Correction automatique d'attachements prépositionnels par utilisation de traits visuels

Sebastien Delecraz[1] Leonor Becerra-Bonache[2]
Benoit Favre[1] Alexis Nasr[1] Frederic Bechet[1]
(1) Aix-Marseille Univ, Université de Toulon, CNRS, LIS, UMR 7020, Marseille, France
(2) Univ Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien, UMR 5516, Saint-Étienne, France
(1) prenom.nom@univ-amu.fr, (2) leonor.becerra@univ-st-etienne.fr

RÉSUMÉ
La désambiguïsation des rattachements prépositionnels est une tâche syntaxique qui demande des connaissances sémantiques, pouvant être extraites d'une image associée au texte traité. Nous présentons et analysons les difficultés de cette tâche pour laquelle nous construisons un système complet entraîné sur une version étendue des annotations du corpus *Flickr30k Entities*. Lorsque la sémantique lexicale n'est pas disponible, l'information visuelle apporte 3 % d'amélioration.

ABSTRACT
PP-attachement resolution using visual features

Resolving prepositionnal attachments is a syntactic task that requires semantic knowledge, which can be extracted from the composition of a picture associated with a text. We present and analyse the difficulties of performing this task, for which we build a full system trained on extended annotations from *Flickr30k Entities*. When lexical semantics are unavailable, visual information brings 3% improvement.

MOTS-CLÉS :
rattachement prépositionnel, analyse syntaxique multimodale, stratégie de correction.

KEYWORDS:
PP-attachment, multimodal parsing, correction strategy.

1 Introduction

Les langues naturelles sont intrinsèquement ambiguës. Une partie de ces ambiguïtés peut être levée en utilisant des indices présents dans la phrase, mais d'autres requièrent un accès au contexte dans lequel elles ont été produites. Pour reprendre l'exemple célèbre « *Jean regard l'homme avec un télescope* », l'ambiguïté du rattachement de la préposition *pour* pourrait facilement être levée si nous avions la possibilité de voir la scène.

Nous proposons dans cet article une méthode de résolution de rattachements prépositionnels fondée sur l'utilisation d'indices visuels. Pour cela, nous utilisons un corpus constitué de paires composées d'une photo et d'une légende décrivant cette dernière. Ce corpus a été annoté manuellement à différents niveaux. Au niveau de l'image, des rectangles (que nous appellerons boîtes), ont été identifiés et une catégorie sémantique a été associée à chacune d'entre elles. Au niveau du texte,

certains groupes nominaux ont été identifiés, ainsi que certains rattachements prépositionnels, pour un sous ensemble de prépositions fréquentes. De plus les boîtes correspondant à des groupes nominaux ont été appariées à ces derniers. Le fait de disposer simultanément de l'analyse de l'image (par l'intermédiaire des boîtes), et du texte, (à travers certains rattachements prépositionnels) ainsi que l'appariement entre boîtes et groupes nominaux permet d'établir un lien entre les deux modalités et d'utiliser des informations provenant de l'image pour traiter le texte.

Le système que nous proposons 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* (*une balle face_à un chien avec collier rouge*), par exemple, correspondant à l'image reproduite en Figure 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 à *ball*, *dog* et *collar*.

La structure de l'article est la suivante. Nous dressons l'état de l'art du problème du rattachement prépositionnel dans la section 2, en se focalisant sur les études les plus pertinentes pour nos travaux. Notre corpus d'étude est présenté dans la section 3. La section 4 traite du problème de la détection des boîtes et de leur appariement avec des groupes nominaux. Le détecteur d'erreurs de rattachement est décrit dans la section 5 et la section 6 décrit les résultats obtenus sur notre corpus d'étude.

2 État de l'art

Le problème du rattachement prépositionnel a 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. Nous passons ici en revue les plus pertinentes pour nos travaux.

Deux sortes de ressources ont largement été utilisées dans la littérature pour résoudre le problème des rattachements prépositionnels : des bases de connaissances sémantiques (Agirre *et al.*, 2008; Dasigi *et al.*, 2017), et des corpus (Rakshit *et al.*, 2016; Mirroshandel & Nasr, 2016; Belinkov *et al.*, 2014; de Kok *et al.*, 2017). 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 *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 différents paires se voient attribuer un score et la paire obtenant le meilleur score est alors sélectionnée. La différence principale entre nos travaux et les leurs 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 (images/légendes) beaucoup plus important (22800 contre 1822).

