de la seconde règle pour la sélection des rattachements ambigus	84
5.3	Interface utilisée pour l'annotation manuelle des rattachements prépositionnels	85
5.4	Exemples d'images pour lesquelles le classifieur a utilisé uniquement les caractéristiques visuelles pour corriger le rattachement de la préposition	97
5.5	Exemples d'images pour lesquelles le classifieur a mal corrigé le rattachement de la préposition avec des caractéristiques visuelles uniquement	98
5.6	Exemples d'images avec un rattachement correct en utilisant des caractéristiques visuelles pour les prépositions alors que l'utilisation de traits lexicaux choisit un mauvais gouverneur	98

Liste des tableaux

2.1	Comparaison de différents CNN	35
2.2	Comparaison des performances de différents systèmes de traduction	45
3.1	Répartition et taille des ensembles du corpus REPERE.	57
3.2	Répartition des tours de parole	58
3.3	Scores de précision et DER des systèmes monomodaux sur l'ensemble de test	63
3.4	Scores de précision et DER pour les fusions tardives a posteriori et au niveau des <i>embeddings</i> suivant différentes combinaisons des modalités audio, texte et image	64
3.5	Performances des modèles neuronaux dans les conditions d'émissions inconnues correspondant à 5 % de l'ensemble de test	64
4.1	Taux d'erreur de notre module d'appariement	78
5.1	Annotation ajoutée au corpus Flickr30k Entities	86
5.2	Statistiques sur les rattachements des prépositions annotées et leur nombre d'occurrences	87
5.3	Taux de rattachements corrects du modèle de correction	94
5.4	Comparaison des taux de rattachement prépositionnel correct sur l'ensemble de test en fonction de l'alignement	96
A.1	Statistiques complètes sur le corpus annoté manuellement	116
B.2	Taux de rattachements correct ssur l'ensemble du corpus de test	119

Liste des acronymes

- ANR** Agence nationale de la recherche [23](#), [57](#)
- ASR** reconnaissance automatique de la parole – ou *automatic speech recognition* – [30](#), [53](#), [62](#)
- AV-ASR** reconnaissance automatique de la parole dans un contexte audiovisuel – ou *audio-visual automatic speech recognition* – [42](#)
- BOW** sac de mots – ou *bag-of-word* – [43](#)
- CBOW** *continuous bag-of-words* [37](#), [38](#)
- CCA** analyse canonique des corrélations – ou *canonical corelation analysis* – [68](#), [71](#)
- CNN** réseaux de neurones convolutifs – ou *convolutional neural network* – [32–36](#), [38](#), [45](#), [54](#), [55](#), [60](#), [63](#), [64](#), [72](#), [74](#), [76](#)
- DER** taux d’erreur de segmentation en locuteur – ou *diarization error rate* – [58](#), [61–63](#), [65](#)
- DGA** Direction générale de l’armement [23](#), [57](#)
- DNN** réseaux de neurones profonds – ou *deep neural network* – [30](#), [31](#), [39](#), [44](#), [47](#), [56](#), [61–64](#), [101](#)
- EHD** descripteur d’histogrammes d’arêtes – ou *edge histogram descriptor* – [32](#), [33](#), [43](#)
- FV** vecteurs de Fisher – ou *Fisher vectors* – [33](#)
- GMM** modèles à mixture de gaussiennes – ou *gaussian mixture model* – [52](#), [53](#), [61](#)
- GPU** *graphics processing unit* [60](#)
- GRU** *gated reccurent unit* [39](#), [40](#), [77](#)
- HMM** modèles de Markov caché – ou *hidden Markov model* – [53](#)
- HOG** histogramme de gradient orienté – ou *histogram of oriented gradients* – [32](#), [43](#), [62](#)
- ILSVRC** ImageNet Large Scale Visual Recognition Challenge [32](#)

- IoU** *intersection over union* 74, 76, voir *intersection over union*
- KCCA** *kernel canonical corelation analysis* 68
- LSA** analyse sémantique latente – ou *latent semantic analysis* – 37
- LSTM** *long short-term memory* 39, 40, 45
- MaxEnt** maximum d'entropie 53
- MFCC** *Mel-frequency cepstral coefficients* 43, 53, 60
- MKL** apprentissage par noyaux multiples – ou *multiple kernel learning* – 43
- MLP** perceptron multicouches – ou *multilayer perceptron* – 33, 34, 41, 58
- MS COCO** Microsoft Common Objects in Context 69
- RBM** machines de Boltzmann restreintes – ou *restricted Boltzmann machines* – 44
- R-CNN** réseaux de neurones convolutifs par région – ou *region-based convolutional neural network* – 74
- ReLU** *rectified linear unit* 34, 55, 56, 61
- RNN** réseaux de neurones récurrents – ou *recurrent neural network* – 38, 39, 45, 65
- SIFT** *scale-invariant feature transform* 32, 43
- SRR** identification du rôle du locuteur – ou *speaker role recognition* – 27, 47, 50–52, 54, 56, 58, 60–62, 67, 78, 101, 102
- SVM** machines à vecteurs de support – ou *support vector machine* – 43, 52, 62, 63, 74
- TAL** traitement automatique des langues 27, 38, 71, 78, 79, 81, 101–103
- UBM** modèle du monde – ou *universal background model* – 60
- WER** taux d'erreur mot – ou *word error rate* – 58

Glossaire

batch Fait référence aux exemples donnés en lot à un modèle d'apprentissage profond ; les poids du réseau sont mis à jour en fonction des erreurs de tous les exemples du lot. [77](#)

cluster En apprentissage, un *cluster* correspond à un ensemble dans lequel des données similaires selon un certain critère ont été regroupées. [37](#), [52](#)

co-occurrence Il s'agit de l'apparition dans un même énoncé de plusieurs éléments linguistiques distincts, par exemple le nombre de fois où deux mots consécutifs sont présents dans un texte. [37](#)

dropout Désactivation de chaque neurone des couches cachées d'un réseau de neurones profonds lors de l'apprentissage avec une certaine probabilité p fixée. [56](#), [59](#)

embedding Dans ce document la notion d'*embedding* correspond au vecteur à valeurs réelles issu du plongement d'un ensemble de caractéristiques d'une modalité dans un espace de représentation. [31](#), [37](#), [38](#), [45](#), [50](#), [54–56](#), [58](#), [60](#), [61](#), [63–65](#), [72](#), [101–103](#)

intersection over union Correspond au rapport entre l'aire de l'intersection et l'aire de l'union de deux surfaces : $\text{IoU} = \frac{A \cap B}{A \cup B}$. Ce score mesure à quel point deux surfaces se ressemblent ; un score de 1 signifiant que les deux surfaces sont identiques. [18](#), [74](#), [76](#)

Approches jointes texte/image pour la compréhension multimodale de documents

Chapitre 1

Introduction générale

Sommaire

1.1	Introduction	23
1.2	Problématique	25
1.3	Organisation du document	27

1.1 Introduction

La majorité des documents analysés chaque jour par les systèmes de traitement automatique de l'information sont multimodaux. Les documents vidéo en sont le parfait exemple puisqu'ils sont la composition d'un signal audio et d'un signal visuel. Mais cette multimodalité se retrouve aussi abondamment dans un contexte texte/image. Aujourd'hui les communications sur Internet ne se limitent plus à de simples échanges de phrases mais sont souvent accompagnées d'une image, d'émoticônes ou encore de liens vers d'autres contenus. Si nous souhaitons que l'analyse de ces documents soit complète, l'ensemble des modalités doit être pris en compte lors de leur traitement automatique.

La multiplication des documents vidéo accessibles sur Internet a rendu nécessaire le développement d'outils permettant d'effectuer des requêtes complexes sur le contenu multimodal de ces documents. Les communautés scientifiques du traitement de l'image et du traitement du signal audio ont chacune proposé de nombreuses façons de caractériser ces documents multimédias, mais chacune dans leurs modalités : identification de locuteurs, transcription de parole pour le signal audio, détection de concepts visuels, classification d'images, reconnaissance d'objets ou de scènes pour la vidéo.

Le défi REPERE¹ (REconnaissance de PERsonnes dans des Emissions audiovisuelles) s'est intéressé à la multimodalité à travers le problème de la détection

1. Campagne d'évaluation de 42 mois lancée début 2011 cofinancé par la [Direction générale de l'armement \(DGA\)](#) et l'[Agence nationale de la recherche \(ANR\)](#) : www.defi-repere.fr

de la présence de personnes dans plusieurs modalités à l'intérieur d'un document vidéo : Qui voit-on ? Qui parle ? Quel nom de personne est écrit à l'écran ? Quel nom de personne est prononcé ?

Les systèmes développés qui ont concouru dans cette évaluation comportent tous à la fois des modules monomodaux tels que des briques d'identification de locuteurs ou de reconnaissance faciale et des modules de fusion multimodale agrégeant les différents descripteurs obtenus (BREDIN, POIGNANT, TAPASWI et al. 2012 ; FAVRE, DAMNATI, BECHET et al. 2013). Les résultats obtenus (GALIBERT et KAHN 2013) lors des différentes phases du défi REPERE ont souligné d'une part la force des briques monomodales, permettant d'obtenir de bons résultats quelle que soit la modalité dans les cas que nous pouvons considérer comme faciles : par exemple lorsque la personne la plus visible à l'écran est aussi celle qui s'exprime oralement ; ou encore lorsqu'un *cartouche* de texte est incrusté dans l'image sous une personne et l'identifie. Le second point soulevé est la nécessité de développer d'autres processus pour les cas plus complexes, par exemple lorsque la personne qui parle n'est pas celle qui est visible à l'écran, et ces processus sont par essence multimodaux.

L'utilisation des différentes modalités semble assez évidente pour des tâches de compréhension de documents audiovisuels. Cependant, la compréhension multimodale peut s'avérer nécessaire pour des tâches qui relèvent d'une seule modalité de prime abord comme l'analyse automatique du langage. Un des grands axes de recherche dans ce domaine est celui de la compréhension du langage, c'est-à-dire la mise en œuvre des procédés automatiques capables de modéliser le sens ou la structure de la langue. Or les langues naturelles sont intrinsèquement ambiguës. Dans le célèbre exemple « Jean regarde l'homme avec un télescope », le seul moyen de savoir à qui de l'homme ou de Jean fait référence le groupe prépositionnel « avec un télescope » est d'avoir un accès visuel de la scène car c'est l'information visuelle de la position du télescope qui est porteuse du sens. Ainsi, dans l'analyse d'un document texte/image, comme une image et sa légende, une compréhension parfaite de celui-ci passe par une analyse des deux modalités. Nous pouvons citer comme application la communication avec des robots, lesquels seraient capables de réagir correctement à des requêtes comme « Attrape cet objet ». Des applications existent en traduction automatique : par exemple entre une image et une légende décrivant le contenu de l'image en une phrase ; entre plusieurs langues avec le support d'une image. Cette dernière application est au centre d'une campagne d'évaluation qui a lieu depuis deux ans lors de la conférence annuelle sur la traduction automatique (WMT) (SPECIA, FRANK, SIMA'AN et al. 2016).

Les modèles capables d'une telle analyse doivent donc mettre en relation les parties de l'image avec les différents groupes nominaux du texte mais aussi savoir transposer la relation entre les mots et les entités de l'image. Ces modèles, lorsqu'ils sont issus du domaine de l'apprentissage automatique se servent de corpus de données afin d'apprendre ce type de relations. Il faudra donc que les données

qui servent à l'apprentissage soient en mesure de couvrir une telle relation à la fois au niveau des mots et au niveau de l'image. Il s'agit d'ailleurs d'une réelle problématique dans le domaine de la multimodalité que celle de pouvoir travailler sur des corpus multimodaux de taille conséquente avec une grande richesse d'annotations. En effet, pendant de nombreuses années, les communautés spécialisées dans le traitement automatique n'ont participé qu'à la création de corpus annotés dans une seule modalité. Le changement ces dernières années avec l'apparition de corpus possédant un grand nombre d'annotations dans chacune des modalités, comme celui du défi REPERE, nous permet de nous intéresser davantage à des méthodes de traitement joint de l'information.

Ainsi, que la tâche d'application d'un système de traitement automatique de document soit multimodale ou non, les difficultés liées à la mise en œuvre de la compréhension multimodale commencent ainsi au plus bas niveau à savoir celui des données, jusqu'à la prise de décision finale.

1.2 Problématique

La compréhension multimodale a pour objectif d'extraire l'information qui est nécessaire à la prise de décision pour une tâche en analysant les différentes modalités à sa disposition.

Le problème de la compréhension de vidéos a déjà été abordé séparément par les communautés scientifiques du traitement automatique de la langue et du traitement d'image. Pour les premiers, des études ont porté sur la segmentation thématique d'émissions vidéo (GUINAUDEAU, GRAVIER et SÉBILLOT 2012) et la reconnaissance du rôle de chaque locuteur dans une émission (DAMNATI et CHARLET 2011). Ces études étaient basées uniquement sur le signal audio et son contenu linguistique à travers les transcriptions automatiques des segments de parole. Pour la communauté du traitement d'image, les études sur l'analyse de scènes (GAIDON, HARCHAOUÏ et SCHMID 2013) permettent d'identifier une action dans une vidéo, uniquement au niveau de l'image, afin de déterminer par exemple : qui parle, qui est assis, qui marche, qui ferme une porte, etc. D'autres travaux tirent parti du script d'un film pour superviser l'apprentissage des actions des personnages à partir des images (BOJANOWSKI, F. BACH, LAPTEV et al. 2013). Des travaux combinant des descripteurs audio et vidéo ont proposé des approches multimodales (DUMONT et QUÉNOT 2012; X. WANG, XIE, LU et al. 2012). Ces approches sont cependant limitées à des tâches de segmentation qui utilisent uniquement de la fusion de descripteurs (caractéristiques extraites des différentes modalités) et non pas des modèles joints de compréhension permettant de prendre en compte la multimodalité dès le processus d'analyse.

Cependant cette multimodalité s'avère indispensable dès que la tâche visée dépasse celle de la simple segmentation. Le défi REPERE portant sur la détection de présence de personnes dans des émissions de télévision en est l'illustration. L'analyse détaillée de la présence de personnes dans ce type d'émissions a montré (BENDRIS, CHARLET et CHOLLET 2010) qu'il n'y avait pas nécessairement de recouvrement exact entre le visage à l'écran et la voix. Ainsi, sur ces contenus, cette non-synchronie des flux audio et vidéo (du point de vue de la personne) oblige à mettre en œuvre des traitements particuliers prenant en compte le déroulement de l'émission (débat, interview, reportage, etc.). Au-delà de la fusion tardive de descripteurs, le problème réside essentiellement dans la compréhension de l'émission et de l'analyse de chaque scène pour déterminer quels en sont les acteurs, actifs ou passifs.

Ainsi nous nous intéressons dans la première partie de ce document à l'étude de la problématique générale : quels sont les processus à mettre en œuvre pour la compréhension multimodale de documents vidéo ? En découle plusieurs interrogations :

- les données brutes devant être transformées afin de les rendre exploitables, quelles sont les transformations efficaces pour en réaliser la fusion par la suite ?
- comment résoudre le problème de non-synchronie de l'information dans les différentes modalités ?
- les modalités ont-elles le même impact dans la prise de décision ?
- les approches mises en œuvre sont-elles facilement généralisables ?

L'objectif de la compréhension multimodale dans un document vidéo peut ainsi être défini comme la capacité d'un système à localiser l'information dont il a besoin pour sa tâche d'application parmi l'ensemble des modalités du document.

À l'inverse, pour des tâches de traitement automatique des langues, l'utilisation de la compréhension multimodale dans le but d'améliorer un système, requiert d'aller chercher une information dans une autre modalité que celle du domaine d'application. Le recours à la modalité visuelle nous permet de supprimer l'ambiguïté sur le propriétaire du télescope dans notre précédent exemple. Cependant, cela requiert une analyse préalable, qui conditionnera la façon dont la vision de la scène pourra être utilisée. Ici il s'agit de la position du télescope qui est importante et c'est cette caractéristique qu'il faudra pouvoir transférer dans l'analyse syntaxique de la phrase.

Nous aborderons ainsi comme seconde problématique générale dans ce document : comment la compréhension multimodale peut-elle être utilisée pour améliorer une tâche monomodale ? Là encore cette question en amène d'autres :

- quelles sont les informations visuelles qui peuvent aider la compréhension de la langue ?

- un alignement étant nécessaire pour transférer l'information d'une modalité à l'autre, quel processus utiliser pour résoudre le problème de l'alignement des modalités ?
- quel est l'impact des erreurs lors de cet alignement ?
- comment intégrer une information visuelle pour une tâche de traitement automatique des langues ?

Que ce soit dans le contexte audiovisuel ou celui du traitement automatique du langage, les travaux décrits dans ce document visent à présenter les moyens d'étudier ces problématiques d'utilisation de la compréhension multimodale ainsi que les différents processus qui la composent.

1.3 Organisation du document

Le chapitre 2 de ce document présente une définition générale de la compréhension multimodale. Il décrit notamment un état de l'art des méthodes utilisant des réseaux de neurones profonds pour l'apprentissage de représentation ainsi que les approches existantes pour la récupération d'informations complémentaires dans les différentes modalités.

Le chapitre 3 décrit les premières contributions réalisées dans le cadre de cette thèse. Nous étudions différentes approches neuronales pour la représentation des modalités d'un document vidéo. Nous proposons une méthode pour la création de représentations multimodales en utilisant des modèles neuronaux. Nous utilisons ces représentations pour la tâche d'identification du rôle du locuteur – ou *speaker role recognition* – (SRR) dans le corpus de vidéos journalistiques REPERE.

Nous nous focalisons par la suite sur les données et les annotations qui sont nécessaires à l'apprentissage de modèles capables de réaliser une compréhension multimodale du document. Le chapitre 4 présente différents corpus de documents texte/image. Nous y abordons la problématique de l'alignement entre les modalités et proposons une chaîne de traitement automatique pour sa réalisation.

Dans le chapitre 5 nous nous intéressons à la mise en œuvre de la compréhension multimodale pour résoudre une problématique du domaine de l'analyse de texte. Nous proposons une approche utilisant l'information visuelle d'une image afin de corriger l'analyse syntaxique de sa légende. L'annotation réalisée sur un corpus texte/image est présentée pour son utilisation dans le cadre d'une tâche de traitement automatique des langues (TAL).

Enfin le chapitre 6 conclut ce document en présentant un bilan des travaux réalisés et en proposant de nouvelles pistes de recherche autour de la compréhension multimodale de documents.

Chapitre 2

Compréhension multimodale

Sommaire

2.1	Définition	29
2.2	Méthodes neuronales pour la représentation des modalités	31
2.2.1	Représentation de l'image	31
2.2.2	Représentation du texte	36
2.3	Récupération d'informations complémentaires dans les différentes modalités	41
2.3.1	Les différents niveaux de fusion	42
2.3.2	Apprentissage par noyaux multiples	43
2.3.3	Approches neuronales	44
2.4	Conclusion	47

2.1 Définition

Le cerveau humain fonctionne naturellement de façon multimodale. Par exemple nous savons qu'il utilise à la fois l'ouïe et la vue dans son mécanisme de compréhension de la parole. MCGURK et MACDONALD (1976) ont mis en lumière ce phénomène en montrant comment la fusion d'informations visuelles et auditives contradictoires pouvaient affecter la perception de la parole avec le célèbre « effet McGurk ». Il s'agit ici d'un des objectifs de la compréhension multimodale à savoir l'intégration de plusieurs sources d'information dans un processus d'analyse de données. Elle vient principalement répondre aux besoins suivants :

- la récupération d'informations complémentaires dans les différentes modalités ;
- le renforcement d'un système pour le rendre robuste à la perte d'une modalité au cours du traitement des données ;
- l'amélioration des prédictions dans le cas où une même information est présente dans l'ensemble des modalités.

Pour une tâche donnée, cela signifie que le système est en mesure de trouver et d'utiliser l'information nécessaire à sa prise de décision, indépendamment de la localisation de celle-ci parmi les différentes sources de données à sa disposition. Le système sera ainsi en mesure de gérer à la fois des informations identiques et contradictoires en fonction des modalités, comme lorsque la personne qui apparaît à l'image dans une vidéo n'est pas la même que celle qui parle.

De nombreux domaines ont ainsi recours à la compréhension multimodale :

- les interfaces homme/machine avec notamment la gestion de la communication non verbale qui représente une part importante dans les échanges humains ;
- la communication qui avec l'explosion des réseaux sociaux ne se limite plus à l'échange de message textuel ;
- l'analyse de documents vidéo qui représente un média très complexe à cause des différents plans que peut avoir une même scène ;
- la traduction de modalité comme la génération automatique de légende pour des images ou encore la [reconnaissance automatique de la parole – ou *automatic speech recognition* – \(ASR\)](#) qui traduit un signal audio en texte.

Cette liste est loin d'être exhaustive mais elle montre la nature très différente que peuvent avoir les données en fonction des modalités, allant d'un signal issu de capteurs pour l'audio à une représentation directement en mot pour le langage naturel.

Ainsi plusieurs transformations peuvent être nécessaires à partir de sa représentation au format brut avant qu'une modalité puisse être utilisée dans un système d'apprentissage automatique, en particulier pour le signal audio et l'image. Ces transformations correspondent à l'extraction de caractéristiques des données qui permettront d'apprécier les différences entre deux unités d'une même modalité. Au départ, ces transformations étaient réalisées de manière explicite, grâce à une expertise sur les données à traiter. Il fallait donc faire en sorte d'utiliser les caractéristiques les plus pertinentes pour la tâche d'application. Une des caractéristiques principales des méthodes d'apprentissage utilisées actuellement est de s'affranchir de cette sélection explicite en effectuant de l'apprentissage implicite de représentation. En effet les [réseaux de neurones profonds – ou *deep neural network* – \(DNN\)](#) permettent d'apprendre des caractéristiques adaptées à la tâche et ce directement depuis le signal.

Nous consacrons la section 2.2 à la présentation de différentes méthodes neuronales pour la représentation des deux modalités centrales à ce document à savoir l'image et le texte. Cependant, la transformation des modalités n'est que la première étape de la compréhension multimodale. En effet dans un second temps il est nécessaire de fusionner les caractéristiques extraites afin de pouvoir récupérer l'information complémentaire dans les différentes modalités. Nous présentons différentes approches qui ont été mises en œuvre dans ce but en section 2.3. Nous

verrons aussi le lien direct avec la prise de décision des systèmes pour différentes tâches. La partie annotation quant à elle, sera étudiée dans le chapitre 4.

2.2 Méthodes neuronales pour la représentation des modalités

Les approches récentes basées sur les DNN ont atteint des performances état de l'art pour de nombreuses tâches de traitement automatique des langues et de l'image. L'avantage de ces modèles est qu'ils sont capables d'apprendre simultanément des représentations abstraites des données ainsi que les fonctions de classification ou de régression pour la tâche qu'ils ont à réaliser. Il sera alors possible de se servir du réseau pour obtenir des représentations de données autres que celles qui ont servies à l'apprentissage. Dans la suite de ce document nous appellerons ces représentations *embeddings*. Ces *embeddings* peuvent ainsi être utilisés soit directement, soient ajustés en utilisant un corpus lié à une tâche en particulier.

Cette capacité à produire des représentations exploitables pour d'autres tâches en font des modèles très intéressants dans le domaine de la compréhension multimodale. En effet, la taille des corpus multimodaux annotés dans toutes les modalités est souvent faible, notamment pour les documents vidéo. Ainsi l'apprentissage des réseaux pourra être effectué sur de grands corpus de données génériques non liés à la tâche. Cela permettra d'obtenir des espaces de représentations pour chacune des modalités qui pourront par la suite être ajustés à la tâche en utilisant un corpus multimodal.

Nous présentons dans cette section, différentes méthodes de transformation des données et notamment des méthodes utilisant des DNN. L'étude de ces approches n'est pas un point central de ce document. Nous nous limitons ainsi à définir les éléments importants relatifs aux techniques nécessaires pour la compréhension des approches mises en œuvre dans nos travaux.

2.2.1 Représentation de l'image

Avant la recrudescence des réseaux de neurones, les modèles d'apprentissage se basaient sur différents descripteurs pour le traitement automatique de l'image. Le rôle de ces descripteurs est de mesurer différentes caractéristiques élémentaires de l'image comme la couleur, la forme, la texture, la position ou encore le mouvement dans le cadre de documents vidéo. Parmi ces descripteurs, il en existe plusieurs issus de la recherche dans le domaine de la détection d'objets dans des images et qui modélisent localement les contours d'un objet.

Parmi ces descripteurs nous pouvons citer par exemple le [descripteur d'histogrammes d'arêtes](#) – ou *edge histogram descriptor* – (EHD) qui modélise la distribution spatiale des arêtes des formes de l'image. Ces descripteurs se basent sur cinq orientations d'arêtes :

- arête verticale ;
- arête horizontale ;
- arête à 45 degrés (par rapport à l'horizon) ;
- arête à 135 degrés (par rapport à l'horizon) ;
- arête non dirigée.

Les histogrammes sont là encore calculés localement. Cependant ces descripteurs modélisent uniquement les contours des objets, ils ne permettent donc pas de discriminer des objets de textures différentes si leur forme est identique.

Le [scale-invariant feature transform \(SIFT\)](#) (LOWE 1999) est aussi très utilisé dans le domaine de la détection. Ces descripteurs proposés à l'origine pour résoudre la problématique de détection d'objets ont rapidement servi pour d'autres tâches comme la recherche d'images par le contenu ou encore l'assemblage de photos. Le principe général de ces descripteurs est d'obtenir une représentation de l'image qui soit indépendante du mode de capture de cette dernière. Ainsi dans le cas d'images n'ayant pas la même résolution, le même angle d'observation ou encore la même exposition, si elles représentent le même objet, alors les descripteurs [SIFT](#) seront quand même relativement proches. Ces descripteurs sont obtenus par une analyse locale de l'image dans laquelle des « points-clés » vont être identifiés et un histogramme des orientations locales des contours, calculés pour chacun d'entre eux.

Un autre descripteur largement utilisé dans le domaine de la détection d'objets dans des images et en particulier pour la détection de personnes est l'[histogramme de gradient orienté](#) – ou *histogram of oriented gradients* – (HOG) (DALAL, TRIGGS et SCHMID 2006). Les histogrammes des gradients de l'image sont calculés sur des petits blocs carrés de quelques pixels et permettent de mesurer localement la forme d'un objet.

Bien que ces descripteurs se soient montrés très performants dans différentes applications d'analyse de l'image, les performances des systèmes ont largement été dépassées avec la recrudescence des méthodes d'apprentissage profond et notamment les modèles basés sur des [réseaux de neurones convolutifs](#) – ou *convolutional neural network* – (CNN). À l'origine les CNN sont inspirés des mécanismes visuels dans les organismes vivants (HUBEL et WIESEL 1968). Aujourd'hui, le traitement automatique de l'image est un domaine dans lequel les CNN sont quasi omniprésents. LECUN, BOTTOU, Y. BENGIO et al. (1998) démontrèrent les performances de cette architecture pour la tâche de reconnaissance de caractères manuscrits. Dans le domaine de la reconnaissance d'objets, ces architectures démontrent chaque année depuis 2012 leurs performances lors de l'[ImageNet Large Scale Visual Recognition Challenge \(ILSVRC\)](#) (RUSSAKOVSKY, DENG, SU et al. 2015).

Il s'agit d'une compétition dont une des tâches principales est la classification d'images issues du corpus ImageNet (DENG, DONG, SOCHER et al. 2009) dans mille catégories.

L'architecture AlexNet proposée par KRIZHEVSKY, SUTSKEVER et HINTON (2012) leur a permis d'obtenir la première place de la compétition en 2012 avec un taux d'erreur de 15,3 % ; soit plus de dix points en moins, par rapport à l'approche non neuronale utilisée par les seconds au classement. L'approche utilisée par ces derniers réalisait la moyenne des prédictions de plusieurs classifieurs entraînés sur des *vecteurs de Fisher* – ou *Fisher vectors* – (FV). Ces architectures permettent ainsi d'apprendre des représentations très performantes pour la tâche de classification d'images. La figure 2.1 présente les différents filtres appris dans la première couche de convolution du réseau AlexNet. Nous pouvons constater l'existence d'une analogie entre ces filtres et les filtres de Gabor ou encore les EHD. Les CNN permettent ainsi de s'affranchir de la conception de tels filtres qui sont appris automatiquement par le réseau.

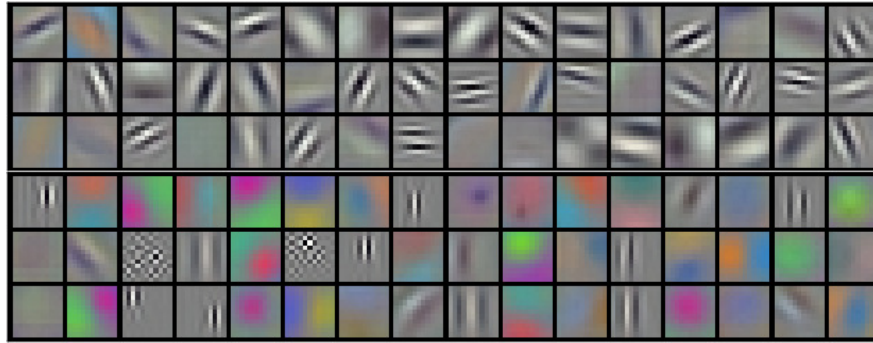


FIGURE 2.1 – Représentation des 96 filtres de taille $11 \times 11 \times 3$ appris par le réseau AlexNet dans sa première couche de convolution (KRIZHEVSKY, SUTSKEVER et HINTON 2012).

Ces architectures sont une évolution du *perceptron multicouches* – ou *multilayer perceptron* – (MLP) dans lesquelles nous trouvons cinq caractéristiques : les couches de convolution, les couches de *pooling*, les couches denses, les fonctions d'activations et la fonction de coût.

Les couches denses correspondent simplement aux couches cachées d'un MLP. Pour rappel, la sortie o d'un neurone d'une couche cachée d'un MLP s'écrit de la forme :

$$y = f\left(\sum_{i=1}^n w_i x_i\right)$$

où un vecteur de poids W est appliqué au vecteur de données X de dimension i puis une fonction d'activation f produit la valeur de sortie.

Les couches de convolution ont pour objectif d'extraire les caractéristiques visuelles de l'image. La différence majeure avec les neurones utilisés dans un [MLP](#) est le nombre de paramètres à apprendre qui est considérablement plus faible. Dans le cas d'un [CNN](#) le vecteur W est remplacé par un noyau K de petite taille. Cela peut se traduire par l'utilisation d'un vecteur de W' dont les poids partagent la même valeur. Dans un [CNN](#), chaque couche de convolution est entraînée à apprendre plusieurs noyaux appelés aussi filtres.

Les couches de *pooling* réalisent une opération de compression de l'information comme un sous-échantillonnage de l'image, réduisant la quantité de calcul nécessaire dans la suite du réseau. Il s'agit d'une opération locale produisant en sortie soit la moyenne des valeurs en entrée (*average pooling*), soit le renvoi de la plus grande des valeurs (*max-pooling*).

Les fonctions d'activations permettent de propager l'information dans le réseau tout en imposant des bornes sur les valeurs de sortie des neurones. Ce sont ces fonctions qui introduisent la non-linéarité dans le modèle. Parmi les plus utilisées nous trouvons :

- la fonction *rectified linear unit* (ReLU) :

$$f(x) = \max(0, x)$$

- la fonction sigmoïd :

$$f(x) = (1 + e^{-x})^{-1}$$

- la fonction tangente hyperbolique :

$$f(x) = \tanh(x)$$

La fonction de coût calcule en sortie de réseau l'erreur d'apprentissage nécessaire à la rétropropagation du gradient sur l'ensemble des poids du modèle. Différentes fonctions sont utilisées en fonction de la tâche d'application du réseau (avec $\hat{y}$ la valeur prédite et y le label) :

- la fonction des moindres carrés :

$$MSE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|^2$$

- la fonction entropie croisée :

$$H_y(\hat{y}) = - \sum_{i=1}^n y_i \log(\hat{y}_i)$$

Noms	# Paramètres	Taux d'erreur
LeNet-5	60 000	—
AlexNet	60 000 000	15,3 %
VGG-16	138 000 000	7,3 %
Inception	23 000 000	6,67 %
ResNet	25 000 000	3,6 %
DenseNet	15 000 000	5,29 %

TABLE 2.1 – Comparaison du nombre de paramètres et du taux d'erreur pour la tâche de classifications du ImageNet Challenge de différents CNN.

— la fonction de coût L_1 :

$$S = \sum_{i=1}^n |y_i - \hat{y}_i|$$

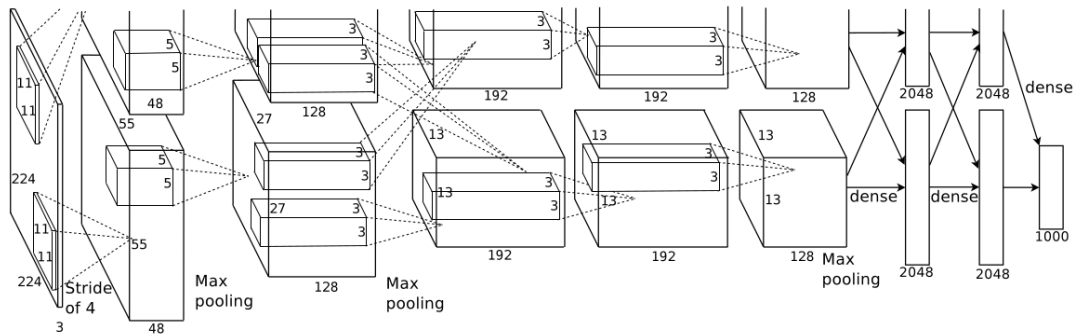


FIGURE 2.2 – Architecture du CNN AlexNet : chaque bloc représente une couche de convolution. Le premier correspond à l'image de taille 224×224 . La largeur d'un bloc correspond au nombre de filtres (48 pour la première convolution) (KRIZHEVSKY, SUTSKEVER et HINTON 2012).

La figure 2.2 représente l'architecture AlexNet. Depuis, de nombreuses architectures de plus en plus complexes voient le jour. Nous présentons dans la table 2.1 le nombre de paramètres et le taux d'erreur pour la tâche de classifications du ImageNet Challenge des six architectures suivantes :

- LeNet-5 : LECUN, BOTTOU, Y. BENGIO et al. (1998) ;
- AlexNet : KRIZHEVSKY, SUTSKEVER et HINTON (2012) ;
- VGG-16 : SIMONYAN et ZISSERMAN (2014) ;
- Inception : SZEGEDY, VANHOUCHE, IOFFE et al. (2016) ;
- ResNet : HE, X. ZHANG, REN et al. (2016) ;
- DenseNet : HUANG, Z. LIU, VAN DER MAATEN et al. (2017).

Aujourd'hui de nombreuses compétitions s'intéressent au traitement automatique de l'image et en particulier à la classification d'images et la reconnaissance

d'objets, la détection d'objets et de personnes, la segmentation d'images, la génération automatique de légendes. Dans chacun de ces domaines, les CNN et les représentations des images qu'ils apprennent sont très largement utilisés et produisent les meilleures performances.

2.2.2 Représentation du texte

Chaque mot d'une langue est associé à un contenu sémantique. L'ensemble des mots d'une langue est appelé un vocabulaire. La façon la plus simple de représenter un mot est la méthode *one-hot*. Il s'agit d'une représentation vectorielle où chaque mot représente une dimension du vecteur. Cette représentation peut aussi être appliquée à un ensemble de mots comme une phrase. Le vecteur prend alors la valeur 1 dans les dimensions des mots de la phrase et la valeur 0 pour le reste. Cependant, derrière chaque mot se trouve un contenu sémantique et morphologique. Avec la représentation *one-hot*, la différence entre deux mots est la même pour n'importe quel couple de mots. Or, la richesse du langage fait que cette différence peut être nuancée de nombreuses façons. Par exemple les mots « grand » et « immense » semblent un peu moins différents que les mots « grand » et « minuscule » d'un point de vue du sens. Ces représentations ne couvrent pas non plus les relations morphologiques des mots comme par exemple les différentes flexions d'un même lemme. Nous présentons dans la suite de cette section des approches pour obtenir des représentations de mots ou de phrases qui permettent de modéliser entre autres ces nuances.

2.2.2.1 Au niveau des mots

Dans les années 50, HARRIS (1954) soumet l'hypothèse que les mots ayant le même contexte ont souvent le même sens. Une représentation des mots doit alors être capable de modéliser les relations sémantiques et syntaxiques entre les mots. Ils peuvent être représentés de façon explicite à l'aide d'un dictionnaire dans lesquels le nombre de mots est fixe. La liste des mots qui constitue un dictionnaire est choisi en fonction de critères linguistiques ou par rapport à une application. À chaque mot est associé une forme qui peut être par exemple orthographique ou encore phonétique ; des informations sont ensuite ajoutées pour chaque forme. Ces informations peuvent être des traits explicites symboliques permettant de représenter soit la nature du mot : sa forme, sa partie de discours (verbe, nom, adjectif, etc.), son lemme, la morphologie, soit des informations fréquentielles.

Concernant la représentation du sens d'un mot, il existe dans la littérature deux approches majeures. La première correspond à l'utilisation de représentations symboliques du sens. La ressource WordNet (MILLER 1995), par exemple, répertorie les mots par catégories morphosyntaxiques et groupes de synonymes (synsets)

ainsi que la relation entre eux (synonymie, antonymie, hyponymie, méronymie, etc.). Une adaptation de cette ressource pour le français est proposée par SAGOT et FIŠER (2008). Ces représentations du sens ont l'avantage d'être très précises au niveau linguistique mais ce sont des ressources statiques, qui d'une part ne s'adaptant pas à un contexte particulier, et d'autre part, proposent des relations qui ne sont pas graduées.

La seconde approche pour la représentation du sens sont les représentations continues. Les premiers modèles allant en ce sens proposent d'apprendre des *clusters* dans lesquels les mots similaires sont regroupés de façon déterministe en suivant des règles ou bien de façon probabiliste (BROWN, DESOUZA, MERCER et al. 1992 ; PEREIRA, TISHBY et LEE 1993 ; BAKER et MCCALLUM 1998). Le principe de ces modèles fréquentiels repose sur l'utilisation d'une matrice de comptes dans laquelle va être répertorié le nombre de fois où chaque mot apparaît dans un ensemble de documents ou dans un contexte. Un schéma de pondération des comptes est appliqué à cette matrice (*Pointwise mutual information*, *Log-Likelihood-Ratios*, etc.). Puis une réduction de dimension de la matrice est réalisée. Chaque ligne dans la matrice réduite représente alors un vecteur de mots. Parmi les méthodes les plus connues nous pouvons citer l'*analyse sémantique latente – ou latent semantic analysis – (LSA)* (DEERWESTER, DUMAIS, FURNAS et al. 1990). Elle consiste à appliquer une décomposition en valeurs singulières de la matrice des *co-occurrences*.

Ces modèles ont été complétés dans les années 2000 par des approches prédictives (Y. BENGIO, DUCHARME, VINCENT et al. 2003 ; SCHWENK 2007 ; COLLOBERT, WESTON, BOTTOU et al. 2011) avec le renouveau des réseaux de neurones et la disponibilité de grand corpus de données. Dans ces représentations, chaque mot est projeté dans un espace multidimensionnel restreint par rapport à la taille du vocabulaire. L'entraînement est réalisé pour que les mots ayant une utilisation proche dans le corpus de données soit aussi proche dans l'espace de représentation. Dans la suite du document nous pourrions faire référence à ce type de représentation sous la dénomination de *word embeddings*.

MIKOLOV, K. CHEN, CORRADO et al. (2013) proposent deux architectures pour apprendre ces deux représentations. Il s'agit de deux modèles symétriques dans leur construction qui utilisent le mot courant et son contexte (c'est-à-dire les mots qui le précèdent et le suivent dans la phrase) afin d'obtenir ces représentations. La figure 2.3 présente ces deux modèles : *continuous bag-of-words (CBOW)* utilise le contexte pour prédire le mot alors que le modèle *skip-gram* prédit le contexte à partir du mot. Plus récemment, PENNINGTON, SOCHER et MANNING (2014) ont proposé le modèle GloVe dans lequel les représentations sont calculées à partir de la matrice de *co-occurrences* des mots du corpus d'apprentissage. Cette matrice est réduite par factorisation afin de produire des représentations de dimension plus petite.

Dans une évaluation comparative entre les représentations issues des modèles fréquentiels et les *word embeddings*, BARONI, DINU et KRUSZEWSKI (2014) montrent que ces dernières représentent le meilleur choix de représentation parmi les modèles sémantiques distributionnels.

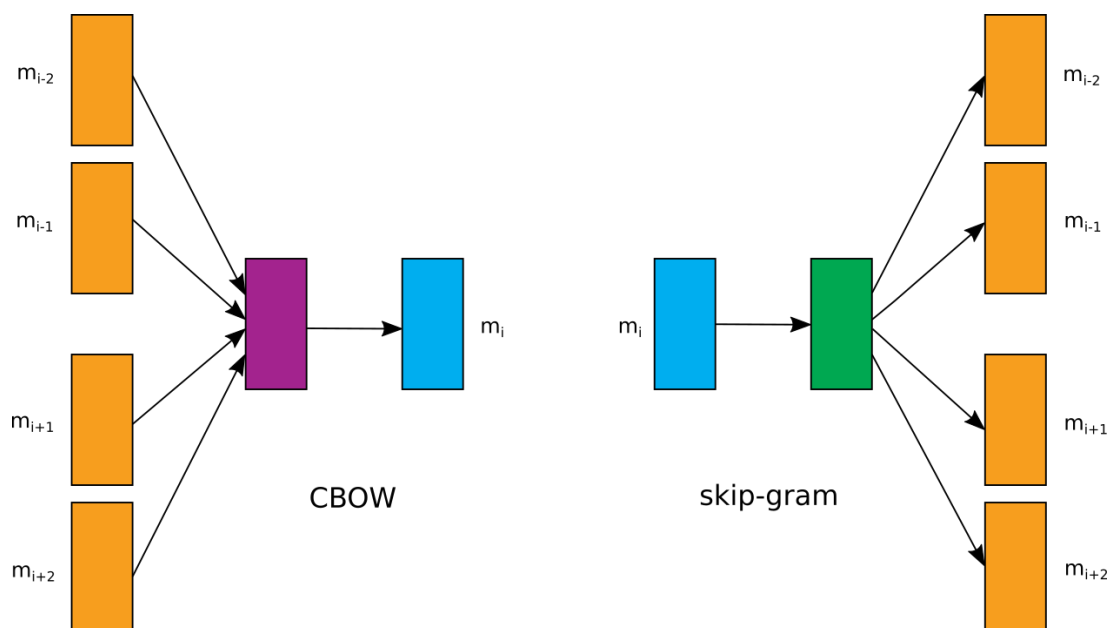


FIGURE 2.3 – Comparaison entre le modèle **CBOW** qui apprend une représentation (en violet) d'un mot m_i à partir des mots de son contexte ; et le modèle *Skip-gram* qui apprend une représentation (en vert) d'un mot, à partir de celui-ci pour en retrouver le contexte.

2.2.2.2 Au niveau d'une séquence

En extension à la représentation des mots, les méthodes neuronales ont largement été utilisées pour la création de représentations de séquences de mots permettant ainsi de capturer l'information de la phrase. COLLOBERT, WESTON, BOTTOU et al. (2011) montrent par exemple comment, appliqué au **TAL**, l'utilisation de **CNN** permet d'extraire les caractéristiques locales au contexte d'un mot pour chaque mot d'une séquence. Pour rappel ; les couches de convolution limitent leur nombre de paramètres en appliquant un filtre de petite taille sur le vecteur d'entrée. Dans le cas d'une phrase, chaque filtre ne traite que quelques mots à la fois. La convolution ne permet ainsi de traiter que le contexte local de chaque mot.

Face à cette problématique, les approches proposées pour des tâches de **TAL** se sont tournées vers l'utilisation de **réseaux de neurones récurrents** – ou *recurrent neural network* – (**RNN**). Ces réseaux ont été introduits formellement par ELMAN

(1990) et se composent d'un vecteur d'entrée x_t , d'une couche cachée h_t , et d'un vecteur de sortie y_t définis par les équations 2.1 et 2.2.

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1} + b_h) \quad (2.1)$$

$$y_t = \sigma_y(W_y h_t + b_y) \quad (2.2)$$

W , U et b étant les paramètres d'apprentissage et σ_y et σ_h des fonctions d'activations. Lors de l'apprentissage, le réseau peut être vu comme un DNN dont les poids sont partagés entre chaque couche. Ces dernières prennent en entrée à la fois l'état caché de la cellule précédente (c'est-à-dire à l'étape $t - 1$) et le mot courant x_t avec $t \in [1, n]$ et n la taille de la séquence. L'état caché de la dernière cellule peut être alors utilisé comme représentation de la séquence.

Cependant, ce formalisme de RNN a de grandes difficultés en pratique à modéliser de longues séquences notamment, à cause du problème de disparition du gradient (Y. BENGIO, SIMARD et FRASCONI 1994; HOCHREITER, Y. BENGIO, FRASCONI et al. 2001). En effet, comme les fonctions d'activations des réseaux sont à valeur dans $[-1, 1]$ ou $[0, 1]$, la rétropropagation du gradient de l'erreur d'apprentissage jusqu'à la première couche du réseau le fait décroître exponentiellement en fonction du nombre de couches. Les cellules *long short-term memory* (LSTM) répondent justement à cette problématique (HOCHREITER et SCHMIDHUBER 1997). Ces cellules disposent d'un état interne c_t , de deux portes F_t (*forget*) et I_t (*input*) lui permettant de contrôler l'information à garder en mémoire en fonction de c_{t-1} , et une porte O_t (*output*) contrôlant si c_t est communiqué en sortie de cellule. Nous détaillons dans les équations 2.3, 2.4, 2.5, 2.6, 2.7 et 2.8 chacune des fonctions de la cellule et donnons un schéma de la cellule en figure 2.4.

$$F_t = \sigma(W_F x_t + U_F h_{t-1} + b_F) \quad (2.3)$$

$$I_t = \sigma(W_I x_t + U_I h_{t-1} + b_I) \quad (2.4)$$

$$O_t = \sigma(W_O x_t + U_O h_{t-1} + b_O) \quad (2.5)$$

$$c_t = F_t \cdot c_{t-1} + I_t \cdot \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (2.6)$$

$$h_t = O_t \cdot \tanh(c_t) \quad (2.7)$$

$$o_t = f(W_o h_t + b_o) \quad (2.8)$$

Le dernier type de cellule récurrente que nous présentons est la cellule *gated reccurent unit* (GRU) (CHO, VAN MERRIËNBOER, GULCEHRE et al. 2014). Il s'agit d'une variation de la cellule LSTM en plus simple. Elle conserve ses capacités de modélisation de mémoire et en facilite l'apprentissage. Les portes F_t et I_t étant par exemple remplacées par une seule porte Z_t (*update*); quant à l'état interne c_t , il est fusionné avec l'état caché h_t . La porte O_t est remplacée par R_t (*reset*). La

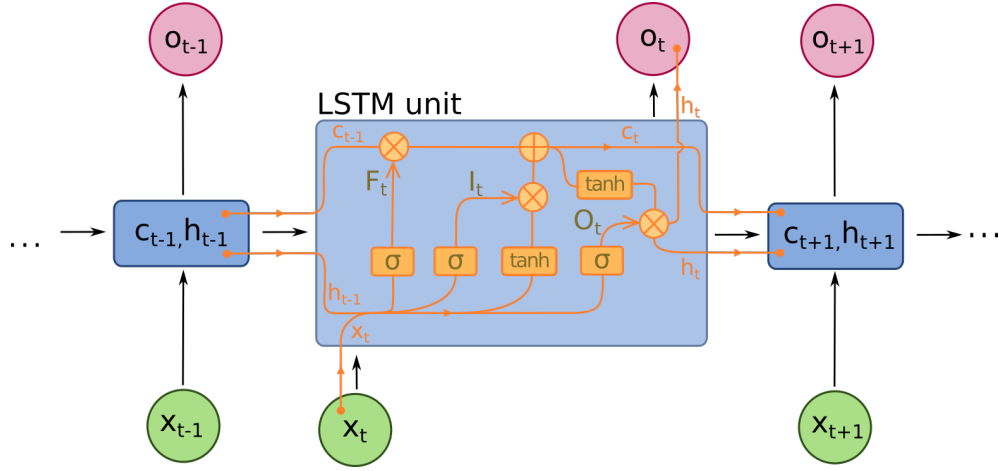


FIGURE 2.4 – Diagramme d’une cellule LSTM (WIKIMEDIA 2018b).

prédiction o_t restant la même. Un schéma de la cellule est dans en figure 2.5 ainsi que les fonctions des portes dans les équations 2.9, 2.10 et 2.11.

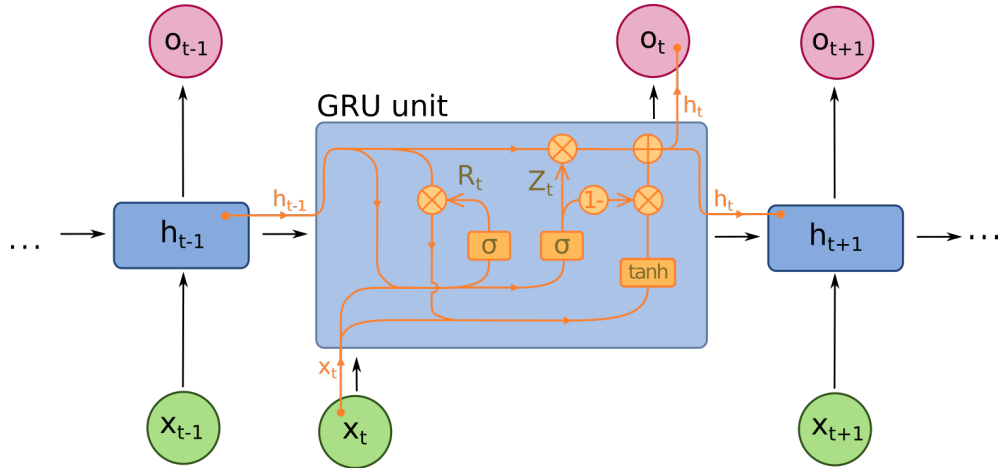


FIGURE 2.5 – Diagramme d’une cellule GRU (WIKIMEDIA 2018a).

$$Z_t = \sigma(W_Z x_t + U_Z h_{t-1} + b_Z) \quad (2.9)$$

$$R_t = \sigma(W_R x_t + U_R h_{t-1} + b_R) \quad (2.10)$$

$$h_t = Z_t \cdot h_{t-1} + (1 - Z_t) \cdot \tanh(W_h x_t + U_h (R_t \circ h_{t-1}) + b_h) \quad (2.11)$$

Cependant, dans ces deux cellules récurrentes, la prédiction y_t ne s’effectue qu’à partir de l’historique courant h_t . BAHDANAU, CHO et Y. BENGIO (2014) proposent alors de remplacer le vecteur de contexte c_t par une somme pondérée

de l'ensemble des états cachés précédents :

$$c_i = \sum_{j=1}^t \alpha_{ij} h_j$$

avec

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^t \exp e_{ik}}$$

$$e_{ij} = v_a^\top \tanh(W_a x_{i-1} + U_a h_j)$$

W_a , U_a et v_a étant les poids du mécanisme d'attention appris avec un [MLP](#) muni d'une seule couche cachée.

2.3 Récupération d'informations complémentaires dans les différentes modalités

Comme l'illustre la figure 2.6, afin de pouvoir prendre une décision à partir des différentes caractéristiques extraites, une des étapes fondamentales de la compréhension multimodale est la fusion des modalités. C'est cette étape qui permet la récupération d'informations complémentaires dans les différentes modalités. Nous consacrons cette section à la présentation de différentes approches pour la mise en œuvre de la fusion des modalités ainsi qu'aux nombreux domaines dans lesquels nous trouvons son utilisation.

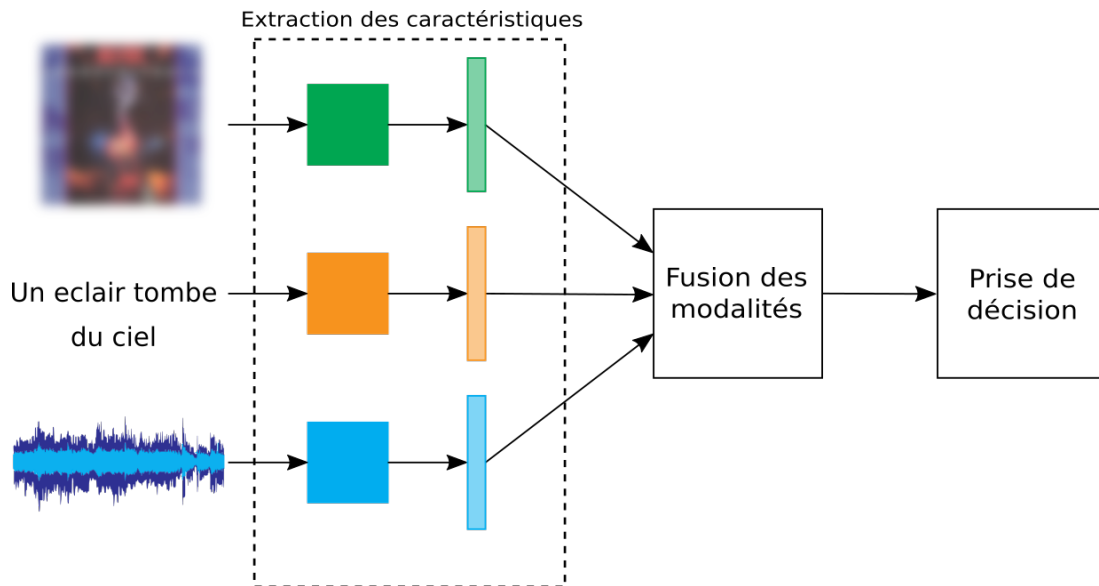


FIGURE 2.6 – Illustration du processus de récupération d'informations complémentaires dans les différentes modalités.

2.3.1 Les différents niveaux de fusion

Dans la littérature, deux grandes familles de méthodes de fusion existent : les approches dites précoces et celles dites tardives. De nombreux travaux existent pour chacun de ces niveaux (ATREY, HOSSAIN, EL SADDIK et al. 2010). Nous présentons ici l'idée générale de ces deux approches.

2.3.1.1 Fusion précoce

Le premier niveau de fusion que nous présentons est celui que nous appellerons fusion précoce. Il correspond à réaliser la fusion directement à partir des caractéristiques extraites des différentes modalités. Cela implique que le système réalise une seule prise de décision en utilisant l'information de toutes les modalités. Cette méthode présente l'avantage de permettre l'observation de corrélations entre les caractéristiques des modalités mais requiert de ces dernières une synchronisation de l'information.

POTAMIANOS, NETI, GRAVIER et al. (2003) proposent une application de la fusion précoce pour la tâche de [reconnaissance automatique de la parole dans un contexte audiovisuel – ou *audio-visual automatic speech recognition* – \(AV-ASR\)](#) dont le but est de traduire le signal audio d'une personne parlante dans une vidéo en langage naturel textuel. Les auteurs réalisent ainsi la fusion des deux modalités après en avoir extrait des caractéristiques pour chacune. Les deux vecteurs contenant respectivement les caractéristiques du signal audio $o_{a,t}$ et du signal vidéo $o_{v,t}$ sont fusionnés par l'intermédiaire d'une concaténation pour obtenir un seul vecteur $o_{av,t}$ contenant l'information des deux modalités :

$$o_{av,t} = o_{a,t} + o_{v,t}$$

Cependant, afin que cette fusion soit possible, les auteurs procèdent à une synchronisation d'un point de vue temporel des deux modalités.

2.3.1.2 Fusion tardive

Par opposition à la fusion précoce, la fusion tardive s'effectue en deux étapes : une prise de décision est réalisée pour chaque modalité puis une décision finale est prise sur l'ensemble des décisions précédentes. Il est ainsi plus simple de fusionner ces premières décisions monomodales car elles ont souvent la même représentation. Il est par exemple fréquent que les hypothèses de prédiction soient associées à une distribution de probabilité. L'avantage d'une telle approche réside aussi dans le fait de pouvoir adapter les modèles de prise de décision aux modalités traitées. Cependant il n'est plus possible d'observer des corrélations entre les caractéristiques des différentes modalités.

La fusion est donc réalisée sur un ensemble de scores issus des classifieurs de chacune des modalités. Plusieurs méthodes existent dans la façon de combiner ces scores. Nous pouvons citer les méthodes de fusion linéaire pondérée qui réalisent la somme ou le produit pondéré des scores (IYENGAR, NOCK et NETI 2003). Dans ce cas, afin de pallier la diversité possible des scores, il est nécessaire d'en réaliser une normalisation (JAIN, NANDAKUMAR et ROSS 2005). Une autre possibilité est d'utiliser ces scores en paramètre d'entrée d'un dernier classifieur (VERLINDE, CHOLLET et ACHEROY 2000).

2.3.2 Apprentissage par noyaux multiples

Les méthodes d'apprentissage par noyaux multiples – ou *multiple kernel learning* – (MKL) (LANCKRIET, CRISTIANINI, BARTLETT et al. 2004), qui sont une extension des machines à vecteurs de support – ou *support vector machine* – (SVM) (BOSER, GUYON et VAPNIK 1992), ont largement été utilisées pour la fusion de modalité (GÖNEN et ALPAYDIN 2011). Cette méthode permet la sélection et la combinaison de noyaux appartenant à un ensemble initial par la pondération de ces derniers. Cette pondération est réalisée dans la fonction noyau générale du SVM qui effectue une combinaison linéaire convexe des noyaux de l'ensemble.

Cette approche est proposée par YEH, T.-C. LIN, CHUNG et al. (2012) pour l'identification d'objets dans des vidéos issues du web. Les auteurs extraient dans un premier temps cinq types de caractéristiques :

- des MFCC pour le signal audio ;
- des descripteurs SIFT ;
- des descripteurs HOG ;
- des filtres de Gabor ;
- des descripteurs EHD.

Les caractéristiques visuelles sont transformées à l'aide d'un modèle de type *sac de mots* – ou *bag-of-word* – (BOW) et les noyaux utilisés dans les MKL sont des noyaux gaussiens.

PORIA, CAMBRIA et GELBUKH (2015) proposent aussi l'utilisation de MKL pour la tâche d'analyse de sentiments de personnes parlantes dans des documents vidéo. Trois modalités sont utilisées dans leur approche : le signal vidéo, le signal audio et le contenu des phrases prononcées dans un format textuel. Les auteurs extraient un certain nombre de caractéristiques (500 pour le texte, 4 568 pour l'image et 6 373 pour l'audio) afin d'obtenir un vecteur de représentation pour chaque modalité. La fusion des modalités est réalisée en concaténant ces vecteurs puis un classifieur MKL est entraîné pour la classification de sentiment à partir de cette représentation multimodale. Les auteurs proposent différentes combinaisons de modalités en entrée du classifieur :

- utilisation d'une seule modalité ;
- utilisation de deux modalités (texte/image, image/audio, audio/texte) ;

— utilisation des trois modalités.

Les performances du classifieur sont directement corrélées avec le nombre de modalités qu'il utilise en entrée. Leurs travaux montrent ainsi que ce type de fusion (c'est-à-dire la concaténation des représentations) donnée en entrée d'un classifieur est une bonne approche pour la compréhension multimodale.

2.3.3 Approches neuronales

L'utilisation des réseaux de neurones appliquée à la fusion multimodale répond, au départ, à la volonté de créer des représentations multimodales capables d'apprendre des corrélations entre les caractéristiques extraites des différentes sources d'informations. Nous trouvons ainsi plusieurs approches mettant en œuvre l'utilisation de *machines de Boltzmann restreintes* – ou *restricted Boltzmann machines* – (RBM).

NGIAM, KHOSLA, KIM et al. (2011) se sont par exemple intéressés à leur utilisation pour l'apprentissage de corrélations cross-modales. Pour ce faire les auteurs proposent une architecture d'auto-encodeur basée sur des RBM. En apprentissage, les auto-encodeurs sont des DNN utilisés pour l'apprentissage de représentations de données. La première partie de ces réseaux projette les données dans un espace de représentation de taille inférieure à celui de l'entrée. La seconde partie quant à elle essaie de reproduire la représentation initiale à partir de l'espace restreint. Dans leurs travaux, les auteurs entraînent donc dans un premier temps une RBM par modalité (audio et vidéo). Une troisième RBM est entraînée sur la concaténation des deux machines précédentes et sert de représentation jointe pour la reconstruction des deux modalités séparées de l'auto-encodeur.

Les RBM aussi ont servi de base aux travaux de SRIVASTAVA et SALAKHUTDINOV (2012) qui proposent de les utiliser en couches pour apprendre des représentations multimodales pour des tâches de classification et de recherche d'informations. Leur modèle transforme les données des différentes modalités dans un espace de représentation joint capable de capturer les caractéristiques nécessaires aux tâches de classification et de recherche.

Comme nous l'avons vu en section 2.2, les réseaux de neurones sont largement utilisés dans chacune des modalités pour l'apprentissage de représentation de données. Ces représentations ont naturellement servi de paramètres d'entrée à des approches neuronales pour la fusion des modalités. NEVEROVA, WOLF, TAYLOR et al. (2016) proposent une architecture en ce sens qui utilise la concaténation des représentations issues de différents réseaux spécialisés dans une modalité pour la reconnaissance de geste. L'originalité de leurs travaux vient de leur module de fusion, qu'ils appellent ModDrop, dans lequel un *dropout* est appliqué à l'ensemble des composants de chacune des modalités avant d'utiliser un DNN pour prendre la décision finale.

Approches	Scores BLEU
BAHDANAU, CHO et Y. BENGIO (2014)	28,45
CHO, VAN MERRIËNBOER, GULCEHRE et al. (2014)	34,54
SUTSKEVER, VINYALS et LE (2014)	34,81

TABLE 2.2 – Comparaison des scores BLEU (PAPINENI, ROUKOS, WARD et al. 2002) de différents systèmes de traduction utilisant des RNN sur le corpus de test (ntst14) de la campagne d'évaluation WMT'14 (BOJAR, BUCK, FEDERMANN et al. 2014).

Nous trouvons aussi l'utilisation de réseaux de neurones dans le contexte multi-modal texte/image avec notamment la problématique de génération de légende d'image. Les RNN ont rapidement montré des résultats état de l'art (comparatif de plusieurs systèmes dans la table 2.2) pour la traduction, en fonctionnant sur le principe encodeur/décodeur pour transformer une phrase d'une langue à une autre.

VINYALS, TOSHEV, S. BENGIO et al. (2015) proposent ainsi de remplacer la partie encodeur par un CNN prenant en entrée une image. Les auteurs entraînent un modèle LSTM à prédire chaque mot de la description à partir de l'image et de l'historique des mots prédits. Le concept qui se trouve derrière ces travaux correspond ainsi d'avantage à une tâche de traduction : l'image étant traduite en langage naturel. Le LSTM propage l'erreur d'apprentissage jusqu'au niveau de l'*embedding* d'image donné en entrée ce qui permet de créer des représentations multimodales. L'espace de représentation des images est adapté par le modèle afin que deux images ayant la même description y soient proches. Des exemples de légendes générées automatiquement par leur modèle Neural Image Caption (NIC) sont présentés en figure 2.7.

Un autre système pour la même problématique a été présenté par XU, BA, KIROS et al. (2015). La différence majeure avec le système précédent est l'ajout d'un mécanisme d'attention. Il est défini formellement par :

$$\hat{z}_t = \phi(\{a_i\}, \{\alpha_i\}) \text{ avec } i \in [1, L] \quad (2.12)$$

où a_i correspond à la caractéristique i parmi les L extraites (chacune correspondant à une certaine partie de l'image), et α_i le vecteur de poids correspondant à la probabilité que la zone i soit pertinente pour prédire le prochain mot. ϕ est une fonction qui crée le vecteur de contexte en appliquant les poids α aux vecteurs a_i , par exemple :

$$\phi(\{a_i\}, \{\alpha_i\}) = \sum_{i=1}^L \alpha_i a_i \quad (2.13)$$

Nous trouvons aussi l'utilisation de la modalité visuelle pour la génération

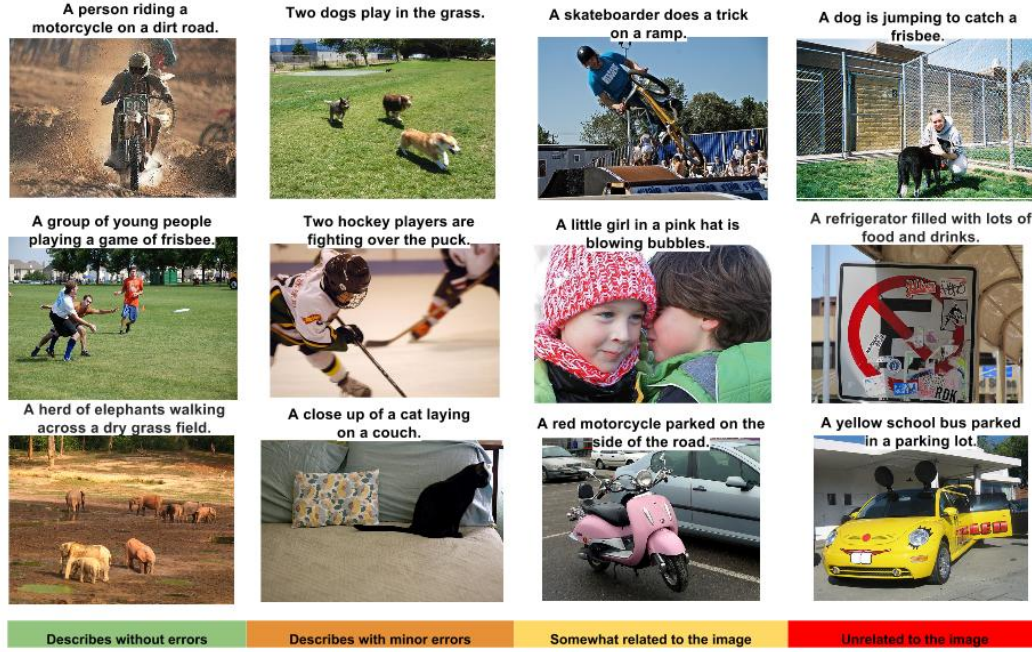


FIGURE 2.7 – Exemples de légendes générées par le modèle NIC (VINYALS, TOSHEV, S. BENGIO et al. 2015).

de texte pour des applications de traduction. CAGLAYAN, ARANSA, Y. WANG et al. (2016) se sont intéressés à la traduction de descriptions d’images dans le cadre de la campagne d’évaluation WMT16 (SPECIA, FRANK, SIMA’AN et al. 2016). Les auteurs proposent un modèle encodeur-décodeur muni d’un mécanisme d’attention. Le mécanisme d’attention partagé (FIRAT, CHO et Y. BENGIO 2016) est utilisé entre les deux modalités. C’est-à-dire qu’il y a un seul modèle qui réalise l’apprentissage des poids de l’attention pour les deux modalités. La fusion est réalisée afin de produire un vecteur de contexte multimodal c_j pour le mot à prédire j :

$$c_j = \tanh\left(\sum_{i=1}^N \alpha_{ij} h_i^{\text{txt}} + \sum_{i=1}^{196} \beta_{ij} h_i^{\text{img}}\right) \quad (2.14)$$

Les termes α_{ij} et β_{ij} sont les poids du mécanisme d’attention entre les modalités sources (texte pour α et image pour β) et le mot cible j . N correspond à la taille du vecteur de contexte de la phrase source (c’est-à-dire le nombre de mots de la phrase source) et 196 à la taille du vecteur de contexte de l’image. Les résultats de ces travaux montrent cependant que l’intégration de caractéristiques visuelles pour le traitement automatique des langues, tel que la traduction, est une tâche difficile. En effet, le système proposé n’arrive pas à produire une traduction meilleure que celles des systèmes qui utilisent uniquement le texte.

2.4 Conclusion

La compréhension multimodale est un axe de recherche dont les domaines d'applications sont nombreux et d'une grande diversité. Nous nous sommes intéressés dans ce chapitre à deux aspects fondamentaux de sa mise en œuvre : la façon d'obtenir des représentations exploitables des différentes modalités puis les méthodes existantes pour les fusionner. Nous avons pu observer que les approches à base de [DNN](#) ont créé un réel changement dans la façon d'obtenir des représentations de données, notamment dans les domaines de l'image et du langage naturel. Nous avons aussi pu constater leurs performances pour la récupération d'informations complémentaires dans différentes modalités ; que ce soit dans un contexte audiovisuel mais aussi pour des documents texte/image.

Nous nous intéressons dans le prochain chapitre à la création de représentations et à la fusion d'informations dans un contexte audiovisuel. Nous proposons une approche neuronale nous permettant d'obtenir une représentation multimodale d'un document vidéo pour la tâche d'[identification du rôle du locuteur – ou *speaker role recognition* – \(SRR\)](#) dans des émissions journalistiques. Nous étudions par la suite les capacités de généralisation d'une telle représentation.

Chapitre 3

Modèles de classification multimodaux

Sommaire

3.1	Introduction	49
3.2	La tâche d'identification du rôle du locuteur	50
3.2.1	Prise de décision sur l'ensemble des tours de parole d'un locuteur	52
3.2.2	Prise de décision à chaque tour de parole	53
3.2.3	Utilisation de la modalité image	53
3.3	Approche neuronale multimodale pour la tâche de SRR	54
3.4	Cadre expérimental	57
3.4.1	Corpus REPERE	57
3.4.2	Segmentation du document	58
3.4.3	Apprentissage des <i>embeddings</i>	58
3.4.4	Métriques d'évaluation	61
3.5	Résultats	62
3.5.1	Évaluation des systèmes monomodaux	62
3.5.2	Évaluation de l'approche multimodale	62
3.5.3	Évaluation de la capacité de généralisation	63
3.6	Conclusion	65

3.1 Introduction

Ce chapitre expose les travaux réalisés à la suite du défi REPERE, une campagne d'évaluation qui s'est intéressée à la multimodalité à travers le problème de la détection de la présence de personnes dans plusieurs modalités, à l'intérieur d'un document vidéo. Nous présentons une approche permettant d'apprendre des représentations multimodales d'un document vidéo. Il s'agit des premières contributions réalisées dans le cadre de cette thèse.

Ce chapitre présente une alternative au paradigme de fusion tardive en proposant une méthode basée sur des *embeddings* multimodaux ajustés pour la tâche d'identification du rôle du locuteur – ou *speaker role recognition* – (SRR). Il s'agit d'une fusion au niveau des espaces de représentation qui caractérise une information multimodale, sans qu'une synchronie entre les modalités ne soit requise. Les expériences que nous avons réalisées sur le corpus français REPERE mettent en évidence le gain de cette approche au regard de stratégies monomodales et des méthodes de fusion tardive.

L'identification du rôle d'une personne dans des émissions de télévision consiste en un problème de classification de personnes (parlant ou visible) selon une liste de rôles (présentateur principal, journaliste, invité, ...). Dans ce contexte, les modalités audio et image sont naturellement complémentaires. En effet nous retrouvons les caractéristiques du rôle dans les trois modalités :

- au niveau du signal audio, par les différences prosodiques notamment ;
- la transcription écrite de la parole, porteuse entre autres du style de langage et du type de discours ;
- enfin, l'image dans laquelle il est possible d'utiliser les caractéristiques des scènes des plateaux d'enregistrement, la mise en forme utilisée pour les séquences de reportage ou d'interview, les tenues vestimentaires des intervenants, leur posture ou encore ce qu'ils sont en train de faire.

De nombreuses approches proposées jusqu'à présent pour l'identification du rôle de la personne ne prennent en compte qu'une seule des modalités. La présence d'une personne n'est pas toujours synchronisée sur les différentes modalités. En effet, les locuteurs ne sont pas toujours visibles et tous les visages visibles à l'écran ne sont pas tous en train de parler. De plus, le manque de données multimodales annotées limite les possibilités pour réaliser un apprentissage joint de systèmes multimodaux qui supposent généralement une parfaite synchronie entre les modalités, en particulier ceux fonctionnant avec de la fusion précoce. Cependant, dans chacune des modalités, des corpus de grande taille existent, comme le corpus d'émission de radio issu de la campagne d'évaluation ESTER (GALLIANO, GEOFFROIS, GRAVIER et al. 2006) ou encore le corpus ImageNet présenté en section 2.2.1.

3.2 La tâche d'identification du rôle du locuteur

L'identification automatique du rôle du locuteur admet que les rôles soient identifiables par des caractéristiques spécifiques acoustiques, visuelles et textuelles, comme le style de langage ou la prosodie. Ici nous voulons classer les locuteurs en quatre rôles en utilisant les modalités audio, image et texte :

- rôle 1 : les présentateurs, caractérisés par leur présence tout au long de l'émission ;

- rôle 2 : les journalistes, des professionnels du monde de la télévision qui apparaissent une fois ou plus au cours d'une émission ;
- rôle 3 : les reporters, proches du rôle R2, ce sont les correspondants qui couvrent les événements en dehors du plateau de l'émission ;
- rôle 4 : les invités et autres, invités pour parler de l'actualité en raison de leur expertise ou de leur renommée, ne prenant pas part à l'organisation et n'étant pas les leaders des débats ; les autres sont toutes les personnes qui peuvent apparaître, comme les personnes interviewées lors d'un reportage.

La figure 3.1 présente les différentes *grammaires* visuelles existant dans le corpus REPERE en fonction des rôles des locuteurs.



(a) Rôle 1 : Présentateur



(b) Rôle 2 : Journaliste



(c) Rôle 3 : Reporter



(d) Rôle 4 : Autre

FIGURE 3.1 – Images extraites du corpus REPERE mettant en scène les différents rôles utilisés dans la tâche de SRR.

Dans la littérature, les méthodes de SRR ont été étudiées dans le but de catégoriser des documents audiovisuels (émissions de débats et d'informations). Les méthodes existantes sont séparées suivant les caractéristiques extraites (audio, texte ou les deux) et la technique de classification (supervisée ou non supervisée).

Enfin une dernière distinction majeure existe entre les différents modèles de [SRR](#) et se situe au niveau de la prise de décision : soit sur l'ensemble des tours de parole d'un locuteur (section [3.2.1](#)) ; soit après chaque tour de parole (section [3.2.2](#)). Ces deux modèles de décision se basent sur les segments audio car les corpus utilisés historiquement dans le domaine étaient issus de la radio.

3.2.1 Prise de décision sur l'ensemble des tours de parole d'un locuteur

Une des approches possibles pour la tâche de reconnaissance du rôle du locuteur est de considérer l'ensemble des segments de parole d'un locuteur avant de lui attribuer son rôle. Les applications utilisant cette approche s'intéressent à l'analyse d'un document ou d'un ensemble de documents. Ainsi plus le nombre de tours de parole pour chaque locuteur est grand, plus il sera facile de prendre une décision sur cet ensemble. DUFOUR, ESTEVE et DELÉGLISE ([2011](#)) proposent une application de la détection de discours spontanés pour la tâche de [SRR](#) en utilisant cette prise de décision. Les auteurs ont mis en évidence une corrélation entre le type de parole et le rôle du locuteur dans des émissions radiophoniques journalistiques. Ils ont ainsi proposé un modèle utilisant des caractéristiques prosodiques et linguistiques extraites pour chaque segment de parole et un modèle contextuel utilisant les segments de parole voisins. Les auteurs utilisent une segmentation manuelle et un découpage en locuteurs de référence pour prendre une décision finale sur l'ensemble des tours de parole d'un locuteur.

HUTCHINSON, B. ZHANG et OSTENDORF ([2010](#)) proposent un système non supervisé de regroupement de locuteurs suivant leur rôle basé sur des caractéristiques structurelles et lexicales. Des algorithmes de sélection *clusters* (programmation linéaire en nombre entier, décomposition en valeur singulière) sont utilisés sur les ensembles de locuteurs, sur des émissions de débats en mandarin et en anglais. W. WANG, YAMAN, PRECODA et al. ([2011](#)) présentent un système de [SRR](#) non supervisé pour l'anglais, l'arabe et le mandarin. Il se base sur l'analyse de caractéristiques structurelles, lexicales, de réseaux sociaux puis d'un classifieur utilisant un algorithme de *boosting*. Une perte de 1,1% de précision est mise en évidence lorsque les caractéristiques sont extraites automatiquement comparé à l'utilisation de phénomènes linguistiques étiquetés manuellement.

BIGOT, FERRANÉ, PINQUIER et al. ([2010](#)) utilisent des caractéristiques temporelles, acoustiques et prosodiques pour classer les rôles au niveau d'un groupe de locuteurs. Les auteurs font la distinction entre les rôles ponctuels et non-ponctuels, et entraînent des [SVM](#) et des [modèles à mixture de gaussiennes – ou *gaussian mixture model* – \(GMM\)](#) (REYNOLDS [2015](#)) comme classifieurs hiérarchiques. De nombreuses méthodes de sélection de caractéristiques ont été comparées (analyse en composantes principales/analyse discriminante/sélection rétrograde séquentielle

de caractéristiques). Des expériences sur un corpus d'émission de radio française ont montré de bons résultats.

3.2.2 Prise de décision à chaque tour de parole

La seconde approche consiste à attribuer un rôle au locuteur à chaque fin de tour de parole. Il s'agit d'une tâche plus difficile car la quantité de données sur laquelle repose la décision est relativement restreinte. Il s'agit cependant d'une tâche plus proche de l'analyse humaine : en quelques secondes, le cerveau humain est capable de comprendre ce qu'il observe et entend sans avoir eu besoin de regarder l'émission depuis le début.

Une approche utilisant des [modèles de Markov caché](#) – ou *hidden Markov model* – ([HMM](#)) ([RABINER 1989](#)) et des classifieurs à [maximum d'entropie](#) ([MaxEnt](#)) est présentée par Y. LIU ([2006](#)). L'analyse du signal audio est réalisée par un modèle de [ASR](#) et la segmentation des tours de parole des locuteurs réalisée manuellement. Dans le modèle à [HMM](#), les rôles des locuteurs sont les états, et les phrases prononcées sont les observations. Le classifieur [MaxEnt](#) utilise la première et la dernière phrase prononcée durant le tour de parole du locuteur pour la classification des rôles. Les meilleurs résultats ont été obtenus par un système basé sur [MaxEnt](#) enrichi avec des informations contextuelles (phrase précédente et suivante). Cependant, la reconnaissance automatique de la parole et la segmentation des tours de parole des locuteurs ont été réalisées manuellement.

DAMNATI et CHARLET ([2011](#)) présentent un système multimodal basé sur des caractéristiques lexicales et acoustiques pour la reconnaissance du rôle. La reconnaissance se fait en deux étapes : d'abord les présentateurs sont déterminés, puis les autres rôles sont affectés. Deux classifieurs sont utilisés : icsiboost (FAVRE, HAKKANI-TÜR et CUENDET [2007](#)) qui est un classifieur basé sur l'algorithme Adaboost (FREUND et R. E. SCHAPIRE [1996](#)), et un [GMM](#) sur les *Mel-frequency cepstral coefficients* ([MFCC](#)). Une régression logistique permet de fusionner leurs sorties.

3.2.3 Utilisation de la modalité image

Pour ce qui est de la modalité visuelle, peu de travaux reposant sur l'image pour la détection du rôle des locuteurs existaient à l'époque où nous nous sommes intéressés à cette problématique. Les recherches portent surtout sur la reconnaissance de type de plan en utilisant des approches à base de réseaux de neurones profonds (FENG, BAI, Z. CHEN et al. [2014](#)).

Dans le cadre du défi REPERE, BECHET, BENDRIS, CHARLET et al. (2015) ont proposé un système de reconnaissance de personnes dans des émissions journalistiques. Il se base sur un modèle d'identification du locuteur enrichi de descripteurs visuels issus de l'analyse des textes incrustés, et d'un module de compréhension de la scène réalisant la détection des caméras et l'identification des plateaux des différentes émissions.

3.3 Approche neuronale multimodale pour la tâche de SRR

Nous présentons dans cette section une approche neuronale multimodale pour la tâche d'identification du rôle du locuteur – ou *speaker role recognition* – (SRR). Le cadre dans lequel nous nous plaçons est celui d'une prise de décision à chaque tour de parole. En effet, ce cadre est le plus intéressant pour la compréhension de la scène sur la modalité visuelle. Il s'agit d'une nouvelle approche par rapport aux travaux présentés dans la section précédente.

Comme nous l'avons présenté dans la section 2.3.1, il existe deux approches utilisées communément pour la fusion multimodale : les fusions précoce et tardive. La fusion tardive considère que les modalités sont indépendantes, en appliquant d'abord des classifieurs séparés pour chaque modalité, puis en fusionnant leurs sorties dans un classifieur de haut niveau. Malheureusement, le classifieur ne peut pas modéliser les corrélations entre les modalités et a seulement accès aux décisions des systèmes monomodaux. L'approche par fusion précoce contourne ce problème en apprenant des caractéristiques et des relations entre les classes pour modéliser les interactions entre les modalités. Toutefois, cette approche nécessite une synchronie entre les caractéristiques extraites dans chaque modalité. Dans la tâche de SRR, cela implique par exemple que la personne visible à l'écran soit forcément celle qui parle. La figure 3.2 présente un aperçu du corpus REPERE et de la diversité de l'information dans les différentes modalités.

L'approche que nous proposons consiste en la création de représentations pour chaque modalité adaptée à la tâche de SRR. Ces représentations sont utilisées en entrée d'un classifieur multimodal qui peut tirer avantage de caractéristiques cross-modales issues de la concaténation des représentations monomodales. Cette approche nous permet ainsi de nous affranchir de la synchronie entre les caractéristiques des modalités, en laissant le réseau apprendre la confiance à accorder à chacune des modalités en fonction du contexte.

Pour la modalité texte, nous entraînons un CNN qui prend des *word embeddings* en entrée. La modalité audio utilise une représentation extraite d'un réseau de neurones modélisé à partir d'un système d'identification du locuteur, mais entraîné sur la tâche de SRR. La modalité image repose sur des représentations extraites



FIGURE 3.2 – Exemple de l’ensemble de l’information disponible dans les différentes modalités d’un document vidéo issu du corpus REPERE.

d’un réseau de neurones d’identification de concepts visuels, puis adaptées pour la tâche cible. La fusion consiste en la concaténation des couches cachées de ces systèmes monomodaux, à laquelle sont ajoutées des couches denses pour créer un espace de représentation multimodal, dans lequel sont fusionnées les caractéristiques monomodales nécessaires à la prise de décision.

Modalité texte Nous avons présenté en section 2.2.2.2 comment les CNN peuvent être utilisés pour apprendre des représentations d’une phrase. Ces représentations se sont montrées très performantes dans le traitement automatique du langage naturel pour des problèmes de classification. Les multiples filtres convolutifs extraient des n-grammes lexicaux de différentes granularités, alors que les couches de *pooling* extraient des caractéristiques sémantiques globales de l’entrée. Par cette approche nous souhaitons ainsi capturer par exemple le style de langage utilisé par les différents locuteurs.

Modalité visuelle Les grammaires visuelles et les émissions de débats et d’informations sont une vraie source d’informations pour l’identification du rôle du locuteur (BECHET, BENDRIS, CHARLET et al. 2015). Les *embeddings* d’image que nous utilisons, sont produits en utilisant le CNN AlexNet (KRIZHEVSKY, SUTSKEVER et HINTON 2012) présenté en section 2.2.1. Cette architecture se compose de cinq couches de convolution, de trois couches denses, des couches de *max-pooling* et de normalisation. Elles utilisent la fonction ReLU comme fonction

d'activation pour accélérer l'apprentissage et du *dropout* après les deux premières couches de neurones denses pour éviter le sur-apprentissage.

Modalité audio Nous utilisons aussi le signal audio dans notre approche. Bien que l'étude de cette modalité ne soit pas centrale dans ce document, la tâche à laquelle nous nous intéressons est relative au locuteur. ROUVIER, BOUSQUET et FAVRE (2015) proposent d'apprendre des caractéristiques acoustiques de haut niveau pour l'identification du locuteur, appelées *speaker embeddings*. Dans la même idée, nous proposons d'apprendre des caractéristiques du rôle du locuteur de haut niveau, appelées *audio embeddings*, en utilisant des modèles profonds entraînés pour la tâche de *SRR*.

Fusion des modalités Nous proposons une approche par fusion précoce, basée sur des réseaux de neurones où les entrées sont les *embeddings* de toutes les modalités pour la tâche spécifiée. Premièrement, les réseaux sont entraînés indépendamment pour chaque modalité pour pouvoir en extraire des représentations monomodales générales (*embeddings* de texte, d'audio et d'image). Deuxièmement, ces *embeddings* sont utilisés en entrée d'un nouveau *DNN* entraîné pour apprendre à classifier le rôle des locuteurs à l'aide de caractéristiques multimodales. Contrairement à la fusion tardive, notre méthode peut tirer avantage de sous-espaces de caractéristiques pertinents de toutes les modalités.

Chaque représentation monomodale est entraînée sur un grand corpus monomodal qui n'est pas nécessairement lié à la tâche de *SRR*. Le corpus multimodal annoté est utilisé seulement lorsque l'on entraîne la fusion. Cette méthode nous permet de tirer parti en même temps des fusions précoces et tardives : avec l'utilisation de grand corpus de données monomodales pour lequel il n'existe pas une synchronie d'annotation dans les autres modalités, comme cela peut être fait en fusion tardive ; ainsi que l'entraînement des classifieurs multimodaux qui permettent la construction de caractéristiques multimodales directement, depuis chaque modalité comme la fusion précoce.

Dans nos expériences, le *DNN* utilisé pour la fusion est composée de deux couches cachées de dimension 1024. La non-linéarité de ces couches cachées est corrigée par une fonction d'activation *ReLU*. Les poids sont mis à jour en utilisant des lots de taille 512, entraînés pendant six époques. Le taux d'apprentissage initialement de 0,01 est réduit à 0,001 lors de la convergence.

Ensembles	Apprentissage	Développement	Test
Tours de parole	18 951	1 402	4 627
Mots	532 131	33 950	118 895
Audio (heures)	41,07	2,72	9,17
Images	18 951	1 402	4 627

TABLE 3.1 – Répartition et taille des ensembles du corpus REPERE.

3.4 Cadre expérimental

3.4.1 Corpus REPERE

L'ensemble des expériences de cette étude a été réalisé sur le corpus multimodal REPERE (GIRAUDEL, CARRÉ, MAPELLI et al. 2012). Ce corpus a été créé dans le cadre du défi REPERE dont le but est le support au développement de systèmes automatiques de reconnaissance multimodal de personnes dans des vidéos. Il s'agit d'une campagne d'évaluation financée par la DGA et l'ANR qui s'est déroulée sur trois ans, de mars 2011 à mars 2014. Ce corpus rassemble 60 heures d'émissions télévisuelles issues de deux chaînes journalistiques françaises : BFMTV et LCP. La figure 3.2 présente un aperçu de l'annotation disponible dans ce corpus.

Chaque tour de parole est manuellement annoté avec la transcription, l'identité des locuteurs, et les rôles des locuteurs selon les classes suivantes : présentateur (R1), journaliste / chroniqueur (R2), reporter terrain (R3) et invité / autre (R4 correspondant à la fusion des rôles R4 et R5 de l'annotation d'origine en raison d'un faible accord inter-annotateur).

Le corpus est divisé en ensembles d'apprentissage, de développement et de test utilisés respectivement pour : l'apprentissage des systèmes ; la validation de la structure des réseaux de neurones et des hyper-paramètres des classifieurs ; et l'évaluation des résultats.

La durée d'un segment moyen est de 7,8 secondes dans le corpus d'apprentissage et les tours de parole ayant une durée inférieure à une seconde non pas été conservés. L'ensemble de test contient à la fois des types d'émissions qui se trouvent aussi dans les ensembles d'apprentissage et de développement (c'est le jour des émissions qui diffère), ainsi que de nouveaux types d'émissions qui sont inconnus des modèles entraînés sur l'ensemble d'apprentissage et de développement ; permettant de vérifier la capacité de généralisation de notre modèle. La répartition des rôles dans l'ensemble de test est la suivante : 23,34 % de présentateurs, 11,28 % de journalistes, 14,22 % de reporters et 51,16 % d'autres.

Rôle	Répartition
Présentateurs	23,34 %
Journalistes	11,28 %
Reporters	14,22 %
Autres	51,16 %

TABLE 3.2 – Répartition des tours de parole par rôle dans l’ensemble de test.

3.4.2 Segmentation du document

La tâche de **SRR** consiste en la classification de segments audios. Nous utilisons ainsi dans nos expériences deux approches pour effectuer cette segmentation. Tout d’abord celle que nous appellerons segmentation *manuelle* et qui correspond à l’utilisation de la segmentation en tour de parole et de la transcription écrite de la parole de référence.

La seconde méthode qui est la segmentation *automatique* utilise pour la segmentation le système LIUM_SpkDiarization (ROUVIER, DUPUY, GAY et al. 2013). Le système produit une segmentation avec un **taux d’erreur de segmentation en locuteur – ou *diarization error rate* – (DER)** de 12,03 % sur ce corpus. La transcription automatique est générée à l’aide du logiciel Kaldi (POVEY, GHOSHAL, BOULIANNE et al. 2011) avec un **taux d’erreur mot – ou *word error rate* – (WER)** sur l’ensemble de test du corpus REPERE est de 19,67 %.

3.4.3 Apprentissage des *embeddings*

La figure 3.3 présente les architectures des différents réseaux. Chaque modalité est entraînée de manière indépendante pour la tâche de **SRR** puis un **MLP** prend une décision multimodale en utilisant les représentations apprises.

3.4.3.1 Pour la modalité texte

Dans nos travaux, nous utilisons la transcription du discours du locuteur courant pour la tâche de **SRR**. Pour chaque mot, nous utilisons des *word embeddings* de dimension 300. Dans nos expériences, ces *word embeddings* ont été entraînés sur Wikipédia avec logiciel word2vec¹ en utilisant le modèle *skip-gram* sur cinq itérations avec une fenêtre de taille sept. Nous souhaitons par cette stratégie caractériser des associations grammaticales et sémantiques entre les mots. Puis, les *word embeddings* pour les mots du tour de parole courant sont passés au travers

1. code.google.com/p/word2vec/

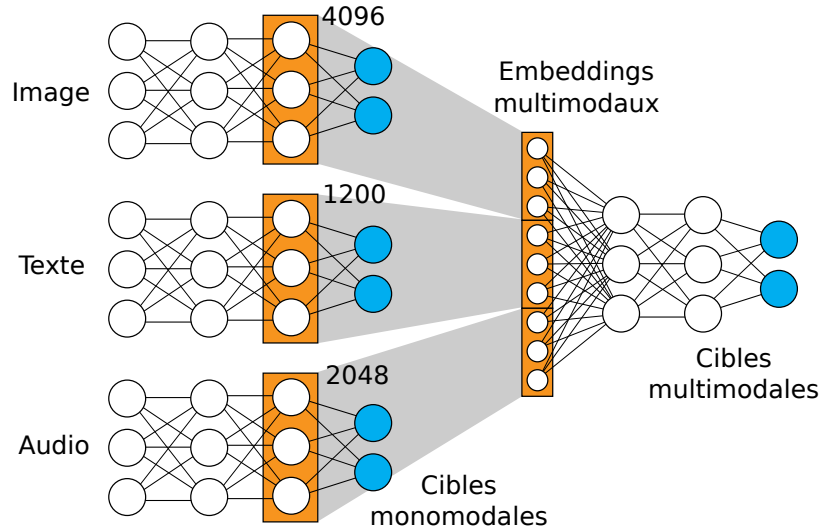


FIGURE 3.3 – Illustration de l’approche de fusion des espaces de représentations : les systèmes monomodaux sont d’abord entraînés de manière indépendante, puis les activations de leurs couches cachées sont concaténées pour servir d’entrée au réseau multimodal.

de trois filtres convolutifs qui sélectionnent les meilleurs 3-grammes, 4-grammes et 5-grammes. Ils sont combinés avec une couche de *max pooling* de dimension quatre cents puis une couche dense avec un *softmax* en sortie. Nous utilisons du *dropout* pour désactiver quarante pour cent des neurones à chaque itération, agissant comme une régularisation.

3.4.3.2 Pour la modalité image

Dans les corpus de type *talk-show* étudiés par BENDRIS, CHARLET et CHOLLET (2010), l’analyse du recouvrement des temps de visages visibles à l’écran et de temps de parole, montrent que le locuteur n’est visible que 60 % du temps, tandis que le visage n’est *parlant* que 30 % de son temps, du fait des plans contenant plusieurs visages, et des différents plans de coupe. Cet asynchronisme représente un réel problème dans le traitement de l’image, car sur un tour de parole de quelques secondes, il est difficile de savoir si un des plans visuels de la séquence contient le locuteur.

Dans notre approche nous avons choisi de n’utiliser qu’une seule image par tour de parole, même si idéalement il faudrait s’assurer de prendre au moins une image pour chaque plan du tour de parole. Ce choix a été réalisé afin de simplifier le traitement de l’image. Comme l’illustre la figure 3.4, pour un tour de parole donné, nous avons donc sélectionné l’image médiane de la plus grande intersection avec la segmentation en plan.

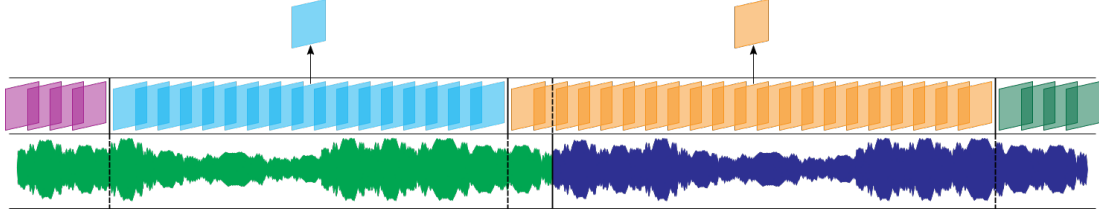


FIGURE 3.4 – Illustration de la sélection de l'image pour chaque tour de parole. Ici deux tours de parole sont représentés avec les différents plans de l'image en parallèle. Les images sélectionnées sont attribuées à l'intégralité du tour de parole.

Pour la création de l'*embedding*, nous avons adapté l'architecture AlexNet (KRIZHEVSKY, SUTSKEVER et HINTON 2012) en changeant la dernière couche de neurones denses pour prédire seulement quatre classes. Nous avons ré-entraîné le CNN pour notre tâche de SRR en initialisant le réseau avec les poids pré-entraînés pour la tâche de classification d'images sur les données ImageNet (DENG, DONG, SOCHER et al. 2009). Le modèle prend en entrée des images carrées redimensionnées de 256 de côté et normalisées en lots de taille 512. Nous avons augmenté le taux d'apprentissage sur les couches de neurones denses (dix fois le taux d'apprentissage global) afin de régulariser l'entraînement. Le réseau a été entraîné pendant 270 époques sur 19 000 images en utilisant le logiciel Caffe (JIA, SHELHAMER, DONAHUE et al. 2014) sur des *graphics processing unit* (GPU). Pour terminer, les *embeddings* d'images sont extraits depuis la seconde couche de neurones denses, ce qui nous fournit des vecteurs de dimension 4 096 qui seront utilisés dans le système de fusion multimodale.

3.4.3.3 Pour la modalité audio

Les *embeddings* audio sont appris de la façon suivante : premièrement, un vecteur de caractéristiques acoustiques de dimension 60 est extrait pour chaque tour de parole avec un taux d'échantillonnage de dix millisecondes (19 MFCC, log énergie et deltas du premier et second ordre). Puis les statistiques du premier ordre, centrées-normalisées, obtenues depuis un *modèle du monde – ou universal background model –* (UBM) sont générées. Ainsi pour chaque composante gaussienne c , la statistique d'ordre du premier rang est extraite de la façon suivante :

$$F^{(c)} = \frac{1}{\sum_t \gamma_c^t} \sum_t \gamma_c^t (o^t - \mu_c) \quad (3.1)$$

où $F^{(c)}$ est la statistique d'ordre du premier rang de composante gaussienne c ; o^t est le vecteur caractéristique d'une trame t ; γ_c^t est la probabilité d'occupation de la gaussienne c pour la trame t et μ_c est la moyenne de la gaussienne c . La statistique de premier ordre complète est $F_i = (F_i^{(1)}, \dots, F_i^{(c)})$. Le modèle du

monde est un **GMM** diagonal à 1024 composantes indépendant du canal, calculé avec le logiciel Kaldi (POVEY, GHOSHAL, BOULIANNE et al. 2011). Ensuite, les statistiques du premier ordre sont utilisées comme entrées d'un **DNN** avec deux couches cachées de 2048 neurones. Les fonctions d'activation de ces couches sont des **ReLU** et la sortie est un *softmax*. L'entraînement est effectué en minimisant l'entropie croisée sur les données d'apprentissage.

L'entraînement est réalisé sur des lots de taille 512, durant huit époques, avec un taux d'apprentissage de 0,04 puis réduit à 0,004 lors de la convergence. Les *embeddings* audio de taille 2048 sont extraits à partir de la dernière couche cachée du **DNN** et utilisés pour la fusion multimodale.

3.4.4 Métriques d'évaluation

La première métrique que nous avons utilisée afin d'évaluer nos approches est l'*accuracy*. Soit S l'ensemble des tours de parole, l'erreur de classification $\text{SRR}_{\text{error}}$ pour un segment $s \in S$ est calculée de la façon suivante :

$$\text{SRR}_{\text{error}}(s) = \begin{cases} 1 & \text{si le rôle prédit est incorrect;} \\ 0 & \text{sinon.} \end{cases}$$

Le score d'*accuracy* se calcule alors de la façon suivante :

$$\text{Acc} = \frac{\sum_{s \in S} \text{SRR}_{\text{error}}(s)}{|S|}$$

Cependant cette métrique ne peut pas être utilisée pour évaluer le système avec une segmentation automatique en locuteur. En effet, un tour de parole généré automatiquement peut se trouver à cheval sur deux tours de parole de la segmentation de référence. Il est alors impossible de savoir si le rôle prédit par le système est correct ou non car il peut y avoir alors deux rôles différents dans la référence.

Pour pallier ce problème, nous utilisons en seconde métrique, le *taux d'erreur de segmentation en locuteur* – ou *diarization error rate* – (**DER**) qui a été introduit par le National Institute of Standards and Technology (NIST) pour des tâches de segmentation en locuteur dans le cadre de la Rich Transcription Meeting Recognition Evaluation. Il s'agit d'une métrique standard dans l'évaluation de systèmes de segmentation en locuteur, correspondant au ratio entre le temps de parole mal analysé par le système et le temps de parole total exprimé en pourcentage. Nous avons choisi d'utiliser la métrique **DER** en l'adaptant à notre tâche en calculant les erreurs de **SRR** au niveau des trames. Soit T l'ensemble des trames d'une vidéo et $\text{dur}(t)$ la durée d'une trame, nous calculons le **DER** de la

façon suivante :

$$\text{DER} = \frac{\sum_{t \in T} \text{dur}(t) \cdot \text{SRR}_{\text{error}}(t)}{\sum_{t \in T} \text{dur}(t)}$$

Cette métrique nous permet ainsi de comparer deux sorties de système de [SRR](#) différentes avec une segmentation en locuteur différente. Dans notre cas, l'utilisation de la segmentation de référence ou alors celle produite automatiquement.

3.5 Résultats

Nous présentons dans cette section les résultats obtenus. Pour en simplifier la lecture, nous définissons la métrique [DER](#)-Man correspondant au [DER](#) calculé sur la segmentation et transcription manuelles, et [DER](#)-Auto correspondant à l'utilisation de la segmentation et de la transcription automatiques.

3.5.1 Évaluation des systèmes monomodaux

Nous avons dans un premier temps comparé, pour chacune des modalités, nos systèmes d'apprentissage profond à des approches non neuronales. Cela permet d'évaluer la qualité des représentations apprises pour la tâche. Parmi les approches de base nous utilisons :

- *Adaboost* : est un classifieur à base de *boosting* (FAVRE, HAKKANI-TÜR et CUENDET 2007) appliqué à des n-grammes de mots de segments textuels pour classifier les rôles ;
- *JFA* : ce système entraîne des modèles *joint factor analysis* (JFA) pour caractériser le rôle du locuteur dans la modalité audio (KENNY, BOULIANNE, OUELLET et al. 2005). JFA est habituellement utilisé pour l'[ASR](#), les données associées à chaque rôle sont considérées ici comme un locuteur ;
- *SVM-HOG* : il s'agit d'un classifieur à base de [SVM](#) entraîné sur des descripteurs [HOG](#) (PANG, YUAN, LI et al. 2011).

Les résultats de ces comparaisons sont présentés dans la table 3.3. Ils indiquent clairement que les approches qui utilisent des [DNN](#) surpassent les systèmes de base pour chacune des modalités. De plus, la modalité audio montre les classifieurs monomodaux les plus performants, passant de 26,76 d'*accuracy* à 77,52.

3.5.2 Évaluation de l'approche multimodale

Nous avons par la suite évalué le modèle de [SRR](#) en utilisant différentes combinaisons de modalités en entrée. Nous donnons à titre de comparaison les scores d'une approche où le rôle le plus fréquent est choisi pour chaque trame de locuteur (système *majorité*). Nous évaluons aussi un système de fusion tardive utilisant

Système	Modalité	Acc-Man	DER-Man	DER-Auto
JFA	Audio	26,76	37,48	42,54
DNN -Audio	Audio	77,52	19,79	25,43
Adaboost	Texte	62,13	28,80	34,33
CNN -Texte	Texte	67,50	29,11	32,66
SVM-HOG	Image	62,76	36,97	42,04
CNN -Image	Image	70,48	25,69	35,25

TABLE 3.3 – Scores de précision et DER des systèmes monomodaux sur l’ensemble de test.

les probabilités données pour chaque modalité par les systèmes monomodaux. Ces probabilités sont regroupées dans un seul vecteur, puis un **SVM** linéaire est entraîné sur ces vecteurs pour prédire le rôle du locuteur.

Les résultats présentés dans la table 3.4 montrent qu’une fusion des décisions au niveau des *embeddings* est plus performante qu’une fusion tardive. Ils justifient aussi l’utilisation de modèles multimodaux pour la tâche : le gain des performances des systèmes multimodaux comparé aux systèmes monomodaux est très important. Le meilleur score de **DER**-Man était de 19,79 % (respectivement 25,43 % pour le **DER**-Auto) pour les systèmes monomodaux alors qu’il est seulement de 13,84 % (respectivement 19,79 %) pour les systèmes multimodaux. Nous pouvons aussi observer que le modèle utilisant les trois modalités produit les meilleurs résultats. Cependant nous n’avons pas vérifié si la différence de performances entre la fusion tardive et avec *embeddings* dans des conditions de segmentation automatique (respectivement 20,97 et 19,79 de **DER**) est significative.

3.5.3 Évaluation de la capacité de généralisation

Afin d’étudier la robustesse de nos méthodes nous avons réalisé une évaluation de notre système sur un sous-ensemble du corpus de test qui correspond à des conditions d’émissions inconnues (c’est-à-dire que le format des émissions est différent). La table 3.5 montre la précision et les **DER** sur ce sous-ensemble.

Le système basé sur des *embeddings* de texte est robuste dans ces conditions alors que les résultats des modalités audio et image diminuent sur ces nouvelles émissions. C’est particulièrement vrai pour la modalité image qui passe d’un **DER** de 25,69 % sur tout le corpus de test, à 43,29 % sur les émissions de format inconnu. Dans ces conditions il n’est pas surprenant que les méthodes de fusion ne donnent pas de meilleurs résultats que la meilleure des modalités seule.

Système	Modalité	Acc-Man	DER-Man	DER-Auto
Majorité	$A + I + T$	51,16	39,77	44,54
Tardive	$A + T$	78,49	18,67	24,11
Tardive	$A + I$	80,98	17,26	22,98
Tardive	$I + T$	78,02	21,16	27,60
Tardive	$A + I + T$	82,36	15,37	20,97
<i>Embedding</i>	$A + T$	80,16	15,90	21,82
<i>Embedding</i>	$A + I$	82,16	15,45	20,65
<i>Embedding</i>	$I + T$	76,01	22,83	28,60
<i>Embedding</i>	$A + I + T$	85,28	13,84	19,79

TABLE 3.4 – Scores de précision et DER pour les fusions tardives a posteriori et au niveau des *embeddings* suivant différentes combinaisons des modalités audio (A), texte (T) et image (I).

Système	Modalité	Acc-Man	DER-Man	DER-Auto
CNN-Texte	Texte	70,07	26,65	28,42
DNN-Audio	Audio	65,69	29,88	37,31
CNN-Image	Image	51,09	43,29	46,47
Tardive	$A + I + T$	70,07	27,77	34,61
<i>Embedding</i>	$A + I + T$	66,42	32,80	34,06

TABLE 3.5 – Performances des modèles neuronaux dans les conditions d’émissions inconnues correspondant à 5 % de l’ensemble de test.

3.6 Conclusion

Dans ce chapitre, nous avons introduit une approche réalisant l'apprentissage de représentations multimodales de documents vidéo. Ces *embeddings* ont été utilisés dans un système d'identification du rôle du locuteur basé sur la fusion d'espaces des représentations multimodaux pour des données asynchrones. Les expériences sur le corpus REPERE en utilisant une segmentation en locuteur et une transcription manuelle ou automatique ont montré que la fusion de caractéristiques textuelles, audio et visuelles améliore considérablement les performances pour la classification en rôles des locuteurs au regard d'approches monomodales. Nos *embeddings* multimodaux ont permis de capturer les caractéristiques du rôle du locuteur sous plusieurs points de vue et l'utilisation d'une fusion au niveau des *embeddings* permet d'obtenir les meilleurs résultats avec 19,79 % de DER sur une segmentation en locuteur automatique.

Notre méthode tire avantage à la fois de la fusion tardive et de la fusion précoce en même temps : nous pouvons utiliser une grande quantité de données monomodales pour lesquelles nous n'avons pas d'annotations synchrones dans les autres modalités comme il se fait en fusion tardive ; nous entraînons des classifieurs multimodaux pour construire des caractéristiques multimodales directement depuis chaque modalité, comme en fusion précoce. Cependant, cette méthode manque de généralisation pour les modèles audio et image dans des conditions de contenus inconnus.

L'augmentation de la robustesse dans ces conditions, ainsi que l'application de notre modèle à d'autres tâches comme l'identification du locuteur, sont des pistes intéressantes pour donner suite aux résultats obtenus. Au niveau de la tâche, une des améliorations serait d'utiliser toutes les images d'un tour de parole. Une architecture de type RNN pourrait être utilisée. Cependant, il sera plus difficile de modéliser les grammaires visuelles propres à ce genre d'émissions, notamment lorsque plusieurs plans apparaissent dans un même tour de parole. En ce qui concerne l'approche, l'utilisation de mécanisme d'attention permettrait d'améliorer la mise en relation des composantes spécifiques des différentes modalités.

Chapitre 4

Alignement dans les corpus texte/image

Sommaire

4.1	Introduction	67
4.2	Corpus multimodaux texte/image	68
4.2.1	Flickr8k et Flickr30k	68
4.2.2	Microsoft Common Objects in Context	69
4.2.3	Flickr30k Entities	69
4.3	Alignement automatique entre les groupes nominaux et l'image . .	71
4.3.1	Reconnaissance des objets de l'image	72
4.3.2	Identification des groupes nominaux	75
4.3.3	Alignement des modalités	75
4.4	Cadre expérimental	76
4.4.1	Traitement de l'image	76
4.4.2	Traitement du texte	77
4.4.3	Appariement	77
4.5	Conclusion	78

4.1 Introduction

Nous avons étudié dans le chapitre précédent la façon dont plusieurs modalités peuvent être fusionnées afin d'obtenir une représentation multimodale d'un document vidéo. Nous avons notamment abordé comment des approches neuronales permettent cette fusion grâce à l'entraînement des réseaux sur des grandes quantités de données puis à leur adaptation pour une tâche de [SRR](#). Cependant, cette adaptation n'est possible qu'avec l'accès à un corpus de vidéos annoté dans l'ensemble des modalités (audio, texte et image). Ce type de corpus est particulièrement rare du fait du coût d'annotation que représente de telles données

alignées selon toutes les modalités annotées. Obtenir automatiquement cet alignement d’annotations mono-modales constitue à la fois une question scientifique intéressante mais aussi une application d’une grande pertinence.

Ce chapitre présente une méthode pour la production d’un alignement entre les différentes modalités d’un document. Nous nous intéressons ici à l’alignement entre les différentes modalités des corpus bimodaux texte/image. Nous présentons dans la section 4.2 plusieurs corpus mettant en relation des images et des phrases les décrivant. Ces corpus plus ou moins riches en termes d’annotations, ne sont souvent pas adaptés pour une utilisation dans une tâche de traitement automatique du langage naturel. La section 4.3 présente une méthode automatique pour l’alignement entre les modalités texte et image. La section 4.4 décrit le cadre expérimental de la mise en œuvre de cette méthode.

4.2 Corpus multimodaux texte/image

4.2.1 Flickr8k et Flickr30k

La construction de corpus d’images associées à des phrases en langage naturel les décrivant est en lien direct avec la volonté d’automatiser cette tâche. La recherche autour de cette tâche a conduit HODOSH, YOUNG et HOCKENMAIER (2013) à proposer en 2013 le corpus Flickr8k. Il se compose de 8 092 images dont chacune est associée à cinq phrases différentes qui fournissent une description des entités et de l’action se déroulant dans la scène. Les phrases ont été produites par des personnes ayant comme tâche de décrire les images et ne sont pas extraites automatiquement d’Internet.

Les auteurs ont eu recours à la création d’un nouveau corpus à cause du manque de ressources dont les descriptions sont génériques. Les images sont issues du site Flickr¹ qui est un service de gestion et de partage de photos en ligne. Les images ont été sélectionnées manuellement dans le but d’obtenir une variété importante de scènes et de situations en évitant les personnalités et les endroits connus. Les descriptions ont été produites en anglais par des personnes américaines ayant réalisé au préalable un test d’orthographe et de grammaire.

Les auteurs proposent d’utiliser leur corpus dans un cadre de description d’images. Ils abordent cette tâche comme le problème de recherche d’une fonction d’affinité capable de mesurer le degré d’association entre des images et des phrases. Ils proposent deux systèmes pour résoudre cette problématique : l’un utilisant une approche par plus proches voisins ; l’autre basé sur une technique appelée *kernel canonical corelation analysis* (KCCA) (F. R. BACH et JORDAN 2002 ; HARDOON, SZEDMAK et SHAW-ET-TAYLOR 2004) qui est une extension de la méthode d’analyse

1. www.flickr.com

canonique des corrélations – ou *canonical correlation analysis* – (CCA) utilisant une fonction à noyau afin de relâcher les contraintes induites par les projections linéaires des données.

Ce corpus a été augmenté à 31 783 images pour 158 915 phrases de descriptions pour des travaux portant sur la dénotation visuelle (YOUNG, LAI, HODOSH et al. 2014). Les auteurs proposent la construction d’un graphe de dénotation visuelle de ce nouveau corpus (appelé Flickr30k) où chaque nœud correspond à une phrase et à un ensemble d’images qui pourraient être décrites par cette phrase. Les arcs dirigés de ce graphe entre le nœud d’une phrase s et celui d’une phrase s' représentent l’existence d’une opération lexicale ou syntaxique qui transforme s en s' .

4.2.2 Microsoft Common Objects in Context

Un autre grand corpus associant des images et leurs descriptions est celui proposé dans le cadre de la tâche de génération de descriptions d’images du challenge *Microsoft Common Objects in Context (MS COCO)* 2015. À l’origine, ce corpus a été créé pour servir la recherche dans le domaine de la détection d’objets dans des images (T.-Y. LIN, MAIRE, BELONGIE et al. 2014). Il s’agit d’un corpus de 328 000 images dans lesquelles ont été détournées manuellement les entités parmi une liste de 91 catégories (par exemple : personne, chien, véhicule, fenêtre, ...). Deux sous-ensembles d’images issus de ce corpus ont été manuellement annotés (X. CHEN, FANG, T.-Y. LIN et al. 2015) :

- un premier ensemble de 159 062 images dans lequel les images ont reçu cinq descriptions ;
- un second ensemble de 5 000 images dans lequel les images ont reçu quarante descriptions.

4.2.3 Flickr30k Entities

PLUMMER, L. WANG, CERVANTES et al. (2017) s’intéressent dans leurs travaux à la compréhension du langage dans un contexte visuel et dans quelle mesure, l’utilisation du langage, peut aider à une meilleure interprétation de l’image. Ils présentent dans ce contexte le corpus Flickr30k Entities qui est une extension du corpus Flickr30k présenté en section 4.2.1.

Il s’agit d’un ajout d’annotations manuelles de 244 035 chaînes de coréférences entre les mentions d’une même entité dans une image et 275 775 boîtes englobantes pour chaque entité. Un exemple de cette annotation est présenté en figure 4.1.

Pour chaque image et les cinq descriptions qui lui sont associées, le corpus nous fournit les informations suivantes :



1. **Someone (people)** is holding out **a punctured ball (other)** in front of **a brown dog (animals)** with **a red collar (clothing)**.
2. **A man (people)** holding out **a deflated soccer ball (other)** in front of **a gray dog (animals)**.
3. **The owner (people)** tries to hand **a deflated ball (other)** to **his dog (animals)**.
4. **Large gray dog (animals)** being handed **a white soccer ball (other)**.
5. **A brown dog (animals)** starring at **a soccer ball (other)**.

FIGURE 4.1 – Exemple d’annotation du corpus Flickr30k Entities : l’image est décrite par cinq phrases, chacune annotée avec les différentes entités ; chaque entité visible dans l’image est reliée à sa boîte englobante correspondante (la coréférence étant représentée par la couleur). Chaque entité est associée à une catégorie sémantique indiquée entre parenthèses.

- les groupes nominaux faisant référence aux différentes entités dans les phrases ;
- une chaîne de coréférence est établie entre tous les groupes nominaux qui font référence à la même entité dans les différentes phrases (représentée par les différentes couleurs) ;
- une classe sémantique est attribuée à chacune des entités parmi la liste suivante : personne, habit, partie du corps, animal, véhicule, instrument, scène et autre ;
- pour chaque entité, si elle est visible dans l'image alors elle est mise en relation avec une ou plusieurs boîtes englobantes qui lui font référence dans l'image.

La création de ce corpus fait écho à la volonté des auteurs de développer des méthodes pour la mise en correspondance entre les groupes nominaux d'une phrase et les régions d'une image. Ils proposent une approche, se basant majoritairement sur les travaux de KLEIN, LEV, SADEH et al. (2014), dans laquelle un espace latent partagé est utilisé pour apprendre des représentations des régions d'image et des caractéristiques des groupes nominaux. La région d'image d'un groupe nominal est ensuite récupérée en utilisant la distance entre les deux représentations (celle de l'image et celle du groupe nominal) dans cet espace. Une CCA est utilisée pour trouver les projections linéaires des deux modalités dans un espace joint de même dimension dans lequel la corrélation entre les vues est maximisée.

4.3 Alignement automatique entre les groupes nominaux et l'image

Les méthodes de TAL font généralement abstraction du contexte multimodal dans lequel est exprimé le langage (vidéo, image avec légendes), nous pouvons alors nous poser la question : que se passe-t-il lorsque cette information visuelle est accessible ? C'est à cette question que nous tentons de répondre dans la suite de ce document en nous intéressant à la problématique de l'alignement entre les modalités textuelles et visuelles.

La figure 4.2 est un exemple de la nécessité d'un tel alignement. Dans l'analyse syntaxique de la phrase « Un homme est debout sur un petit radeau avec un drapeau » nous cherchons à mettre en relation les mots afin d'en comprendre la structure. Pour l'analyse du groupe prépositionnel « avec un drapeau » il y a deux possibilités : soit il s'agit d'un drapeau attaché au radeau, soit c'est l'homme qui le tient. Face à l'image nous sommes capables d'émettre plusieurs observations quant à la position de chaque groupe nominal (à savoir l'homme, le radeau et le drapeau) par exemple la position de chaque objet entre eux. Si maintenant nous souhaitons utiliser ces informations dans notre analyse, il faut être en mesure de

savoir quel objet de l'image correspond à quel mot de la phrase, pour comprendre à quoi s'appliquent nos différentes observations.



FIGURE 4.2 – Illustration de la phrase « Un homme est debout sur un petit radeau avec un drapeau ».

L'utilisation de la modalité visuelle pour la désambiguïsation des rattachements prépositionnels suppose qu'un alignement entre des morceaux de l'image, qui peuvent être représentés par des boîtes, et des groupes nominaux des légendes existe. Nous présentons ici la chaîne de traitement utilisée pour la réalisation de cet appariement automatique qui se décompose en quatre étapes :

- la détection des boîtes dans l'image et l'association de chaque boîte à une classe sémantique ;
- la détection des groupes nominaux dans la légende ;
- la transformation de chaque boîte et chaque groupe nominal en *embedding* ;
- l'appariement des deux modalités afin de produire un ensemble de couples boîte/groupe nominal.

La figure 4.3 illustre cette chaîne de traitement que nous décrivons dans les trois sections suivantes.

4.3.1 Reconnaissance des objets de l'image

La tâche de détection de boîtes dans une image consiste à prédire la présence ou l'absence d'un objet dans une image, étant donné une liste d'objets que le système est en mesure de reconnaître. Lorsqu'un objet est reconnu, les coordonnées de la boîte qui le contient sont produites.

L'utilisation massive des CNN dans le domaine du traitement de l'image a fait émerger de nouvelles méthodes pour cette tâche et notamment l'utilisation de

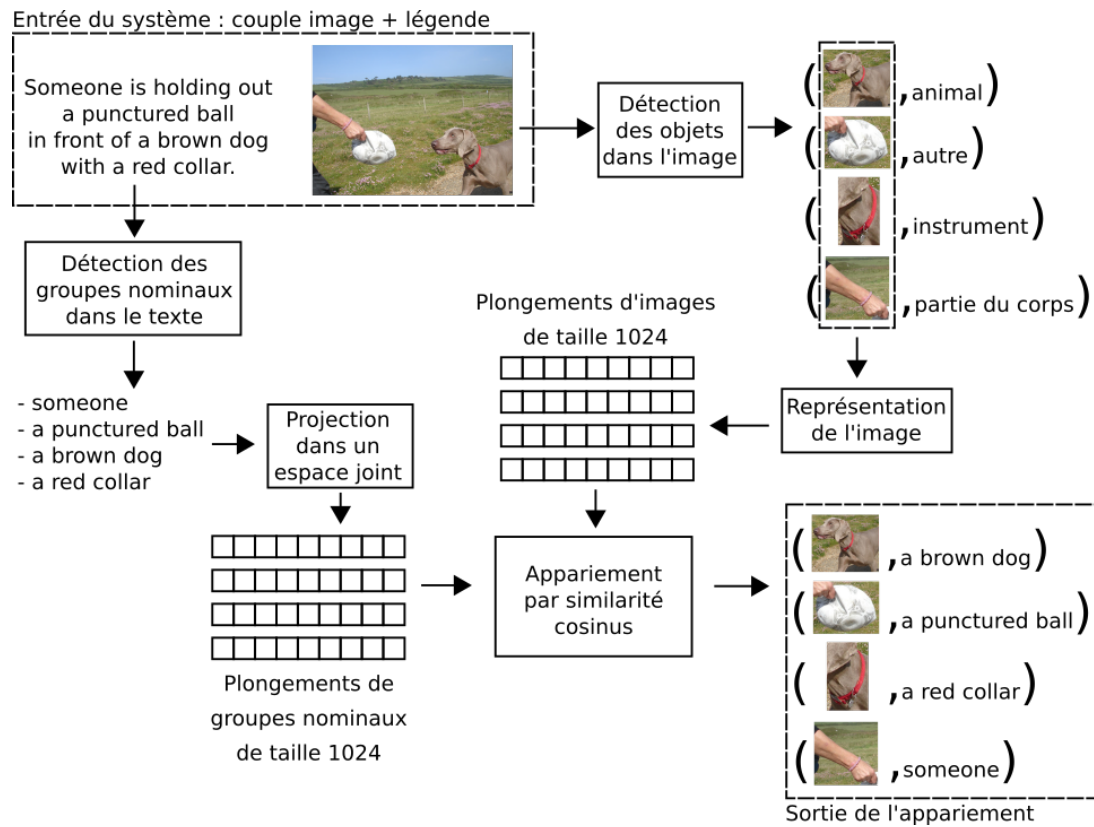


FIGURE 4.3 – Chaîne de traitement automatique pour l'appariement entre les boîtes des entités d'une image et les groupes nominaux de sa description.

réseaux de neurones convolutifs par région – ou *region-based convolutional neural network* – (R-CNN). GIRSHICK, DONAHUE, DARRELL et al. (2014) proposent un modèle qui se décompose de la façon suivante :

- la première étape dans laquelle un ensemble de régions d'intérêts de l'image est sélectionné (UIJLINGS, VAN DE SANDE, GEVERS et al. 2013) ;
- par la suite chaque région est passée en entrée à un CNN pour en récupérer une représentation ;
- un SVM par classe attribue un score à chaque représentation de région ;
- une méthode naïve est utilisée pour éliminer les régions en comparant leur score deux à deux pour celles ayant un score d'*intersection over union* (IoU) supérieur à 0,3.

L'avantage d'un tel modèle est qu'il limite le nombre de régions à traiter par le CNN ce qui très coûteux en termes de temps. Cette problématique du temps de calcul a été étudiée par GIRSHICK, DONAHUE, DARRELL et al. (2014). Ils proposent un nouveau modèle appelé Fast R-CNN dans lequel sont ajoutées de nombreuses optimisations pour améliorer la vitesse de traitement ainsi que les performances de détection. Ces modèles ont évolué dans les travaux de REDMON, DIVVALA, GIRSHICK et al. (2016) qui s'intéressent à la problématique de détection d'objets dans des flux vidéo. Ils y présentent un système dans lequel la détection de région a été supprimée. Une partie réalisant la prédiction des régions a été ajoutée au CNN afin de réaliser les deux tâches en une seule évaluation de l'image.

Le modèle que nous avons utilisé pour réaliser la partie détection d'objets de notre tâche d'alignement est le système YOLOv2 (REDMON et FARHADI 2017) qui est une extension du modèle de REDMON, DIVVALA, GIRSHICK et al. (2016). Ce système se décompose de la façon suivante : il prend en entrée une image puis la découpe en grille. Pour chaque cellule de la grille, le système prédit un nombre fixe de boîtes englobantes et un score de confiance pour chaque boîte. Ce score correspond à la confiance du modèle en la présence d'un objet dans la boîte et à quel point les dimensions de la boîte prédite correspondent à celle d'un objet de référence. Le système admet aussi en entrée une liste d'objets qu'il pourra reconnaître et fournit pour chaque cellule de la grille une distribution de probabilité sur ces classes. La prédiction finale est générée en multipliant les probabilités des différentes classes par le score de confiance. Le modèle produit ainsi un ensemble de couples (boîte, classe) avec un score de confiance pour chaque couple qui représente la probabilité qu'un objet de cette classe soit dans cette boîte et que la taille de la boîte soit parfaitement adaptée à l'objet.

Chaque couple ainsi extrait doit être aligné sur le groupe nominal qui lui correspond dans la légende. La prochaine section décrit la méthode utilisée pour l'identification des différents groupes nominaux de la légende.

4.3.2 Identification des groupes nominaux

Bien que la détection de groupes nominaux soit une tâche largement étudiée, les groupes cibles dans notre travail correspondent à des objets visuels et peuvent différer par nature, des groupes nominaux typiques issus d'une analyse syntaxique. Pour cette raison, le système est entraîné directement sur les groupes nominaux du corpus Flickr30k Entities.

Il consiste en un simple détecteur de début et de fin de groupes nominaux, qui associe à tout mot de la phrase, une étiquette de la forme *B* (*begin*), *I* (*inside*) et *O* (*outside*) selon que le mot débute un groupe nominal, qu'il se trouve à l'intérieur d'un groupe nominal sans en être le premier mot ou qu'il se trouve à l'extérieur d'un groupe nominal. La prédiction est réalisée à l'aide d'un perceptron moyenné qui repose sur les mots de la phrase ainsi que sur leurs parties de discours. Une évaluation sur un ensemble de test indique un taux d'erreur de 2,2 % par mot.

4.3.3 Alignement des modalités

Le problème de l'appariement consiste à déterminer pour chaque groupe nominal à quel objet détecté il correspond dans l'image. Par exemple, pour la légende de la figure 4.3, il faut retrouver parmi les boîtes correspondant aux objets détectés (le ballon, le bras, le chien, le collier) à quels groupes nominaux ils correspondent (*someone, a punctured ball, a brown dog, a red collar*). C'est un problème de vision artificielle difficile du fait de la nature très différente des objets appariés : les pixels de l'image d'un côté et une séquence de caractères de l'autre. Il est rendu encore plus difficile par le fait que certains groupes nominaux peuvent correspondre à plusieurs objets dans l'image (par exemple *children playing soccer*) ; certains objets ne sont représentés que partiellement dans la photo (*people standing in a train*), et le détecteur d'objets peut avoir détecté des objets non représentés dans la légende.

Nous abordons cette tâche comme deux sous-tâches : la première consiste à calculer un score d'association entre chaque objet visuel et chaque groupe nominal ; la seconde est de décider parmi ces associations potentielles lesquelles seront conservées pour la suite. Dans un souci de simplification nous ne traitons pas le problème des associations multiples.

Le score d'association entre un objet visuel et une séquence de mots est calculé en projetant les pixels de l'image et les mots de la légende vers un même espace de représentation. Chaque objet visuel et chaque séquence textuelle sont représentés par des vecteurs dans cet espace commun ; ce qui permet de calculer une similarité entre les vecteurs pour obtenir un score d'association. Les paramètres de ce réseau peuvent être entraînés à partir de paires connues (**image**, **texte**) à l'aide d'une méthode fondée sur les plongements visuels sémantiques (*visual semantic*

embeddings) de (FAGHRI, FLEET, KIROS et al. 2017). Cette dernière tire parti d'un CNN pour créer des représentations d'images et de réseaux de neurones récurrents pour créer des représentations des séquences de mots.

Une fois le score d'appariement obtenu pour chaque paire (*image*, *texte*), il faut déterminer une association globale, sachant qu'elle n'est ni injective ni surjective (certains éléments ne sont pas associés, d'autres ont des associations multiples), ce qui empêche d'utiliser des algorithmes optimaux trouvant une bijection comme l'algorithme hongrois de KUHN (1955). Cette association est réalisée à l'aide de l'heuristique suivante : les paires de plus fort score sont sélectionnées itérativement de manière gloutonne, chaque boîte pouvant être attribuée à un groupe nominal au maximum. Seules les paires de score supérieur à 0,3 sont considérées (seuil déterminé sur un corpus de développement).

Nous présentons dans la prochaine section le cadre expérimental utilisé pour la réalisation de l'alignement automatique entre les groupes nominaux d'une légende d'image avec les parties de l'image correspondante.

4.4 Cadre expérimental

Les expériences ont été réalisées sur le corpus Flickr30k Entities augmenté des annotations présentées en section 5.2. Il a été subdivisé en trois ensembles : apprentissage (23 254 prépositions), développement (2 907 prépositions) et test (2 907 prépositions).

4.4.1 Traitement de l'image

Côté visuel, nous avons ré-entraîné le modèle YOLOv2 sur le corpus Flickr30k Entities en utilisant comme initialisation les poids fournis par les auteurs et en limitant le nombre de catégories aux huit catégories sémantiques du corpus Flickr30k Entities. Seules les prédictions avec un score de confiance supérieur à 0,1 ont été retenues. Sur les 14 229 boîtes des images issues de notre corpus de test, le système en a détecté 7 110. Un objet est considéré comme détecté si l'IoU entre sa boîte de référence et la boîte prédite est supérieur à 0,5.

Le détecteur atteint, sur l'ensemble de test, un rappel de 0,49 et une précision de 0,29. Si nous prenons en compte les catégories sémantiques prédites par le modèle YOLOv2 avec chaque boîte dans le calcul des scores, les performances descendent respectivement à 0,25 et 0,15. Ces résultats nous montrent qu'il s'agit d'une tâche difficile et que le traitement automatique de l'image dans cette tâche de détection représente une première barrière à l'utilisation de l'information visuelle.

Ensuite, le contenu de chaque boîte est redimensionné en 224 par 224 pixels, puis passé en entrée d'un réseau de type ResNet-512 (HE, X. ZHANG, REN et al. 2016) ; pré-entraîné pour la tâche de classification d'images dans mille catégories de scènes du ImageNet Large Scale Visual Recognition Challenge (RUSSAKOVSKY, DENG, SU et al. 2015). Les paramètres du réseau sont ajustés pour générer des représentations dans l'espace commun image/texte. La dernière couche du réseau est remplacée par une couche dense de taille 1 024. Grâce à ces traitements, chaque boîte est associée à un vecteur de taille 1024 qui est utilisé comme représentation de la partie image dans la méthode d'alignement décrite précédemment.

4.4.2 Traitement du texte

Côté texte, les mots des groupes nominaux sont d'abord projetés dans un espace de représentation de taille 300 qui fournit des entrées à une couche récurrente de type GRU (dont la tâche est la prédiction du mot suivant) avec une représentation cachée de taille 1 024. La représentation cachée du réseau récurrent, à l'issue de la lecture des mots d'un groupe nominal est utilisée comme représentation pour la modalité textuelle.

4.4.3 Appariement

Pour chaque représentation des deux modalités obtenues précédemment, une normalisation ℓ_2 est appliquée afin qu'elles puissent être comparées à l'aide du produit scalaire ; ceci est équivalent au calcul de la similarité cosinus entre les deux vecteurs. L'apprentissage est réalisé sur des *batches* de taille 48 pendant trente époques à l'aide de la méthode d'optimisation Adam (KINGMA et BA 2014). Il s'agit d'un modèle dans la famille *triplet ranking*. L'apprentissage est effectué en calculant la similarité entre une paire (**image**, **texte**) existant dans les données d'apprentissage et une paire aléatoire avec l'un des deux membres en commun. Les poids du modèle sont ajustés pour que le score de la paire valide soit supérieur à celui de la paire invalide.

La table 4.1 présente les performances du système d'appariement entre boîtes et groupes nominaux. Nous avons utilisé dans cette évaluation les boîtes de référence et non celles prédites par le modèle YOLOv2. Le taux d'erreur est calculé de la manière suivante : chaque groupe nominal est associé à la boîte prédite automatiquement avec laquelle il obtient le score d'appariement le plus élevé ; s'il s'agit de la boîte qui lui correspond dans les données de référence, alors l'association est considérée comme correcte. Les résultats sont calculés selon deux méthodes d'appariement : VSE et VSE++, qui diffèrent par la fonction de coût utilisée pour l'apprentissage (FAGHRI, FLEET, KIROS et al. 2017) et selon le

Méthode	Entraînement	Taux d'erreur
VSE++	Flickr30k	42,07 %
VSE++	MS COCO	38,90 %
VSE	MS COCO	37,17 %
VSE	ré-entraîné	21,47 %

TABLE 4.1 – Taux d’erreur d’appariement sur notre ensemble de test en comparant les boîtes de référence et les groupes nominaux de référence, selon le modèle (VSE, VSE++) et le corpus d’entraînement : Flickr30k et Microsoft COCO sont les modèles fournis avec l’outil, entraînés sur des paires (**phrase complète, image complète**) plutôt que des groupes nominaux et le contenu des boîtes.

modèle utilisé : ceux fournis avec l’outil VSEpp² entraînés sur les corpus Flickr30k et Microsoft Common Objects in Context (T.-Y. LIN, MAIRE, BELONGIE et al. 2014) ou le modèle ré-entraîné sur les données de notre tâche (Flickr30k Entities). Les modèles disponibles avec l’implémentation VSEpp ont été entraînés sur des images complètes et des phrases de description complètes. Leurs performances s’écroulent sur les boîtes n’englobant qu’un objet dans notre corpus, doublant le nombre d’erreurs d’association, comparé au même modèle ré-entraîné sur les données cibles (de 21 % à 37 %) ; cela démontre l’importance de ré-entraîner le modèle dans des conditions identiques à celles du test.

4.5 Conclusion

Ce chapitre présente plusieurs corpus pouvant être utilisés dans le domaine de l’apprentissage multimodal texte/image. S’il est des tâches pour lesquelles l’alignement n’est pas nécessaire, comme la tâche de **SRR**, entre le texte et l’image, il est souvent obligatoire pour des tâches de **TAL**. En effet, de nombreuses tâches s’intéressent à comprendre la structure d’une phrase, notamment d’un point de vue sémantique, en analysant les relations entre les mots. La description d’une scène elle aussi, passe par la mise en relation des différentes parties d’une image.

La compréhension multimodale doit pouvoir identifier les relations établies entre les différentes entités d’un document dans l’ensemble des modalités à sa disposition. Cela passe ainsi par l’accès à un alignement des entités entre chaque modalité.

Nous avons présenté dans ce chapitre une méthode automatique pour la réalisation d’un tel alignement. Ce dernier est utilisé dans la suite de ce document pour

2. github.com/fartashf/vsepp

une application de la compréhension multimodale à une tâche de [TAL](#).

Chapitre 5

Compréhension multimodale pour le TAL

Sommaire

5.1	Introduction	81
5.2	Adaptation d'un corpus multimodal pour la tâche de désambiguïsation	83
5.2.1	Traitement linguistique automatique des descriptions . . .	83
5.2.2	Sélection des rattachements ambigus	84
5.2.3	Annotation manuelle de rattachements prépositionnels . .	84
5.3	Désambiguïsation de rattachements prépositionnels	86
5.3.1	Bases de connaissances sémantiques	86
5.3.2	Utilisation de corpus	88
5.3.3	Approches multimodales	88
5.4	Stratégie de correction de rattachements prépositionnels	89
5.4.1	Détection des erreurs de rattachement	90
5.4.2	Stratégie de correction	91
5.5	Expériences	93
5.5.1	Résultats	93
5.5.2	Analyse d'erreurs	96
5.6	Conclusion	99

5.1 Introduction

Nous présentons dans ce chapitre les travaux sur l'utilisation de la fusion multimodale pour une tâche de traitement automatique du langage. Ce chapitre décrit une approche dans laquelle, l'information présente dans une modalité est récupérée puis exploitée afin d'aider à la prise de décision pour une tâche dans une autre modalité. Il s'agit ici de comprendre comment une image peut être exploitée afin d'aider dans une tâche de TAL. Dans le chapitre précédent nous avons proposé une méthode pour aligner des groupes nominaux sur des éléments

visuels dans une image. Cette méthode permet de spatialiser les composants d’une phrase, ce qui peut permettre une meilleure compréhension de celle-ci, comme dans l’exemple initial de cette thèse : « Jean regarde l’homme avec un télescope ». Ce chapitre porte sur l’étude de l’utilisation de tels indices visuels pour la désambiguïsation de phrase. Nous nous intéresserons à un phénomène particulier qui est la désambiguïsation des rattachements prépositionnels lors d’un processus d’analyse syntaxique.

Les rattachements prépositionnels sont une source importante d’ambiguïté. Ils constituent une des constructions les plus difficiles à traiter (KUMMERFELD, D. L. W. HALL, CURRAN et al. 2012) par un analyseur syntaxique et représentent environ vingt pour cent des erreurs de ce dernier (KOK et HINRICHS 2016). Une part de la difficulté provient du fait que cette ambiguïté ne peut souvent pas être levée en utilisant uniquement des indices linguistiques (comme dans l’exemple précédent). Même si cette ambiguïté peut parfois être levée par des contraintes sémantiques comme dans la phrase « Je commande une glace à la serveuse », un analyseur ne dispose souvent pas de ces contraintes permettant d’écarter des rattachements incorrects. Cependant, ajouter de l’information venant d’une source visuelle à l’analyse d’une phrase est une tâche difficile car il faut d’une part extraire de l’information pertinente de cette source, et d’autre part, pouvoir la mettre en relation avec la phrase.

Les travaux que nous présentons dans ce chapitre s’articulent autour d’un double objectif. Premièrement, proposer une méthode de résolution de rattachements prépositionnels fondée sur l’utilisation d’indices visuels. Et, dans un second temps, analyser l’impact de l’ajout d’informations visuelles sur la tâche d’analyse syntaxique. Pour cela, nous utilisons un corpus constitué de paires composées d’une photo et d’une légende décrivant cette dernière. Il s’agit du corpus Flickr30k Entities décrit dans la section 4.2.3. Ce corpus met à disposition simultanément l’analyse de l’image (par l’intermédiaire des boîtes), du texte, (à travers certains rattachements prépositionnels) ainsi que l’alignement entre boîtes et groupes nominaux. Cela permet d’établir un lien entre les deux modalités et d’utiliser des informations provenant de l’image pour traiter le texte.

L’approche que nous proposons pour améliorer le rattachement des groupes prépositionnels dans un analyseur syntaxique à l’aide d’informations provenant de l’image, repose sur un détecteur d’erreurs de rattachement, proposant un rattachement alternatif s’il détecte une erreur. L’originalité de ce détecteur est qu’il permet de prendre en entrée, des indices lexicaux mais aussi visuels et conceptuels. Dans le groupe nominal *a ball in front of a dog with a red collar* (un ballon face à un chien avec collier rouge), par exemple, correspondant à l’image reproduite en figure 4.1 ; la décision de rattacher *with* à *dog* plutôt qu’à *ball* peut se fonder sur des indices lexicaux évidents, mais pourrait aussi se fonder sur des indices visuels, en étudiant par exemple, les positions relatives des boîtes correspondant aux mots *ball*, *dog* et *collar*.

L'ensemble des contributions que nous présentons dans ces travaux est une chaîne de traitement complète de légende d'image, dans laquelle nous utilisons de manière originale l'image correspondante pour une tâche de traitement automatique du langage.

La section 5.2 décrit les ressources utilisées et leur adaptation pour notre tâche. Nous dressons l'état de l'art du problème des rattachements prépositionnels dans la section 5.3, en se focalisant sur les études les plus pertinentes pour nos travaux. La section 5.4 décrit notre stratégie de correction et les différents modules qui la composent. Nous détaillons le cadre expérimental en section 5.5 et les résultats obtenus sur notre corpus d'étude en section 5.5.1. Enfin nous présentons une analyse d'erreurs de notre système en section 5.5.2.

5.2 Adaptation d'un corpus multimodal pour la tâche de désambiguïsation

Le corpus Flickr30k Entities est très intéressant pour son utilisation dans l'étude du traitement automatique du langage naturel grâce à l'annotation dont il dispose. Cependant, l'adaptation de ce corpus pour son utilisation dans un cadre de traitement automatique du langage, requiert une quantité de ressources relativement importante. Comme nous l'avons expliqué en introduction de ce chapitre, la modalité visuelle est parfois nécessaire pour des tâches de traitement automatique du langage, comme pour la désambiguïsation de rattachements prépositionnels. Les rattachements prépositionnels sont une des sources d'erreurs principales des analyseurs syntaxiques. Nous avons donc procédé à une annotation de ce corpus en nous limitant aux rattachements prépositionnels dans des contextes ambigus.

5.2.1 Traitement linguistique automatique des descriptions

Toutes les légendes du corpus Flickr30k Entities sont analysées par un étiqueteur en partie de discours de la chaîne de traitement linguistique séquentielle Macaon (NASR, BÉCHET, REY et al. 2011) entraîné sur le corpus Penn Treebank (MARCUS, MARCINKIEWICZ et SANTORINI 1993). Nous avons aussi réalisé l'analyse syntaxique en dépendances, pour identifier automatiquement les rattachements prépositionnels dans les légendes, toujours à l'aide de Macaon. Ce module a été entraîné sur le corpus Penn Treebank (MARCUS, MARCINKIEWICZ et SANTORINI 1993). La figure 5.1 présente les différentes informations obtenues.

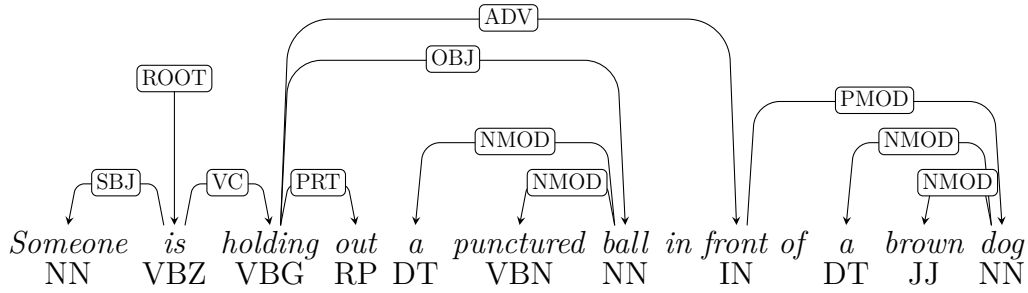


FIGURE 5.1 – Informations linguistiques produites automatiquement par l’outil Macaon : pour chaque mot l’analyseur nous donne sa partie de discours (NN, VB, IN, etc.), les arcs représentent les fonctions syntaxiques.

5.2.2 Sélection des rattachements ambigus

Afin de limiter l’annotation aux rattachements potentiellement ambigus, nous avons procédé à leur sélection en utilisant deux règles simples sur les étiquettes en partie de discours. Nous considérons un rattachement prépositionnel comme ambigu si la préposition est précédée dans la phrase par au moins deux noms ou un verbe et un nom. Ces deux règles peuvent être formalisées par les deux expressions régulières suivantes :

- $X^* N X^* N X^* p X^*$
- $X^* V X^* N X^* p X^*$

où p correspond à la préposition cible analysée, N et V correspondent respectivement aux étiquettes en partie de discours **NOM** et **VERBE** et X à n’importe quelle autre étiquette. L’étoile $*$ représentant le quantificateur zéro ou plusieurs. Un exemple d’utilisation de la seconde règle est présenté en figure 5.2.



FIGURE 5.2 – Exemple de l’utilisation de la seconde règle pour la sélection des rattachements ambigus.

Par ces règles, nous avons sélectionné 22 800 légendes dans 15 700 images différentes, soit un total de 29 068 prépositions.

5.2.3 Annotation manuelle de rattachements prépositionnels

Chaque préposition a été manuellement rattachée à son gouverneur syntaxique. L’annotation a été réalisée par une seule personne. La figure 5.3 montre les informations limitées mises à la disposition de l’annotateur, à savoir : la description

en entier, la préposition cible qui devait être rattachée à son gouverneur (indiquée par le drapeau *target*) et l'image correspondante.



FIGURE 5.3 – Interface utilisée pour l’annotation manuelle des rattachements prépositionnels : la légende est présentée avec l’image qui lui correspond ; la préposition cible est indiquée par le mot-clé « target » ; l’annotateur doit alors sélectionner le gouverneur de la préposition.

L’annotation ainsi produite est présentée dans la table 5.1 :

- les colonnes *Lemme*, *PoS* (*part of speech*), *Gov.* (gouverneur) et label *Label* ont été produites automatiquement par la chaîne de traitement linguistique séquentielle Macaon ; nous y trouvons dans l’ordre : le lemme du mot, sa partie de discours, le numéro du mot qui est son gouverneur dans l’arbre syntaxique et la fonction syntaxique ;
- la colonne *Entité* correspond au numéro de l’entité dans l’annotation du corpus Flickr30k Entites ;
- enfin la colonne *Ref.* correspond au numéro du gouverneur annoté manuellement.

La table 5.2 présente un résumé de quelques statistiques du corpus pour les vingt prépositions les plus fréquentes ainsi que sur l’ensemble du corpus en dernière ligne. Les prépositions sont ordonnées par occurrence, le taux de rattachement à un nom et à un verbe est donné, ainsi que la distance moyenne en nombre de mots entre la préposition et son gouverneur. Les statistiques sur l’intégralité du corpus sont données en annexe A de ce document.

Nous pouvons observer que 64 % des prépositions sont rattachées à un verbe et que la distance moyenne entre une préposition et son gouverneur est de 2,37 mots. Les dix prépositions les plus fréquentes représentent 61,22 % du corpus annoté. 25 prépositions ont une distance moyenne avec leur gouverneur syntaxique de plus de trois mots (soit 4 497 rattachements), et leur gouverneur est un verbe dans 74 % des cas.

Les deux dernières colonnes de cette table correspondent au nombre d’occurrences et au score de rattachement de l’analyseur pour chaque préposition dans l’ensemble de test (tel que défini dans la section 4.4). Ces chiffres montrent que sur le corpus de test, l’analyseur réalise 25 % d’erreurs sur les rattachements prépositionnels. Ce nombre important d’erreurs justifie de développer une stratégie de correction d’erreurs.

ID	Mot	Lemme	PoS	Gov.	Label	Entité	Ref.
1	Someone	someone	NN	2	SBJ	227018	–
2	is	be	VBZ	0	ROOT	–	–
3	holding	hold	VBG	2	VC	–	–
4	out	out	RP	3	PRT	–	–
5	a	a	DT	7	NMOD	227019	–
6	punctured	puncture	VDN	7	NMOD	227019	–
7	ball	ball	NN	3	OBJ	227019	–
8	in front of	in front of	IN	3	ADV	–	3
9	a	a	DT	11	NMOD	227017	–
10	brown	brown	JJ	11	NMOD	227017	–
11	dog	dog	NN	8	PMOD	227017	–
12	with	with	IN	11	NMOD	–	11
13	a	a	DT	15	NMOD	227021	–
14	red	red	JJ	15	NMOD	227021	–
15	collar	collar	NN	12	PMOD	227021	–
16	.	.	.	2	P	–	–

TABLE 5.1 – Ensemble des informations ajoutées au corpus Flickr30k Entities pour son utilisation dans un cadre de traitement automatique du langage.

5.3 Désambiguïsation de rattachements prépositionnels

Le problème du rattachement prépositionnel, fait l’objet d’un grand nombre d’études en traitement automatique des langues. Il constitue un problème important et difficile pour les analyseurs syntaxiques. De nombreuses sources d’informations et méthodes ont été proposées pour le résoudre.

5.3.1 Bases de connaissances sémantiques

Une des ressources très utilisée dans la littérature pour résoudre le problème des rattachements prépositionnels est les bases de connaissances sémantiques. Plusieurs approches ont montré l’efficacité de l’ajout d’informations sémantiques, notamment issues de la ressource WordNet (MILLER 1995). Ces dernières sont décrites dans le paragraphe suivant.

Les travaux de AGIRRE, BALDWIN et MARTINEZ (2008) montrent l’apport de l’utilisation d’informations sémantiques pour la désambiguïsation de rattachements prépositionnels. Cette approche consiste à remplacer lors de l’entraînement

Préposition	Occ.	Nom	Verbe	Distance	Occ. Test	Parser
<i>in</i>	4 191	0,59	0,41	2,21	369	0,76
<i>with</i>	3 018	0,59	0,41	2,47	310	0,65
<i>for</i>	1 777	0,36	0,64	1,57	168	0,73
<i>near</i>	1 452	0,65	0,35	1,69	159	0,33
<i>through</i>	1 420	0,05	0,95	2,01	145	0,95
<i>on</i>	1 359	0,26	0,74	2,03	143	0,85
<i>next to</i>	1 342	0,08	0,92	3,23	137	0,89
<i>from</i>	1 172	0,30	0,70	2,43	140	0,76
<i>into</i>	1 123	0,08	0,92	2,32	116	0,89
<i>over</i>	941	0,39	0,61	2,38	111	0,66
<i>by</i>	890	0,10	0,90	2,60	76	0,84
<i>at</i>	720	0,20	0,80	2,75	63	0,84
<i>of</i>	700	0,97	0,03	1,03	72	0,93
<i>around</i>	589	0,18	0,82	2,26	59	0,73
<i>in front of</i>	570	0,11	0,89	3,72	51	0,90
<i>under</i>	544	0,18	0,82	3,29	58	0,84
<i>behind</i>	544	0,35	0,65	1,78	35	0,74
<i>along</i>	500	0,37	0,63	1,79	50	0,52
<i>during</i>	423	0,14	0,86	5,08	41	0,71
<i>across</i>	415	0,11	0,89	2,21	49	0,88
Corpus	29 068	0,36	0,64	2,37	2 907	0,75

TABLE 5.2 – Statistiques sur les rattachements des prépositions annotées et leur nombre d’occurrences : les colonnes *Nom* et *Verbe* correspondent respectivement au taux de rattachement à un nom ou un verbe. Les deux dernières colonnes présentent les taux de rattachements corrects de l’analyseur en dépendances uniquement sur le corpus de test.

de l'analyseur des mots par des classes sémantiques basées sur la ressource WordNet (MILLER 1995) dans laquelle les mots sont organisés par catégories morphosyntaxiques et groupes de synonymes (*synsets*). Chaque *synset* est aussi classé dans des groupes sémantiques (*semantic file*) en fonction de sa catégorie syntaxique et sémantique. Par exemple le mot « *ball* » est associé au *synset* « *round object that is hit or thrown or kicked in games* », et appartient au *semantic file* « *Artifact* ». Leurs expériences montrent que l'utilisation de cette information sémantique permet d'obtenir de meilleurs résultats pour l'analyse de rattachements prépositionnels.

DASIGI, AMMAR, DYER et al. (2017) se sont aussi intéressés à l'ajout d'une telle information sémantique en utilisant les *synsets* de la ressource WordNet pour la construction originale de plongements de mots. En utilisant un modèle à base de réseaux de neurones récurrents, ils montrent que l'utilisation de bases de connaissances sémantiques améliore leurs résultats pour la prédiction de rattachements prépositionnels.

5.3.2 Utilisation de corpus

Une autre approche consiste en l'utilisation de corpus. RAKSHIT, SONTAKKE, BHATTACHARYYA et al. (2016) présentent une méthode leur permettant d'améliorer leurs résultats pour la désambiguïsation de rattachements prépositionnels de dix pour cent en utilisant l'alignement d'une phrase dans deux langues différentes (anglais et hindi). Ils utilisent un algorithme qui vise à faire converger les arbres d'analyse syntaxique dans les deux langues. Nous retrouvons aussi plusieurs travaux utilisant des plongements de mots (BELINKOV, LEI, BARZILAY et al. 2014; KOK, MA, DIMA et al. 2017).

5.3.3 Approches multimodales

À notre connaissance, peu de travaux utilisent des informations multimodales pour traiter ce problème. Les travaux les plus pertinents pour nous sont ceux de CHRISTIE, LADDHA, AGRAWAL et al. (2016); leur approche consiste à réaliser simultanément l'analyse visuelle (identification de boîtes) et l'analyse syntaxique pour des paires (**image**, **phrase**) puis ils considèrent le produit cartésien des analyses syntaxiques et visuelles. Les différentes paires se voient attribuer un score et la paire obtenant le meilleur score est alors sélectionnée. La différence principale entre leurs travaux et les nôtres est que nous produisons une unique analyse syntaxique et que cette dernière est corrigée en fonction des informations visuelles. De plus, nous menons des expériences sur un nombre de paires (**image**, **phrase**) beaucoup plus important (22 800 contre 1 822).

La désambiguïsation des rattachements prépositionnels à l'aide d'informations visuelles est aussi en lien avec l'apprentissage de relations visuelles. Les travaux les plus pertinents sont ceux de PEYRE, LAPTEV, SCHMID et al. (2017) dans lesquels les auteurs développent de nouveaux descripteurs visuels pour la représentation de relations entre les objets d'une image. Leur modèle repose sur des représentations multimodales des configurations d'objets pour chaque relation, et entraînent des classifieurs sur la relation des objets avec une supervision au niveau de l'image seulement (c'est-à-dire des annotations du niveau image comme *person on bike*, sans annoter les objets impliqués dans la relation). Alors que nous pourrions utiliser leurs classifieurs de relations spatiales, l'objectif de notre travail est différent. Nous nous intéressons au problème de la désambiguïsation des rattachements prépositionnels. Nous utilisons des caractéristiques visuelles similaires pour représenter la configuration spatiale des objets ; cependant, les objets sont détectés et représentés d'une manière différente. Notre système réalise aussi l'alignement de segments de phrases (plus précisément des groupes nominaux) avec des boîtes dans l'image afin de pouvoir utiliser des caractéristiques multimodales, sans que cela soit notre objectif principal.

5.4 Stratégie de correction de rattachements prépositionnels

Comme nous l'avons vu dans la table 5.2, l'analyseur syntaxique en dépendances réalise 25 % d'erreurs de rattachements sur notre corpus de test. Nous proposons alors une approche afin de corriger ces rattachements incorrects en utilisant des caractéristiques visuelles. Le modèle que nous proposons, prend en entrée une image et une phrase et fournit une analyse syntaxique de la phrase en suivant les étapes décrites ci-dessous :

1. détection des groupes nominaux dans la phrase ;
2. détection d'entités dans l'image ;
3. alignement automatique entre les deux objets précédents ;
4. analyse syntaxique automatique de la phrase ;
5. détection de rattachements prépositionnels incorrects ;
6. correction des rattachements détectés comme incorrects.

La détection des entités dans l'image et sa légende ainsi que leur alignement sont présentés en section 4.3. Cet alignement permet d'intégrer des caractéristiques visuelles pour la tâche de correction des rattachements prépositionnels, ces derniers étant produits par un analyseur syntaxique. L'analyse syntaxique automatique de la phrase est décrite dans la section 5.2.1. Cette section présente le module de correction qui se décompose en deux étapes. Une première étape réalisant la détection des erreurs de rattachements produites par l'analyseur syntaxique à

l'aide d'un classifieur. Une seconde étape, correspondant à la stratégie de correction dans laquelle des candidats gouverneurs à la préposition cible sont sélectionnés puis évalués là encore, avec un classifieur ; le gouverneur au score le plus élevé étant sélectionné.

Le choix d'une telle stratégie plutôt que l'intégration des indices visuels directement au niveau de l'analyseur, repose sur deux facteurs majeurs. Premièrement, nous souhaitons étudier l'influence de l'information visuelle indépendamment de l'analyse sémantique de la légende. Dans une seconde mesure, la mise en œuvre d'une intégration au niveau de l'analyse est beaucoup plus complexe que l'utilisation d'un système de correction en sortie d'analyse.

5.4.1 Détection des erreurs de rattachement

La détection des erreurs de rattachement est réalisée par un classifieur utilisant trois catégories de caractéristiques : lexicales, visuelles conceptuelles et visuelles spatiales. Ces caractéristiques portent sur la préposition p , son gouverneur G et son objet O . Lorsque le gouverneur est un verbe, c'est le sujet du verbe qui fait office de G . Ainsi, dans la phrase « Jean mange avec des gants. », on obtient $G = \text{Jean}$, $p = \text{avec}$ et $O = \text{gants}$.

Pour les caractéristiques lexicales, nous utilisons en partant de l'arbre en dépendance produit par un analyseur : le lemme et la catégorie grammaticale du gouverneur et de l'objet, la distance entre la préposition et son gouverneur. Soit $G \rightarrow p \rightarrow O$ un rattachement prépositionnel issu de l'analyse syntaxique d'une phrase ; nous définissons les fonctions : $POS(X)$ qui attribue à un mot sa partie de discours, $LEM(X)$ son lemme et $FCT(X)$ sa fonction syntaxique. Nous définissons aussi la fonction $DIST(X,Y)$ qui représente la distance (en nombre de mots) entre deux mots X et Y . L'ensemble des caractéristiques lexicales utilisé est initialisé avec le lemme de la préposition cible $LEM(p)$ puis augmenté avec les caractéristiques lexicales L_i suivantes :

$$\begin{array}{lll} L_1 = POS(G) & L_4 = LEM(G) & L_7 = FCT(p) \\ L_2 = POS(D) & L_5 = LEM(D) & L_8 = DIST(G,p) \\ L_3 = POS(G) + POS(D) & L_6 = LEM(G) + LEM(D) & \end{array}$$

Les caractéristiques visuelles conceptuelles et spatiales sont calculées à partir des boîtes englobantes que le système d'appariement a associé au gouverneur et à l'objet de la préposition. Les caractéristiques conceptuelles (personne, partie du corps, animal, vêtement, instrument, véhicule et autre) correspondent à celles utilisées dans le corpus Flickr30k Entities ; elles sont prédites lors de la détection d'objet dans l'image. Par ailleurs dans le cas où le module d'appariement n'a sélectionné aucune boîte pour un des deux groupes nominaux sélectionnés

(gouverneur ou objet), la valeur *UNK* est utilisée pour représenter le concept de ce groupe nominal et aucune des caractéristiques spatiales n'est calculée.

En ce qui concerne les caractéristiques visuelles spatiales, étant donné la boîte du gouverneur $b_G = [x_g, y_g, w_g, h_g]$ et la boîte de l'objet $b_O = [x_d, y_d, w_d, h_d]$ de la préposition, où (x, y) sont les coordonnées du centre de la boîte, et (w, h) sont la largeur et la hauteur de la boîte, nous utilisons les caractéristiques proposées par PEYRE, LAPTEV, SCHMID et al. (2017) :

$$\begin{aligned} V_{S1} &= \frac{x_d - x_g}{\sqrt{w_g h_g}} & V_{S3} &= \sqrt{\frac{w_d h_d}{w_g h_g}} & V_{S5} &= \frac{w_g}{h_g} \\ V_{S2} &= \frac{y_d - y_g}{\sqrt{w_g h_g}} & V_{S4} &= \frac{b_g \cap b_d}{b_g \cup b_d} & V_{S6} &= \frac{w_d}{h_d} \end{aligned}$$

Les caractéristiques V_{S1} et V_{S2} représentent respectivement les positions relatives horizontale et verticale entre les centres de deux boîtes. V_{S3} est le rapport entre les tailles des boîtes, V_{S4} le rapport de l'intersection et l'union entre les deux boîtes, et V_{S5} et V_{S6} le rapport de forme des deux boîtes. L'annotation du corpus Flickr30k Entities peut fournir plusieurs boîtes pour une même entité. Afin de faciliter nos calculs, nous avons fait le choix de ne garder qu'une seule boîte par groupe nominal. Cette sélection se fait en gardant la boîte dont le centre est le plus proche du barycentre de toutes les boîtes de l'entité.

La détection d'erreurs de rattachement est réalisée à l'aide du classifieur Icsi-boost (FAVRE, HAKKANI-TÜR et CUENDET 2007) qui est fondé sur l'algorithme AdaBoost (FREUND, R. SCHAPIRE et ABE 1999). Ce classifieur est une combinaison de classifieurs faibles qui apprennent un seuil pour les caractéristiques continues, et un indicateur binaire pour les caractéristiques discrètes. L'apprentissage minimise la fonction de perte exponentielle en sélectionnant de façon gloutonne le meilleur classifieur et repondérant l'ensemble d'apprentissage pour se focaliser sur les exemples mal classés. Ce type de classifieur a deux avantages : les modèles sont simples à interpréter par rapport à d'autres familles de modèles ; la sélection gloutonne du classifieur, choisit efficacement les caractéristiques pertinentes et est moins affecté par le bruit. En se basant sur toutes les caractéristiques, le classifieur vérifie alors si le rattachement proposé par l'analyseur syntaxique est correct ou non.

5.4.2 Stratégie de correction

Afin d'augmenter la précision de l'analyseur syntaxique, nous utilisons une stratégie de correction qui consiste à changer le rattachement proposé par l'analyseur syntaxique en utilisant un correcteur d'erreur (ANGUIANO et CANDITO 2011 ;

K. HALL et NOVÁK 2005 ; ATTARDI et CIARAMITA 2007). Lorsqu'un rattachement est détecté comme incorrect par le classifieur, nous appliquons des règles à l'arbre syntaxique généré par l'analyseur pour obtenir un ensemble de rattachements alternatifs. Ces nouveaux rattachements possibles sont donnés au classifieur pour prendre une décision finale en sélectionnant celui dont la probabilité de rattachement est la meilleure.

Nous faisons le choix de travailler sur la génération de rattachement alternatif plutôt que sur les n-meilleures sorties de l'analyseur. En effet, nous sommes intéressés uniquement par les possibilités alternatives sur les rattachements prépositionnels et pas sur les autres mots. Ainsi, l'ensemble des règles que nous présentons permettent de couvrir la plupart des groupes nominaux ou des verbes de la phrase et nous assure de conserver une structure arborescente à l'arbre syntaxique produit par l'analyseur.

Étant donné une phrase S , une analyse syntaxique T de S et une préposition cible p , nous définissons un ensemble de gouverneurs candidats G_p . L'ensemble G_p est initialisé avec g le gouverneur actuel de la préposition p dans l'analyse syntaxique T . Nous appliquons l'ensemble de règles suivantes à T pour ajouter de nouveaux gouverneurs potentiels à l'ensemble G_p :

$$\text{Initialisation} \Rightarrow G_p = \{g\} \quad (5.1a)$$

$$N \leftarrow V \rightarrow p \Rightarrow G_p = G_p \cup \{N\} \quad (5.1b)$$

$$N \leftarrow P \leftarrow V \rightarrow p \Rightarrow G_p = G_p \cup \{N\} \quad (5.1c)$$

$$N' \leftarrow N \rightarrow p \Rightarrow G_p = G_p \cup \{N'\} \quad (5.1d)$$

$$N' \leftarrow P \leftarrow N \rightarrow p \Rightarrow G_p = G_p \cup \{N'\} \quad (5.1e)$$

$$N' \rightarrow X \rightarrow N \rightarrow p \Rightarrow G_p = G_p \cup \{N'\} \quad (5.1f)$$

$$N \rightarrow N \rightarrow p \Rightarrow G_p = G_p \cup \{N\} \quad (5.1g)$$

$$V \rightarrow N \rightarrow p \Rightarrow G_p = G_p \cup \{V\} \quad (5.1h)$$

Nous pouvons traduire la règle (5.1a) de la façon suivante : si une préposition cible p a un verbe comme gouverneur qui a un nom N comme dépendant direct, alors N est un gouverneur candidat.

En prenant pour exemple la phrase « *a man throws a child into the air at the beach* » et la préposition cible *at*. Si l'analyseur syntaxique nous donne comme gouverneur le mot *child*, alors nous pouvons appliquer les règles (5.1e) et (5.1h) :

$$\begin{array}{llllllll} (5.1a) & man & throws & \underline{child} & into & air & at & \Rightarrow G_p = \{child\} \\ (5.1e) & & & N & P & N* & p & \Rightarrow G_p = G_p \cup \{air\} \\ (5.1h) & & V* & N & & & p & \Rightarrow G_p = G_p \cup \{throws\} \\ & & & & & & & \Rightarrow G_p = \{child, air, throws\} \end{array}$$

5.5 Expériences

Les expériences présentées ici font l’hypothèse d’un cas dégradé dans lequel l’information lexicale n’est pas disponible (la sémantique des mots n’est pas connue par le système), afin de mettre en lumière les informations exploitables dans la partie visuelle de la tâche. Le cas où les caractéristiques lexicales sont utilisées est alors vu comme une borne supérieure. Dans ce contexte, nous nous intéressons à deux questions : quel est l’impact des catégories sémantiques dans la modalité visuelle, et quel est celui de l’information spatiale dans cette même modalité ?

Les expériences ont été réalisées sur le corpus présenté dans la section précédente. Il a été subdivisé de la même manière que dans la section 4.4 en trois ensembles : apprentissage (23 254 prépositions), développement (2 907 prépositions) et test (2 907 prépositions). Le détecteur d’erreurs de rattachement prépositionnel est quant à lui entraîné sur notre corpus d’apprentissage. Les paramètres du classifieur ont été ajustés en fonction de ses performances sur le corpus de développement.

Nous avons aussi utilisé le corpus de développement pour évaluer les performances des règles que nous utilisons pour la recherche de gouverneurs potentiels G_p . En sortie d’analyseur syntaxique, les règles nous permettent de récupérer le bon gouverneur annoté manuellement pour la préposition dans l’ensemble G_p dans 92,28 % des cas. Ce score représente donc une borne supérieure pour les rattachements prépositionnels.

La couverture non maximale s’explique par le choix que nous avons fait de travailler directement sur l’analyse produite. Nous ne pouvons pas simplement prendre l’ensemble des mots comme candidat gouverneur. En effet, cela implique une modification de l’analyse déjà produite pour s’assurer de lui converser une structure arborescente. Une possibilité pour obtenir une couverture de 100 % serait d’utiliser une stratégie de re-scoring à la place de la stratégie de correction. Cette approche consiste à attribuer un score à l’ensemble des analyses syntaxiques en dépendances possibles et de modifier ces scores à l’aide des indices visuels, puis de choisir l’analyse ayant le meilleur score.

5.5.1 Résultats

La table 5.3 présente le taux de bon rattachement pour les vingt prépositions les plus courantes du corpus de test. Les résultats pour l’ensemble des prépositions ayant un nombre d’occurrences supérieur à dix sont donnés en annexe B. Le corpus Flickr30k Entities étant majoritairement composé de légendes compréhensibles pour un humain sans voir l’image associée (sémantiquement non ambiguës), les cas réellement ambigus sont donc rares.

Pour chacune de ces prépositions, six configurations différentes sont évaluées :

Prépositions	#	Parser	V_C	V_S	$V_C + V_S$	L	$V_C + V_S + L$
<i>in</i>	369	0,76	0,76	0,77	0,76	0,85	0,84
<i>with</i>	310	0,65	0,68	0,68	0,70	0,78	0,78
<i>for</i>	168	0,73	0,73	0,73	0,72	0,82	0,83
<i>near</i>	159	0,33	0,53	0,50	0,59	0,84	0,84
<i>through</i>	145	0,95	0,95	0,95	0,95	0,96	0,96
<i>on</i>	143	0,85	0,85	0,85	0,85	0,89	0,87
<i>from</i>	140	0,76	0,76	0,76	0,76	0,86	0,85
<i>next to</i>	137	0,89	0,89	0,89	0,89	0,90	0,89
<i>into</i>	116	0,89	0,89	0,89	0,89	0,92	0,95
<i>over</i>	111	0,66	0,64	0,66	0,68	0,85	0,84
<i>by</i>	76	0,84	0,84	0,84	0,84	0,86	0,86
<i>of</i>	72	0,93	0,93	0,93	0,93	0,93	0,93
<i>at</i>	63	0,84	0,84	0,84	0,84	0,87	0,89
<i>around</i>	59	0,73	0,73	0,73	0,71	0,85	0,80
<i>under</i>	58	0,84	0,84	0,84	0,84	0,84	0,84
<i>in front of</i>	51	0,90	0,90	0,90	0,90	0,88	0,90
<i>along</i>	50	0,52	0,60	0,54	0,58	0,88	0,86
<i>across</i>	49	0,88	0,88	0,88	0,88	0,96	0,98
<i>down</i>	41	0,63	0,59	0,56	0,56	0,73	0,76
<i>during</i>	41	0,71	0,71	0,71	0,71	0,76	0,76
Corpus	2907	0,75	0,77	0,77	0,78	0,85	0,85

TABLE 5.3 – Taux de rattachements corrects sur le test : # indique le nombre d’occurrences de la préposition ; la colonne *parser* correspond au score de l’analyseur syntaxique ; V_C représente les concepts visuels, V_S les caractéristiques spatiales, $V_C + V_S$ la combinaison des deux et L sont les caractéristiques lexicales.

- *parser* qui correspond au score obtenu par l’analyseur sans correction ;
- V_C correspond au score obtenu après correction en n’utilisant que les caractéristiques conceptuelles correspondant aux neuf classes sémantiques distinguées dans le corpus ;
- la configuration V_S où seules les six caractéristiques spatiales sont utilisées ;
- $V_C + V_S$ où l’ensemble des caractéristiques visuelles sont utilisées ;
- L qui correspond à l’utilisation des caractéristiques linguistiques uniquement ;
- et enfin $V_C + V_S + L$ où l’ensemble des caractéristiques sont prises en compte.

Comme nous pouvons le voir dans la table 5.3, la précision globale de l’analyseur syntaxique est de 75 %. Ce score est inférieur aux 86 % de précision présenté dans (ANGUIANO et CANDITO 2011) sur le Penn Treebank en utilisant la même catégorie d’analyseur. Cette différence n’est cependant pas surprenante étant donné la différence de nature de ces deux corpus ; le Penn Treebank étant composé de phrases de type journalistique. Nous pouvons noter que selon les prépositions, nous obtenons des performances très différentes avec notre analyseur, variant de 95 % pour la préposition *through*, à 33 % pour la préposition *near*.

Si un seul type de caractéristique est utilisé, les meilleurs résultats sont obtenus avec les caractéristiques textuelles ; elles augmentent la précision de onze points. Ce n’est pas surprenant car il est bien connu que certaines ambiguïtés peuvent être résolues en utilisant uniquement des informations syntaxiques. Il semble donc assez prévisible que l’information lexicale (L) ait un effet drastique : les performances augmentent fortement pour la plupart des prépositions. Nous pouvons considérer ces scores comme la « *topline* » de notre système. En effet nous pouvons faire une analogie entre l’utilisation des indices lexicaux uniquement et le ré-entraînement de l’analyseur sur les phrases de notre corpus. Le classifieur utilisé dans notre approche étant ainsi capable d’apprendre des régularités lexicales sur le corpus. Le problème est que ce type d’information n’est pas toujours disponible. La question suivante peut alors se poser : d’autres caractéristiques sont-elles utiles pour la résolution des rattachements prépositionnels ?

Alors que les concepts visuels (V_C) et l’information spatiale (V_S) apportent des améliorations variables selon les prépositions, le gain semble surtout toucher des prépositions locatives comme *near* ou des prépositions qui ont un usage très différent du Penn Treebank comme *with*. Nous pourrions penser que les catégories conceptuelles soient suffisantes pour résoudre le problème, mais il faut noter que la tâche de catégorisation visuelle est difficile (et donc le classifieur souvent erroné), et que les catégories sont plutôt grossières et ne permettent peut-être pas de lever toutes les ambiguïtés. La combinaison de V_C et V_S donne les meilleurs résultats et permet de corriger environ trois pour cent relatif des erreurs.

L’apport entre l’utilisation de l’information lexicale uniquement et l’utilisation de toutes les modalités ($V_C + V_S + L$) est nul en moyenne. Un apport qui souligne cependant que la fusion des caractéristiques issues des modalités texte et image, tire parti de la modalité la plus forte sans pâtir d’une modalité aux performances

Caractéristiques	Parser	V_C	V_S	$V_C + V_S$	L	$V_C + V_S + L$
Référence	0,75	0,82	0,77	0,82	0,85	0,86
Automatique	0,75	0,77	0,77	0,77	0,85	0,85

TABLE 5.4 – Comparaison des taux de rattachement prépositionnel correct sur l’ensemble de test en utilisant les alignements de référence et ceux produits automatiquement.

plus faibles. Par ailleurs, les images reprennent souvent le même type d’objet. Un apprentissage sur une partie du corpus permet ainsi de résoudre la plupart des ambiguïtés sémantiques.

À titre de comparaison, nous présentons dans la table 5.4 les taux de rattachements corrects en sortie du module de correction avec un alignement entre le texte et l’image qui n’est pas issu du module d’appariement automatique mais extrait directement depuis le corpus. Il faut donc comprendre que ce sont à la fois les boîtes englobantes des objets et leur classe sémantique qui sont issues de l’annotation de référence.

Il est à noter que les caractéristiques visuelles utilisées sont légèrement différentes de celles décrites précédemment. En effet, dans ce cas, elles se concentrent sur la position relative du coin supérieur droit et inférieur gauche de la boîte du dépendant de la préposition, par rapport à ceux de la boîte du gouverneur, au lieu de se concentrer sur le centre des boîtes. Ces caractéristiques produisent cependant les mêmes performances : 77 % de rattachements corrects. En revanche, les classes sémantiques de référence apportent un gain substantif par rapport à l’utilisation de celles prédites automatiquement.

Là encore ces résultats mettent en évidence les capacités de la fusion des modalités. L’utilisation de l’ensemble des caractéristiques visuelles ($V_C + V_S$: concepts sémantiques et spatiaux) tire parti de la modalité la plus performante. Lors de la fusion de toutes les modalités, le système gagne un point de performance par rapport à l’utilisation de caractéristiques lexicales uniquement, passant de 0,85 à 0,86 avec l’alignement de référence. Il s’agit d’un gain à nuancer, aucun test de significativité n’ayant été réalisé.

5.5.2 Analyse d’erreurs

Nous présentons ici quelques exemples de couples (**image**, **texte**) pour lesquels l’image a permis, ou non, de réaliser un rattachement correct en utilisant des ensembles de traits différents. Dans l’ensemble des exemples, la préposition cible est

en gras, le gouverneur choisi par l'analyseur est souligné et le nouveau gouverneur après correction est entre crochets.

Les figure 5.4a et 5.4b montrent des phrases pour lesquelles l'analyseur a effectué un mauvais rattachement mais que le classifieur a permis de corriger en utilisant uniquement des informations visuelles. Dans la figure 5.4a, l'analyseur propose le mot *wall* comme gouverneur de la préposition *with*, et le classifieur corrige le rattachement en choisissant *sitting* comme gouverneur. Dans la figure 5.4a la préposition *near* est incorrectement rattachée à *area* par l'analyseur. Là encore, le classifieur permet de réaliser le rattachement correct. Ces exemples constituent la justification de cette étude : corriger de mauvais rattachements grâce à des informations visuelles.



(a) A boy [*sitting*] on a concrete wall **with** a hat on. (b) Two children [*are*] in a grassy area **near** two horses.

FIGURE 5.4 – Exemples d'images pour lesquelles le classifieur a utilisé uniquement les caractéristiques visuelles pour corriger le rattachement de la préposition.

Les figures 5.5a et 5.5b montrent des phrases pour lesquelles l'utilisation de caractéristiques visuelles uniquement, n'a pas permis de corriger un rattachement erroné. Dans la figure 5.5a, la préposition *on* est incorrectement rattachée au mot *building* et dans la figure 5.5b la préposition *in* est incorrectement rattachée au mot *crosswalk*. Dans les deux cas, le système d'appariement n'a pas trouvé de boîtes englobantes pour au moins un des deux groupes nominaux. Ces exemples constituent l'une des limites de cette étude : la difficulté de la phase de détection et d'appariement entre des boîtes et des groupes nominaux limite l'impact des traits visuels dans la correction des analyses erronées.

Même si les traits lexicaux sont les plus performants, si le corpus d'apprentissage ne contient pas assez d'exemples pour certaines entités, les caractéristiques visuelles



(a) *Two people sitting in front of an older building on a bench*

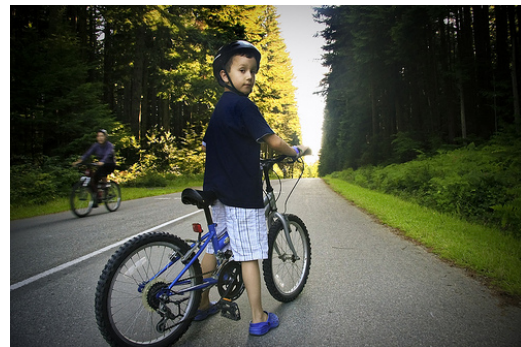


(b) *Two younger women and an older man in a red sweatshirt are walking across a crosswalk **in** San Francisco*

FIGURE 5.5 – Exemples d’images pour lesquelles le classifieur a mal corrigé le rattachement de la préposition avec des caractéristiques visuelles uniquement.



(a) *A dog is wearing [jeans] and a blue and yellow shirt **with** a black vehicle in the background*



(b) *A boy with a black shirt and white shorts **on** a [bike] is turning to look behind himself*

FIGURE 5.6 – Exemples d’images avec un rattachement correct en utilisant des caractéristiques visuelles pour les prépositions alors que l’utilisation de traits lexicaux choisit un mauvais gouverneur.

peuvent s'avérer plus performantes. Ainsi les figures 5.6a et 5.6b présentent des phrases pour lesquelles l'utilisation de caractéristiques visuelles uniquement permet d'effectuer le bon rattachement, alors que l'utilisation de caractéristiques linguistiques seules produit une erreur. Dans la figure 5.6a, la préposition *with* est incorrectement rattachée au mot *jeans* à la place du mot *wearing*. Dans la figure 5.6b, la préposition *in* est rattachée au mot *bike* au lieu du mot *boy*.

5.6 Conclusion

Les travaux présentés dans ce chapitre explorent la possibilité d'utiliser l'information d'une modalité pour aider à l'analyse d'une autre modalité dans un même document. Nous avons proposé d'utiliser des images pour désambiguïser les rattachements prépositionnels dans des phrases les décrivant. Les caractéristiques visuelles améliorent en moyenne de trois points les performances selon les prépositions, et parfois de manière drastique comme pour la préposition *near*.

Trois raisons permettent d'expliquer les résultats modestes obtenus par les caractéristiques visuelles. Dans certains cas, il n'y a rien dans l'image qui permet de lever une ambiguïté d'attachement. D'autre part, dans les cas où l'information visuelle permet de lever l'ambiguïté, les caractéristiques utilisées ne représentent pas toujours les traits pertinents de l'image pour la lever. Finalement, les caractéristiques visuelles sont prédites. Cette prédiction est difficile : les boîtes ne sont pas toujours bien détectées et la classe sémantique qui leur correspond n'est pas toujours bien prédite. L'information lexicale quant à elle, est beaucoup plus fiable. Elle n'est pas prédite mais donnée, et dans de nombreux cas les mots suffisent à lever l'ambiguïté.

Lorsque nous combinons toutes les caractéristiques (visuelles et lexicales), il n'y a pas de gain en moyenne sur les prépositions sélectionnées. Cependant, nous observons des améliorations comme pour la préposition *into*, qui dénote généralement des relations spatiales, et pour laquelle le détecteur arrive à exploiter les caractéristiques visuelles.

Nous pourrions penser que les catégories conceptuelles sont suffisantes pour résoudre le problème, mais il faut noter que la tâche de catégorisation visuelle est difficile (et donc le classifieur souvent erroné), et que les catégories sont assez grossières et ne permettent peut-être pas de lever toutes les ambiguïtés.

Une meilleure utilisation de l'information provenant de l'image est en ce sens une piste majeure d'amélioration du système ; avec notamment, l'intégration de l'information issue directement des pixels comme l'utilisation d'espace de représentation pour l'image ou directement au niveau des boîtes englobantes. Par ailleurs, l'alignement automatique tel qu'il est utilisé n'est clairement pas adapté. Le conditionnement de la détection des objets est une amélioration envisageable.

Chapitre 6

Conclusion générale

La compréhension multimodale a pour objectif de comprendre l'information qui est nécessaire à la prise de décision pour une tâche en analysant les différentes modalités à sa disposition. Il s'agit d'une faculté que le cerveau humain utilise à chaque instant pour comprendre le monde qui l'entoure. Cependant, transférer cette faculté dans une machine soulève de nombreuses problématiques auxquelles nous nous sommes intéressés dans ce document.

6.1 Conclusion

Le chapitre 2 présente une définition générale de la compréhension multimodale pour le traitement automatique de documents. Les deux étapes clés de sa mise en œuvre sont d'une part l'extraction de caractéristiques dans les différentes modalités afin d'en obtenir des représentations. Les performances des DNN pour cette tâche dans le traitement de l'image et de texte ont été abordées en section 2.2. D'autre part, il s'agit de la fusion des modalités. La section 2.3, présente différentes approches de fusion et leur application dans les domaines du traitement automatique de documents vidéos ou du TAL. Nous avons pu constater que cette étape est la plus difficile à cause de l'influence de chacun des choix pris pour sa mise en œuvre, chacun possédant leurs avantages et inconvénients.

Le travail décrit dans le chapitre 3 a pour but de proposer une méthode pour la compréhension multimodale de documents vidéos. Nous avons proposé dans la section 3.3 une approche pour l'apprentissage de représentations multimodales. Basée sur l'utilisation de DNN, l'avantage majeur de cette approche est qu'elle permet d'apprendre des représentations pour chacune des modalités sur de grands corpus de données monomodales, puis d'ajuster ces représentations en fonction d'une tâche pour obtenir des *embeddings* multimodaux. Nous avons étudié ces représentations dans la section 3.4 en les utilisant pour la tâche de SRR dans le corpus de vidéos journalistiques REPERE. Ces représentations ont permis de capturer les caractéristiques du rôle du locuteur sous plusieurs points de

vue. Cependant, une fois ajustées à un ensemble de données, il semble que ces représentations aient de faibles capacités de généralisation.

La suite de ce document se focalise davantage sur le traitement joint texte/image. Pour la tâche de [SRR](#), l'analyse des différentes modalités est indépendante car ce sont les caractéristiques propres à chaque modalité que nous souhaitons utiliser pour la compréhension multimodale de la vidéo. Cependant, la compréhension d'une modalité peut être utilisée pour aider à l'analyse d'une autre. Le chapitre [4](#) pose la problématique de l'alignement des modalités qui est alors requis pour cette configuration. Le corpus Flickr30k Entities dispose d'une annotation fournissant cet alignement pour l'apprentissage de modèle joint texte/image. Nous présentons une méthode permettant l'alignement automatique des différentes entités d'une image et sa légende afin d'obtenir un lien entre chaque groupe nominal et les pixels de l'image les représentant. Les taux d'erreurs obtenus par ce système, montrent qu'il s'agit d'une tâche difficile et qu'elle peut représenter un frein à la mise en œuvre de la compréhension multimodale pour des tâches de [TAL](#).

L'application de la compréhension multimodale pour une tâche de [TAL](#) est par la suite étudiée dans le chapitre [5](#). Nous proposons une approche permettant de corriger l'analyse d'une description d'image en utilisant l'information visuelle de la scène. Nous avons proposé une méthode de correction de rattachements prépositionnels pour l'analyse syntaxique en dépendance des descriptions d'images. La section [5.2](#) décrit le processus d'annotation réalisé pour adapter le corpus Flickr30k Entities pour son utilisation en [TAL](#). Les expériences réalisées sur ce corpus montrent que l'information visuelle a pu être correctement exploitée pour aider à la désambiguïsation des rattachements. Cependant il semble d'une part que l'information nécessaire pour la tâche soit difficile à extraire, et d'autre part que les erreurs d'alignement jouent un rôle essentiel dans les performances du système.

6.2 Perspectives

Une des problématiques majeures de la compréhension multimodale est que les techniques actuelles pour la représentation des données se trouvent être la plupart du temps des approches par réseaux de neurones. L'ensemble des travaux mettant en œuvre des processus de fusion ainsi que ceux présentés dans ce document, montrent clairement que l'information est là et que les systèmes sont capables de l'identifier et de l'utiliser afin d'améliorer leurs performances.

En ce qui concerne l'apprentissage de représentations multimodales, nous avons vu que leur capacité de généralisation est relativement faible. Ainsi, il est encore difficile pour l'instant d'imaginer des *embeddings* génériques de documents multimodaux, de la même façon que nous apprenons des *embeddings* de mots.

Cependant, dans l'espace de représentation des *word embeddings*, deux mots seront proches si leur sens est proche. La création de telles représentations pour des documents multimodaux doit ainsi passer par l'identification de cette notion de rapprochement, en fonction de laquelle deux contenus différents seront jugés proches ou non. Pour ce faire, il se peut que l'utilisation d'un mécanisme d'attention sur ces *embeddings* multimodaux améliore la modélisation de l'information dans les différentes modalités.

Nous avons abordé la compréhension multimodale pour le TAL et notamment l'importance de l'alignement des modalités dans ce domaine. Nos travaux ont montré que cet alignement représente une réelle difficulté pour la mise en œuvre de traitements joints texte/image. Une première approche pour résoudre cette problématique serait de conditionner la détection des entités de l'image en fonction des groupes nominaux identifiés dans la phrase.

Dans le cadre de la tâche de correction des rattachements prépositionnels nous pouvons aussi envisager un système complet qui laisserait un mécanisme d'attention mettre en avant les régions pertinentes pour la désambiguïsation. Il s'agirait de créer une distribution d'attention sur des paires de régions d'images correspondant aux groupes nominaux du gouverneur et du dépendant. Q. CHEN, ZHU, LING et al. (2018) proposent un modèle qui réalise de l'attention sur des paires de mots et l'augmente avec des relations de synonymie issues de WordNet. Un tel modèle pourrait alors être utilisé pour ajouter des caractéristiques modélisant les relations spatiales entre les paires de régions d'images.

Plus généralement, il serait intéressant d'étudier les approches proposées dans ce document pour d'autres tâches de TAL comme la traduction. Les travaux sur la traduction multimodale présentés par CAGLAYAN, ARANSA, Y. WANG et al. (2016) sont une première étape dans ce sens. Ce domaine regroupe aussi l'ensemble des travaux sur la génération automatique de description d'images qui est la traduction d'une image dans une langue naturelle.

Bibliographie

- AGIRRE, Eneko, Timothy BALDWIN et David MARTINEZ (2008). « Improving Parsing and PP Attachment Performance with Sense Information ». In : *Proceedings of ACL-08 : HLT*. Columbus, Ohio : Association for Computational Linguistics, p. 317-325 (cf. p. 86).
- ANGUIANO, Enrique Henestroza et Marie CANDITO (2011). « Parse correction with specialized models for difficult attachment types ». In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, p. 1222-1233 (cf. p. 91, 95).
- ATREY, Pradeep K, M Anwar HOSSAIN, Abdulmotaleb EL SADDIK et Mohan S KANKANHALLI (2010). « Multimodal fusion for multimedia analysis : a survey ». In : *Multimedia systems* 16.6, p. 345-379 (cf. p. 42).
- ATTARDI, Giuseppe et Massimiliano CIARAMITA (2007). « Tree revision learning for dependency parsing ». In : *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, p. 388-395 (cf. p. 92).
- BACH, Francis R et Michael I JORDAN (2002). « Kernel independent component analysis ». In : *Journal of machine learning research* 3.Jul, p. 1-48 (cf. p. 68).
- BAHDANAU, Dzmitry, Kyunghyun CHO et Yoshua BENGIO (2014). « Neural machine translation by jointly learning to align and translate ». In : *arXiv preprint arXiv :1409.0473* (cf. p. 40, 45).
- BAKER, L Douglas et Andrew Kachites MCCALLUM (1998). « Distributional clustering of words for text classification ». In : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 96-103 (cf. p. 37).
- BARONI, Marco, Georgiana DINU et Germán KRUSZEWSKI (2014). « Don't count, predict A systematic comparison of context-counting vs. context-predicting semantic vectors ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. T. 1, p. 238-247 (cf. p. 38).
- BECHET, Frederic, Meriem BENDRIS, Delphine CHARLET, Geraldine DAMNATI, Benoit FAVRE, Mickael ROUVIER, Remi AUGUSTE, Benjamin BIGOT, Richard DUFOUR, Corinne FREDOUILLE, Georges LINARES, Jean MARTINET, Gregory SENAY et Pierre TIRILLY (2015). « Identification de personnes dans des flux multimédia ». In : *CORIA* (cf. p. 54, 55).
- BELINKOV, Yonatan, Tao LEI, Regina BARZILAY et Amir GLOBERSON (2014). « Exploring compositional architectures and word vector representations for

- prepositional phrase attachment ». In : *Transactions of the Association for Computational Linguistics* 2, p. 561-572 (cf. p. 88).
- BENDRIS, Meriem, Delphine CHARLET et Gerard CHOLLET (2010). « Talking faces indexing in TV-content ». In : *Content-Based Multimedia Indexing (CBMI), 2010 International Workshop on*. IEEE, p. 1-6 (cf. p. 26, 59).
- BENGIO, Yoshua, Réjean DUCHARME, Pascal VINCENT et Christian JAUVIN (2003). « A neural probabilistic language model ». In : *Journal of machine learning research* 3.Feb, p. 1137-1155 (cf. p. 37).
- BENGIO, Yoshua, Patrice SIMARD et Paolo FRASCONI (1994). « Learning long-term dependencies with gradient descent is difficult ». In : *IEEE transactions on neural networks* 5.2, p. 157-166 (cf. p. 39).
- BIGOT, Benjamin, Isabelle FERRANÉ, Julien PINQUIER et Régine ANDRÉ-OBRECHT (2010). « Speaker Role Recognition to help Spontaneous Conversational Speech Detection ». In : *SCSS* (cf. p. 52).
- BOJANOWSKI, Piotr, Francis BACH, Ivan LAPTEV, Jean PONCE, Cordelia SCHMID et Josef SIVIC (2013). « Finding actors and actions in movies ». In : *Proceedings of the IEEE International Conference on Computer Vision*, p. 2280-2287 (cf. p. 25).
- BOJAR, Ondrej, Christian BUCK, Christian FEDERMANN, Barry HADDOW, Philipp KOEHN, Johannes LEVELING, Christof MONZ, Pavel PECINA, Matt POST, Herve SAINT-AMAND, Radu SORICUT, Lucia SPECIA et Aleš TAMCHYNA (2014). « Findings of the 2014 Workshop on Statistical Machine Translation ». In : *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA : Association for Computational Linguistics, p. 12-58 (cf. p. 45).
- BOSER, Bernhard E, Isabelle M GUYON et Vladimir N VAPNIK (1992). « A training algorithm for optimal margin classifiers ». In : *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, p. 144-152 (cf. p. 43).
- BREDIN, Hervé, Johann POIGNANT, Makarand TAPASWI, Guillaume FORTIER, Viet Bac LE, Thibault NAPOLEON, Hua GAO, Claude BARRAS, Sophie ROSSET, Laurent BESACIER, Jakob VERBEEK, Georges QUÉNOT, Frédéric JURIE et Hazim Kemal EKENEL (2012). « Fusion of speech, faces and text for person identification in TV broadcast ». In : *European Conference on Computer Vision*. Springer, p. 385-394 (cf. p. 24).
- BROWN, Peter F, Peter V DESOUZA, Robert L MERCER, Vincent J Della PIETRA et Jenifer C LAI (1992). « Class-based n-gram models of natural language ». In : *Computational linguistics* 18.4, p. 467-479 (cf. p. 37).
- CAGLAYAN, Ozan, Walid ARANSA, Yaxing WANG, Marc MASANA, Mercedes GARCÍA-MARTÍNEZ, Fethi BOUGARES, Loïc BARRAULT et Joost VAN DE WEIJER (2016). « Does multimodality help human and machine for translation and image captioning? » In : *arXiv preprint arXiv :1605.09186* (cf. p. 46, 103).
- CHEN, Qian, Xiaodan ZHU, Zhen-Hua LING, Diana INKPEN et Si WEI (2018). « Neural natural language inference models enhanced with external knowledge ».

-
- In : *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. T. 1, p. 2406-2417 (cf. p. [103](#)).
- CHEN, Xinlei, Hao FANG, Tsung-Yi LIN, Ramakrishna VEDANTAM, Saurabh GUPTA, Piotr DOLLÁR et C Lawrence ZITNICK (2015). « Microsoft COCO captions : Data collection and evaluation server ». In : *arXiv preprint arXiv :1504.00325* (cf. p. [69](#)).
- CHO, Kyunghyun, Bart VAN MERRIËNBOER, Caglar GULCEHRE, Dzmitry BAH-DANAU, Fethi BOUGARES, Holger SCHWENK et Yoshua BENGIO (2014). « Learning phrase representations using RNN encoder-decoder for statistical machine translation ». In : *arXiv preprint arXiv :1406.1078* (cf. p. [39](#), [45](#)).
- CHRISTIE, Gordon, Ankit LADDHA, Aishwarya AGRAWAL, Stanislaw ANTOL, Yash GOYAL, Kevin KOCHERSBERGER et Dhruv BATRA (2016). « Resolving Language and Vision Ambiguities Together : Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes ». In : *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1493-1503 (cf. p. [88](#)).
- COLLOBERT, Ronan, Jason WESTON, Léon BOTTOU, Michael KARLEN, Koray KAVUKCUOGLU et Pavel KUKSA (2011). « Natural language processing (almost) from scratch ». In : *Journal of Machine Learning Research* 12.Aug, p. 2493-2537 (cf. p. [37](#), [38](#)).
- DALAL, Navneet, Bill TRIGGS et Cordelia SCHMID (2006). « Human detection using oriented histograms of flow and appearance ». In : *European conference on computer vision*. Springer, p. 428-441 (cf. p. [32](#)).
- DAMNATI, Geraldine et Delphine CHARLET (2011). « Multi-view approach for speaker turn role labeling in TV Broadcast News shows ». In : *Twelfth Annual Conference of the International Speech Communication Association* (cf. p. [25](#), [53](#)).
- DASIGI, Pradeep, Waleed AMMAR, Chris DYER et Eduard HOVY (2017). « Ontology-Aware Token Embeddings for Prepositional Phrase Attachment ». In : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. T. 1, p. 2089-2098 (cf. p. [88](#)).
- DEERWESTER, Scott, Susan T DUMAIS, George W FURNAS, Thomas K LANDAUER et Richard HARSHMAN (1990). « Indexing by latent semantic analysis ». In : *Journal of the American society for information science* 41.6, p. 391-407 (cf. p. [37](#)).
- DENG, Jia, Wei DONG, Richard SOCHER, Li-Jia LI, Kai LI et Li FEI-FEI (2009). « Imagenet : A large-scale hierarchical image database ». In : *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, p. 248-255 (cf. p. [33](#), [60](#)).
- DUFOUR, Richard, Yannick ESTEVE et Paul DELÉGLISE (2011). « Investigation of spontaneous speech characterization applied to speaker role recognition ». In : *Interspeech* (cf. p. [52](#)).

- DUMONT, Emilie et Georges QUÉNOT (2012). « Automatic story segmentation for tv news video using multiple modalities ». In : *International journal of digital multimedia broadcasting* 2012 (cf. p. 25).
- ELMAN, Jeffrey L (1990). « Finding structure in time ». In : *Cognitive science* 14.2, p. 179-211 (cf. p. 38).
- FAGHRI, Fartash, David J. FLEET, Ryan KIROS et Sanja FIDLER (2017). « VSE++ : Improved Visual-Semantic Embeddings ». In : *CoRR* abs/1707.05612 (cf. p. 76, 77).
- FAVRE, Benoit, Geraldine DAMNATI, Frederic BECHET, Meriem BENDRIS, Delphine CHARLET, Remi AUGUSTE, Stephane AYACHE, Benjamin BIGOT, Alexandre DELTEI, Richard DUFOUR, Corinne FREDOUILLE, Georges LINARES, Jean MARTINET, Gregory SENAY et Pierre TIRILLY (2013). « PERCOLI : a person identification system for the 2013 REPERE challenge ». In : *First Workshop on Speech, Language and Audio in Multimedia* (cf. p. 24).
- FAVRE, Benoit, Dilek HAKKANI-TÜR et Sebastien CUENDET (2007). *Icsiboost*. <https://github.com/benob/icsiboost.git> (cf. p. 53, 62, 91).
- FENG, Bailan, Jinfeng BAI, Zhineng CHEN, Xiangsheng HUANG et Bo XU (2014). « Anchor Shot Detection with Deep Neural Network ». In : *PCM* (cf. p. 53).
- FIRAT, Orhan, Kyunghyun CHO et Yoshua BENGIO (2016). « Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism ». In : *Proceedings of NAACL-HLT*, p. 866-875 (cf. p. 46).
- FREUND, Yoav et Robert E SCHAPIRE (1996). « Experiments with a new boosting algorithm ». In : *ICML*. T. 96. Citeseer, p. 148-156 (cf. p. 53).
- FREUND, Yoav, Robert SCHAPIRE et Naoki ABE (1999). « A short introduction to boosting ». In : *Journal-Japanese Society For Artificial Intelligence* 14.771-780, p. 1612 (cf. p. 91).
- GAIDON, Adrien, Zaid HARCHAOUI et Cordelia SCHMID (2013). « Temporal localization of actions with actoms ». In : *IEEE transactions on pattern analysis and machine intelligence*, p. 1 (cf. p. 25).
- GALIBERT, Olivier et Juliette KAHN (2013). « The first official repere evaluation ». In : *First Workshop on Speech, Language and Audio in Multimedia* (cf. p. 24).
- GALLIANO, Sylvain, Edouard GEOFFROIS, Guillaume GRAVIER, Jean-François BONASTRE, Djamel MOSTEFA et Khalid CHOUKRI (2006). « Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news ». In : *Proceedings of LREC*. T. 6, p. 315-320 (cf. p. 50).
- GIRAUDEL, Aude, Matthieu CARRÉ, Valérie MAPELLI, Juliette KAHN, Olivier GALIBERT et Ludovic QUINTARD (2012). « The REPERE Corpus : a multimodal corpus for person recognition ». In : *LREC* (cf. p. 57).
- GIRSHICK, Ross, Jeff DONAHUE, Trevor DARRELL et Jitendra MALIK (2014). « Rich feature hierarchies for accurate object detection and semantic segmentation ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 580-587 (cf. p. 74).

-
- GÖNEN, Mehmet et Ethem ALPAYDIN (2011). « Multiple kernel learning algorithms ». In : *Journal of machine learning research* 12.Jul, p. 2211-2268 (cf. p. 43).
- GUINAUDEAU, Camille, Guillaume GRAVIER et Pascale SÉBILLOT (2012). « Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation ». In : *Computer Speech & Language* 26.2, p. 90-104 (cf. p. 25).
- HALL, Keith et Václav NOVÁK (2005). « Corrective modeling for non-projective dependency parsing ». In : *Proceedings of the Ninth International Workshop on Parsing Technology*. Association for Computational Linguistics, p. 42-52 (cf. p. 92).
- HARDOON, David R, Sandor SZEDMAK et John SHAWE-TAYLOR (2004). « Canonical correlation analysis : An overview with application to learning methods ». In : *Neural computation* 16.12, p. 2639-2664 (cf. p. 68).
- HARRIS, Zellig S (1954). « Distributional structure ». In : *Word* 10.2-3, p. 146-162 (cf. p. 36).
- HE, Kaiming, Xiangyu ZHANG, Shaoqing REN et Jian SUN (2016). « Deep residual learning for image recognition ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770-778 (cf. p. 35, 77).
- HOCHREITER, Sepp, Yoshua BENGIO, Paolo FRASCONI et Jürgen SCHMIDHUBER (2001). *Gradient flow in recurrent nets : the difficulty of learning long-term dependencies* (cf. p. 39).
- HOCHREITER, Sepp et Jürgen SCHMIDHUBER (1997). « Long short-term memory ». In : *Neural computation* 9.8, p. 1735-1780 (cf. p. 39).
- HODOSH, Micah, Peter YOUNG et Julia HOCKENMAIER (2013). « Framing image description as a ranking task : Data, models and evaluation metrics ». In : *Journal of Artificial Intelligence Research* 47, p. 853-899 (cf. p. 68).
- HUANG, Gao, Zhuang LIU, Laurens VAN DER MAATEN et Kilian Q WEINBERGER (2017). « Densely Connected Convolutional Networks. » In : *CVPR*. T. 1. 2, p. 3 (cf. p. 35).
- HUBEL, David H et Torsten N WIESEL (1968). « Receptive fields and functional architecture of monkey striate cortex ». In : *The Journal of physiology* 195.1, p. 215-243 (cf. p. 32).
- HUTCHINSON, Brian, Bin ZHANG et Mari OSTENDORF (2010). « Unsupervised broadcast conversation speaker role labeling ». In : *ICASSP* (cf. p. 52).
- IYENGAR, Giridharan, Harriet J NOCK et Chalapathy NETI (2003). « Audio-visual synchrony for detection of monologues in video archives ». In : *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. T. 1. IEEE, p. I-329 (cf. p. 43).
- JAIN, Anil, Karthik NANDAKUMAR et Arun ROSS (2005). « Score normalization in multimodal biometric systems ». In : *Pattern recognition* 38.12, p. 2270-2285 (cf. p. 43).
- JIA, Yangqing, Evan SHELHAMER, Jeff DONAHUE, Sergey KARAYEV, Jonathan LONG, Ross GIRSHICK, Sergio GUADARRAMA et Trevor DARRELL (2014).

- « Caffe : Convolutional architecture for fast feature embedding ». In : *ICM* (cf. p. 60).
- KENNY, Patrick, Gilles BOULIANNE, Pierre OUELLET et Pierre DUMOUCHEL (2005). « Factor Analysis Simplified. » In : *ICASSP* (cf. p. 62).
- KINGMA, Diederik P et Jimmy BA (2014). « Adam : A method for stochastic optimization ». In : *arXiv preprint arXiv :1412.6980* (cf. p. 77).
- KLEIN, Benjamin, Guy LEV, Gil SADEH et Lior WOLF (2014). « Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation ». In : *arXiv preprint arXiv :1411.7399* (cf. p. 71).
- KOK, Daniël de et Erhard W. HINRICHS (2016). « Transition-based dependency parsing with topological fields ». In : *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 2 : Short Papers* (cf. p. 82).
- KOK, Daniël de, Jianqiang MA, Corina DIMA et Erhard HINRICHS (2017). « PP Attachment : Where do We Stand ? » In : *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*. T. 2, p. 311-317 (cf. p. 88).
- KRIZHEVSKY, Alex, Ilya SUTSKEVER et Geoffrey E. HINTON (2012). « ImageNet Classification with Deep Convolutional Neural Network ». In : *NIPS* (cf. p. 33, 35, 55, 60).
- KUHN, Harold W (1955). « The Hungarian method for the assignment problem ». In : *Naval research logistics quarterly* 2.1-2, p. 83-97 (cf. p. 76).
- KUMMERFELD, Jonathan K., David Leo Wright HALL, James R. CURRAN et Dan KLEIN (2012). « Parser Showdown at the Wall Street Corral : An Empirical Investigation of Error Types in Parser Output ». In : *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012*, p. 1048-1059 (cf. p. 82).
- LANCKRIET, Gert RG, Nello CRISTIANINI, Peter BARTLETT, Laurent El GHAOUI et Michael I JORDAN (2004). « Learning the kernel matrix with semidefinite programming ». In : *Journal of Machine learning research* 5.Jan, p. 27-72 (cf. p. 43).
- LECUN, Yann, Léon BOTTOU, Yoshua BENGIO et Patrick HAFFNER (1998). « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11, p. 2278-2324 (cf. p. 32, 35).
- LIN, Tsung-Yi, Michael MAIRE, Serge BELONGIE, James HAYS, Pietro PERONA, Deva RAMANAN, Piotr DOLLÁR et C Lawrence ZITNICK (2014). « Microsoft coco : Common objects in context ». In : *European conference on computer vision*. Springer, p. 740-755 (cf. p. 69, 78).
- LIU, Yang (2006). « Initial Study on Automatic Identification of Speaker Role in Broadcast News Speech ». In : *NAACL* (cf. p. 53).
- LOWE, David G (1999). « Object recognition from local scale-invariant features ». In : *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. T. 2. Ieee, p. 1150-1157 (cf. p. 32).

-
- MARCUS, Mitchell P, Mary Ann MARCINKIEWICZ et Beatrice SANTORINI (1993). « Building a large annotated corpus of English : The Penn Treebank ». In : *Computational linguistics* 19.2, p. 313-330 (cf. p. 83).
- MCGURK, Harry et John MACDONALD (1976). « Hearing lips and seeing voices ». In : *Nature* 264.5588, p. 746 (cf. p. 29).
- MIKOLOV, Tomas, Kai CHEN, Greg CORRADO et Jeffrey DEAN (2013). « Efficient Estimation of Word Representations in Vector Space ». In : *CoRR* (cf. p. 37).
- MILLER, George A (1995). « WordNet : a lexical database for English ». In : *Communications of the ACM* 38.11, p. 39-41 (cf. p. 36, 86, 88).
- NASR, Alexis, Frédéric BÉCHET, Jean-François REY, Benoît FAVRE et Joseph LE ROUX (2011). « Macaon : An nlp tool suite for processing word lattices ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Systems Demonstrations*. Association for Computational Linguistics, p. 86-91 (cf. p. 83).
- NEVEROVA, Natalia, Christian WOLF, Graham TAYLOR et Florian NEBOUT (2016). « Moddrop : adaptive multi-modal gesture recognition ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8, p. 1692-1706 (cf. p. 44).
- NGIAM, Jiquan, Aditya KHOSLA, Mingyu KIM, Juhan NAM, Honglak LEE et Andrew Y. NG (2011). « Multimodal Deep Learning ». In : *ICML* (cf. p. 44).
- PANG, Yanwei, Yuan YUAN, Xulong LI et Jing PAN (2011). « Efficient HOG human detection ». In : *Signal Processing* 91.4, p. 773-781 (cf. p. 62).
- PAPINENI, Kishore, Salim ROUKOS, Todd WARD et Wei-Jing ZHU (2002). « BLEU : a method for automatic evaluation of machine translation ». In : *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, p. 311-318 (cf. p. 45).
- PENNINGTON, Jeffrey, Richard SOCHER et Christopher MANNING (2014). « Glove : Global vectors for word representation ». In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532-1543 (cf. p. 37).
- PEREIRA, Fernando, Naftali TISHBY et Lillian LEE (1993). « Distributional clustering of English words ». In : *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 183-190 (cf. p. 37).
- PEYRE, Julia, Ivan LAPTEV, Cordelia SCHMID et Josef SIVIC (2017). « Weakly-Supervised Learning of Visual Relations ». In : *IEEE International Conference on Computer Vision, ICCV 2017*, p. 5189-5198 (cf. p. 89, 91).
- PLUMMER, Bryan A, Liwei WANG, Chris M CERVANTES, Juan C CAICEDO, Julia HOCKENMAIER et Svetlana LAZEBNIK (2017). « Flickr30k Entities : Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models ». In : *International Journal of Computer Vision* 123.1, p. 74-93 (cf. p. 69).
- PORIA, Soujanya, Erik CAMBRIA et Alexander GELBUKH (2015). « Deep convolutional neural network textual features and multiple kernel learning for utterance-

- level multimodal sentiment analysis ». In : *Proceedings of the 2015 conference on empirical methods in natural language processing*, p. 2539-2544 (cf. p. 43).
- POTAMIANOS, Gerasimos, Chalapathy NETI, Guillaume GRAVIER, Ashutosh GARG et Andrew W SENIOR (2003). « Recent advances in the automatic recognition of audiovisual speech ». In : *Proceedings of the IEEE* 91.9, p. 1306-1326 (cf. p. 42).
- POVEY, Daniel, Arnab GHOSHAL, Gilles BOULIANNE, Lukas BURGET, Ondrej GLEMBEK, Nagendra GOEL, Mirko HANNEMANN, Petr MOTLICEK, Yanmin QIAN, Petr SCHWARZ, Jan SILOVSKY, Georg STEMMER et Karel VESELY (2011). « The Kaldi Speech Recognition Toolkit ». In : *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No. : CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society (cf. p. 58, 61).
- RABINER, Lawrence R (1989). « A tutorial on hidden Markov models and selected applications in speech recognition ». In : *Proceedings of the IEEE* 77.2, p. 257-286 (cf. p. 53).
- RAKSHIT, Geetanjali, Sagar SONTAKKE, Pushpak BHATTACHARYYA et Gholamreza HAFFARI (2016). « Prepositional Attachment Disambiguation Using Bilingual Parsing and Alignments ». In : *arXiv preprint arXiv :1603.08594* (cf. p. 88).
- REDMON, Joseph, Santosh DIVVALA, Ross GIRSHICK et Ali FARHADI (2016). « You only look once : Unified, real-time object detection ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 779-788 (cf. p. 74).
- REDMON, Joseph et Ali FARHADI (2017). « YOLO9000 : Better, Faster, Stronger ». In : *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, p. 6517-6525 (cf. p. 74).
- REYNOLDS, Douglas (2015). « Gaussian mixture models ». In : *Encyclopedia of biometrics*, p. 827-832 (cf. p. 52).
- ROUVIER, Mickael, Pierre-Michel BOUSQUET et Benoit FAVRE (2015). « Speaker Diarization Through Speaker Embeddings ». In : *EUSIPCO* (cf. p. 56).
- ROUVIER, Mickael, Gregor DUPUY, Paul GAY, Elie KHOURY, Teva MERLIN et Sylvain MEIGNIER (2013). « An Open-source State-of-the-art Toolbox for Broadcast News Diarization. » In : *InterSpeech* (cf. p. 58).
- RUSSAKOVSKY, Olga, Jia DENG, Hao SU, Jonathan KRAUSE, Sanjeev SATHEESH, Sean MA, Zhiheng HUANG, Andrej KARPATHY, Aditya KHOSLA, Michael BERNSTEIN, Alexander C. BERG et Li FEI-FEI (2015). « ImageNet Large Scale Visual Recognition Challenge ». In : *International Journal of Computer Vision (IJCV)* 115.3, p. 211-252 (cf. p. 32, 77).
- SAGOT, Benoît et Darja FIŠER (2008). « Building a free French wordnet from multilingual resources ». In : *OntoLex* (cf. p. 37).
- SCHWENK, Holger (2007). « Continuous space language models ». In : *Computer Speech & Language* 21.3, p. 492-518 (cf. p. 37).

-
- SIMONYAN, Karen et Andrew ZISSERMAN (2014). « Very deep convolutional networks for large-scale image recognition ». In : *arXiv preprint arXiv :1409.1556* (cf. p. 35).
- SPECIA, Lucia, Stella FRANK, Khalil SIMA'AN et Desmond ELLIOTT (2016). « A shared task on multimodal machine translation and crosslingual image description ». In : *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*. T. 2, p. 543-553 (cf. p. 24, 46).
- SRIVASTAVA, Nitish et Ruslan R SALAKHUTDINOV (2012). « Multimodal learning with deep boltzmann machines ». In : *Advances in neural information processing systems*, p. 2222-2230 (cf. p. 44).
- SUTSKEVER, Ilya, Oriol VINYALS et Quoc V LE (2014). « Sequence to sequence learning with neural networks ». In : *Advances in neural information processing systems*, p. 3104-3112 (cf. p. 45).
- SZEGEDY, Christian, Vincent VANHOUCKE, Sergey IOFFE, Jon SHLENS et Zbigniew WOJNA (2016). « Rethinking the inception architecture for computer vision ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 2818-2826 (cf. p. 35).
- UIJLINGS, Jasper RR, Koen EA VAN DE SANDE, Theo GEVERS et Arnold WM SMEULDERS (2013). « Selective search for object recognition ». In : *International journal of computer vision* 104.2, p. 154-171 (cf. p. 74).
- VERLINDE, Patrick, Gerard CHOLLET et Marc ACHEROY (2000). « Multi-modal identity verification using expert fusion ». In : *Information fusion* 1.1, p. 17-33 (cf. p. 43).
- VINYALS, Oriol, Alexander TOSHEV, Samy BENGIO et Dumitru ERHAN (2015). « Show and tell : A neural image caption generator ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3156-3164 (cf. p. 45, 46).
- WANG, Wen, Sibel YAMAN, Kristin PRECODA et Colleen RICHEY (2011). « Automatic identification of speaker role and agreement/disagreement in broadcast conversation ». In : *ICASSP* (cf. p. 52).
- WANG, Xiaoxuan, Lei XIE, Mimi LU, Bin MA, Eng Siong CHNG et Haizhou LI (2012). « Broadcast news story segmentation using conditional random fields and multimodal features ». In : *IEICE TRANSACTIONS on Information and Systems* 95.5, p. 1206-1215 (cf. p. 25).
- WIKIMEDIA, Commons (2018a). *File :Gated Recurrent Unit.svg* — *Wikimedia Commons, the free media repository*. [Online ; accessed 9-October-2018]. URL : https://commons.wikimedia.org/w/index.php?title=File:Gated_Recurrent_Unit.svg&oldid=277862579 (cf. p. 40).
- (2018b). *File :Long Short-Term Memory.svg* — *Wikimedia Commons*. [Online ; accessed 9-octobre-2018]. URL : https://commons.wikimedia.org/w/index.php?title=File:Long_Short-Term_Memory.svg&oldid=281467835 (cf. p. 40).
- XU, Kelvin, Jimmy BA, Ryan KIROS, Kyunghyun CHO, Aaron COURVILLE, Ruslan SALAKHUDINOV, Rich ZEMEL et Yoshua BENGIO (2015). « Show,

- attend and tell : Neural image caption generation with visual attention ». In : *International conference on machine learning*, p. 2048-2057 (cf. p. 45).
- YEH, Yi-Ren, Ting-Chu LIN, Yung-Yu CHUNG et Yu-Chiang Frank WANG (2012). « A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection ». In : *IEEE Transactions on multimedia* 14.3, p. 563-574 (cf. p. 43).
- YOUNG, Peter, Alice LAI, Micah HODOSH et Julia HOCKENMAIER (2014). « From image descriptions to visual denotations : New similarity metrics for semantic inference over event descriptions ». In : *Transactions of the Association for Computational Linguistics* 2, p. 67-78 (cf. p. 69).

Annexes

A Statistiques complètes sur le corpus annoté manuellement

TABLE A.1 – Statistiques sur les rattachements des prépositions annotées et leur nombre d’occurrences : les colonnes *Nom* et *Verbe* correspondant respectivement au taux de rattachement à un nom ou un verbe.

Prépositions	Occ.	Nom	Verbe	Distance
<i>in</i>	4 191	0,59	0,41	2,21
<i>with</i>	3 018	0,59	0,41	2,47
<i>for</i>	1 777	0,36	0,64	1,57
<i>near</i>	1 452	0,65	0,35	1,69
<i>through</i>	1 420	0,05	0,95	2,01
<i>on</i>	1 359	0,26	0,74	2,03
<i>next to</i>	1 342	0,08	0,92	3,23
<i>from</i>	1 172	0,30	0,70	2,43
<i>into</i>	1 123	0,08	0,92	2,32
<i>over</i>	941	0,39	0,61	2,38
<i>by</i>	890	0,10	0,90	2,60
<i>at</i>	720	0,20	0,80	2,75
<i>of</i>	700	0,97	0,03	1,03
<i>around</i>	589	0,18	0,82	2,26
<i>in front of</i>	570	0,11	0,89	3,72
<i>under</i>	544	0,18	0,82	3,29
<i>behind</i>	544	0,35	0,65	1,78
<i>along</i>	500	0,37	0,63	1,79
<i>during</i>	423	0,14	0,86	5,08
<i>across</i>	415	0,11	0,89	2,21
<i>down</i>	393	0,66	0,34	2,53
<i>against</i>	365	0,39	0,61	1,56
<i>outside</i>	356	0,38	0,62	2,28
<i>towards</i>	276	0,08	0,92	2,41
<i>out of</i>	252	0,13	0,87	2,08
<i>beside</i>	245	0,03	0,97	3,51
<i>above</i>	241	0,43	0,57	2,90
<i>in the middle of</i>	240	0,12	0,88	3,63
<i>onto</i>	210	0,08	0,92	2,34
<i>outside of</i>	206	0,12	0,88	3,30
<i>inside</i>	197	0,13	0,87	3,72
<i>between</i>	189	0,72	0,28	1,29
<i>past</i>	170	0,19	0,81	2,95

Prépositions	Occ.	Nom	Verbe	Distance
<i>toward</i>	167	0,34	0,66	1,53
<i>on top of</i>	166	0,14	0,86	2,93
<i>like</i>	159	0,34	0,66	1,32
<i>among</i>	142	0,37	0,63	2,13
<i>after</i>	126	0,10	0,90	3,12
<i>away from</i>	109	0,04	0,96	1,75
<i>off</i>	104	0,31	0,69	3,72
<i>up</i>	96	0,82	0,18	1,93
<i>up to</i>	86	0,12	0,88	2,35
<i>before</i>	71	0,10	0,90	4,14
<i>atop</i>	60	0,23	0,77	2,97
<i>about</i>	54	0,59	0,41	1,52
<i>along with</i>	54	0,15	0,85	4,98
<i>underneath</i>	49	0,18	0,82	3,16
<i>without</i>	46	0,54	0,46	2,98
<i>out</i>	42	0,29	0,71	4,43
<i>at the top of</i>	40	0,10	0,90	2,23
<i>inside of</i>	39	0,23	0,77	3,72
<i>amongst</i>	33	0,09	0,91	3,18
<i>close to</i>	33	0,15	0,85	4,42
<i>upon</i>	31	0,13	0,87	1,71
<i>amidst</i>	28	0,29	0,71	3,18
<i>beneath</i>	26	0,12	0,88	3,54
<i>within</i>	24	0,33	0,67	3,62
<i>below</i>	23	0,52	0,48	1,83
<i>at the bottom of</i>	22	0,18	0,82	2,86
<i>amid</i>	22	0,18	0,82	5,00
<i>in between</i>	18	0,22	0,78	2,89
<i>up against</i>	16	0,06	0,94	3,69
<i>ahead of</i>	14	0,00	1,00	1,71
<i>together with</i>	13	0,08	0,92	2,54
<i>such as</i>	13	0,38	0,62	5,31
<i>besides</i>	12	0,42	0,58	2,33
<i>beyond</i>	10	0,70	0,30	3,30
<i>on the top of</i>	10	0,10	0,90	2,70
<i>while</i>	10	0,20	0,80	6,40
<i>near to</i>	10	0,10	0,90	4,10
<i>to the left of</i>	9	0,33	0,67	3,22
<i>though</i>	6	0,00	1,00	1,00

Prépositions	Occ.	Nom	Verbe	Distance
<i>aboard</i>	4	0,00	1,00	3,25
<i>astride</i>	4	0,00	1,00	1,75
<i>via</i>	4	0,00	1,00	5,00
<i>opposite</i>	3	0,00	1,00	1,00
<i>despite</i>	3	0,00	1,00	5,33
<i>in line with</i>	2	0,00	1,00	2,50
<i>instead of</i>	2	0,00	1,00	6,00
<i>on the bottom of</i>	2	0,00	1,00	2,50
<i>throughout</i>	2	0,00	1,00	3,00
<i>except for</i>	2	0,50	0,50	3,50
<i>according to</i>	2	0,00	1,00	7,00
<i>to the right of</i>	2	0,00	1,00	4,00
<i>than</i>	2	0,50	0,50	3,00
<i>in back of</i>	2	0,00	1,00	1,50
<i>per</i>	1	1,00	0,00	1,00
<i>except</i>	1	1,00	0,00	1,00
<i>unto</i>	1	0,00	1,00	4,00
<i>far from</i>	1	0,00	1,00	4,00
<i>thanks to</i>	1	0,00	1,00	5,00
<i>in view of</i>	1	0,00	1,00	9,00
<i>because of</i>	1	1,00	0,00	1,00
<i>apart from</i>	1	0,00	1,00	1,00
Total	29 068	0,36	0,64	2,37

B Performances du système de correction de rattachements prépositionnels

TABLE B.2 – Taux de rattachements corrects sur le corpus de test : # indique le nombre d’occurrences de la préposition ; *parser* correspond au score produit par l’analyseur syntaxique ; V_C représente les concepts visuels, V_S les caractéristiques spatiales, V la combinaison des deux et L sont les caractéristiques lexicales. Le score est donné pour les prépositions ayant un nombre d’occurrences supérieur ou égal à dix. Le total est quand à lui donné sur l’ensemble du corpus de test.

Prépositions	Occ.	<i>Parser</i>	V_C	V_S	$V_c + V_S$	L	$V_c + V_S + L$
<i>in</i>	369	0,76	0,76	0,77	0,76	0,85	0,84
<i>with</i>	310	0,65	0,68	0,68	0,70	0,78	0,78
<i>for</i>	168	0,73	0,73	0,73	0,72	0,82	0,83
<i>near</i>	159	0,33	0,53	0,50	0,59	0,84	0,84
<i>through</i>	145	0,95	0,95	0,95	0,95	0,96	0,96
<i>on</i>	143	0,85	0,85	0,85	0,85	0,89	0,87
<i>from</i>	140	0,76	0,76	0,76	0,76	0,86	0,85
<i>next to</i>	137	0,89	0,89	0,89	0,89	0,90	0,89
<i>into</i>	116	0,89	0,89	0,89	0,89	0,92	0,95
<i>over</i>	111	0,66	0,64	0,66	0,68	0,85	0,84
<i>by</i>	76	0,84	0,84	0,84	0,84	0,86	0,86
<i>of</i>	72	0,93	0,93	0,93	0,93	0,93	0,93
<i>at</i>	63	0,84	0,84	0,84	0,84	0,87	0,89
<i>around</i>	59	0,73	0,73	0,73	0,71	0,85	0,80
<i>under</i>	58	0,84	0,84	0,84	0,84	0,84	0,84
<i>in front of</i>	51	0,90	0,90	0,90	0,90	0,88	0,90
<i>along</i>	50	0,52	0,60	0,54	0,58	0,88	0,86
<i>across</i>	49	0,88	0,88	0,88	0,88	0,96	0,98
<i>down</i>	41	0,63	0,59	0,56	0,56	0,73	0,76
<i>during</i>	41	0,71	0,71	0,71	0,71	0,76	0,76
<i>behind</i>	35	0,74	0,74	0,74	0,74	0,86	0,91
<i>outside</i>	35	0,63	0,63	0,69	0,69	0,71	0,74
<i>out of</i>	32	0,78	0,78	0,78	0,78	0,78	0,78
<i>against</i>	31	0,77	0,77	0,77	0,77	0,94	0,94
<i>towards</i>	30	0,90	0,90	0,90	0,90	0,93	0,93
<i>beside</i>	29	0,93	0,93	0,93	0,93	0,93	0,93
<i>above</i>	27	0,44	0,52	0,44	0,48	0,81	0,78
<i>in the middle of</i>	27	0,89	0,89	0,89	0,89	0,93	0,93
<i>like</i>	23	0,78	0,78	0,78	0,83	0,74	0,74

Prépositions	Occ.	<i>Parser</i>	V_C	V_S	$V_c + V_S$	L	$V_c + V_S + L$
<i>on top of</i>	22	0,86	0,86	0,86	0,86	0,86	0,86
<i>inside</i>	21	0,81	0,81	0,81	0,81	0,81	0,81
<i>between</i>	20	0,40	0,40	0,55	0,40	0,75	0,80
<i>outside of</i>	20	0,80	0,80	0,80	0,80	0,85	0,85
<i>toward</i>	20	0,65	0,70	0,70	0,70	0,90	0,90
<i>among</i>	19	0,79	0,84	0,63	0,79	0,84	0,84
<i>onto</i>	16	0,88	0,88	0,88	0,88	1,00	1,00
<i>up</i>	14	0,86	0,86	0,86	0,86	0,86	0,86
<i>off</i>	11	0,73	0,64	0,73	0,73	0,64	0,64
<i>after</i>	10	0,70	0,70	0,70	0,70	0,70	0,70
Total	2 907	0,75	0,77	0,77	0,77	0,85	0,85

Approches jointes texte/image pour la compréhension multimodale de documents

Résumé : Les mécanismes de compréhension chez l'être humain sont par essence multimodaux. Comprendre le monde qui l'entoure revient chez l'être humain à fusionner l'information issue de l'ensemble de ses récepteurs sensoriels. La plupart des documents utilisés en traitement automatique de l'information sont multimodaux. Par exemple, du texte et des images dans des documents textuels ou des images et du son dans des documents vidéo. Cependant, les traitements qui leurs sont appliqués sont le plus souvent monomodaux. Le but de cette thèse est de proposer des traitements joints s'appliquant principalement au texte et à l'image pour le traitement de documents multimodaux à travers deux études : l'une portant sur la fusion multimodale pour la reconnaissance du rôle du locuteur dans des émissions télévisuelles, l'autre portant sur la complémentarité des modalités pour une tâche d'analyse linguistique sur des corpus d'images avec légendes. Pour la première étude nous nous intéressons à l'analyse de documents audiovisuels provenant de chaînes d'information télévisuelle. Nous proposons une approche utilisant des réseaux de neurones profonds pour la création d'une représentation jointe multimodale pour les représentations et la fusion des modalités. Dans la seconde partie de cette thèse nous nous intéressons aux approches permettant d'utiliser plusieurs sources d'informations multimodales pour une tâche monomodale de traitement automatique du langage, afin d'étudier leur complémentarité. Nous proposons un système complet de correction de rattachements prépositionnels utilisant de l'information visuelle, entraîné sur un corpus multimodal d'images avec légendes.

Mots clés : Compréhension multimodale, Apprentissage automatique multimodal, Réseaux de neurones profonds, Traitement automatique des langues, Traitement automatique de l'image

Text/image joint approaches for multimodal understanding of documents

Abstract: The human faculties of understanding are essentially multimodal. To understand the world around them, human beings fuse the information coming from all of their sensory receptors. Most of the documents used in automatic information processing contain multimodal information, for example text and image in textual documents or image and sound in video documents, however the processings used are most often monomodal. The aim of this thesis is to propose joint processes applying mainly to text and image for the processing of multimodal documents through two studies: one on multimodal fusion for the speaker role recognition in television broadcasts, the other on the complementarity of modalities for a task of linguistic analysis on corpora of images with captions. In the first part of this study, we interested in audiovisual documents analysis from news television channels. We propose an approach that uses in particular deep neural networks for representation and fusion of modalities. In the second part of this thesis, we are interested in approaches allowing to use several sources of multimodal information for a monomodal task of natural language processing in order to study their complementarity. We propose a complete system of correction of prepositional attachments using visual information, trained on a multimodal corpus of images with captions.

Keywords: Multimodal understanding, Multimodal machine learning, Deep neural networks, Natural language processing, Image processing