De nombreux travaux se sont intéressés à la mise en correspondance d'un segment de phrase et d'une partie d'image, pour différentes sortes d'applications, tel la génération de légende (Vinyals *et al.*, 2015; Fang *et al.*, 2015; Karpathy & Fei-Fei, 2015) et la recherche d'image (Coyne & Sproat, 2001; Chang *et al.*, 2015). 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.

Nos travaux sont aussi en lien avec ceux sur l'apprentissage de relations visuelles. Les travaux les plus pertinents sont ceux de Peyre *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 (*i.e.* 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, mais les objets sont détectés et représentés d'une manière différente (en utilisant *YOLOv2* vs *Fast R-CNN*).

De nombreux chercheurs en psycholinguistique et en psychologie cognitive ont également étudié l'interaction entre la vision et le langage lors du traitement des phrases par l'humain (Spivey *et al.*, 2002; Coco & Keller, 2015). Ces travaux démontrent la pertinence de l'information visuelle pour les humains afin de résoudre l'ambiguïté linguistique. Cette information est également d'une grande importance au cours des premières étapes de l'acquisition du langage chez l'enfant, puisque la plupart des phrases reçues par les enfants sont liées à leur environnement visuel immédiat (Snow, 1972; Shaerlaekens, 1973). Même si les objectifs sont éloignés, nos travaux s'inspirent de ces idées.

3 Le corpus *Flickr30k Entities*

Il existe peu de corpus multimodaux (texte, image) qui associent des régions de l'image à des séquences de mots du texte. Dans cet article nous avons utilisé le corpus multimodal *Flickr30K Entities* (Plummer *et al.*, 2017) (*F30kE*) qui fournit ce type d'annotations et qui constitue une extension du corpus *Flickr30k* (Young *et al.*, 2014), une référence bien connue pour la description d'images par des phrases.

F30kE est composé de 32K images, chacune associée à cinq phrases la décrivant. Les chaînes de coréférences se rapportant aux mêmes entités sont annotées et liées aux boîtes englobantes des objets correspondant dans l'image (244K chaînes de coréférence et 276K boîtes sont fournies). De plus, chaque boîte est associée à une des catégories sémantiques suivantes : personnes, parties du corps, animaux, vêtements, instruments, véhicules, scène, et autres. Un exemple issu de ce corpus est reproduit dans la Figure 1.

Nous avons enrichi ce corpus avec une analyse syntaxique automatique des phrases, suivi d'une vérification manuelle des 29068 rattachements prépositionnels potentiellement ambigus du corpus. La correction de rattachement a été faite par un seul annotateur, qui avait à sa disposition uniquement la préposition cible dans la phrase et l'image.

4 Appariement automatique de boîtes et de groupes nominaux

Le modèle de correction que nous proposons suppose un appariement entre les boîtes détectées dans l'image et des groupes nominaux des légendes. Cette tâche se décompose en trois étapes : la détection des boîtes dans l'image, la détection des groupe nominaux dans la légende et, enfin, leur appariement.

FIGURE 1 – Exemple de l'annotation du corpus *F30kE*. L'image est décrite par 5 légendes, chacune annotée avec des entités. Les entités coréférentes à un élément visuel sont lié à la boîte correspondante.

Elles sont décrites successivement dans les trois sections suivantes et illustrés dans la Figure 2.

4.1 Détection des boîtes

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. Nous avons utilisé ici le modèle de détection d'objets temps-réel à base de réseau de neurones *YOLOv2* (Redmon & Farhadi, 2017) qui produit, pour une image donnée, une liste de boîtes.

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, un score de confiance pour chaque boîte, et une probabilité pour chaque catégorie. Les prédictions finales sont prises en multipliant les scores de confiance aux probabilité des catégories. Nous avons ré-entraîné le modèle *YOLOv2* sur le corpus *F30kE* en utilisant comme initialisation les poids fournis par les auteurs et en limitant le nombre de catégories aux huit catégories sémantiques du corpus *F30kE*. Seules les prédictions avec un score de confiance supérieur à 0.1 ont été retenues.

Sur les 14229 boîtes des images issues de notre corpus de test, le système en a détecté 7110 (un objet est considéré comme détecté si le score d'*Intersection over Union* (*IoU* : ratio entre l'aire de l'intersection et l'aire de l'union de deux boîtes) entre sa boîte de référence et la prédiction 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 on prend en compte les catégories sémantiques, ces 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.

4.2 Détection 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

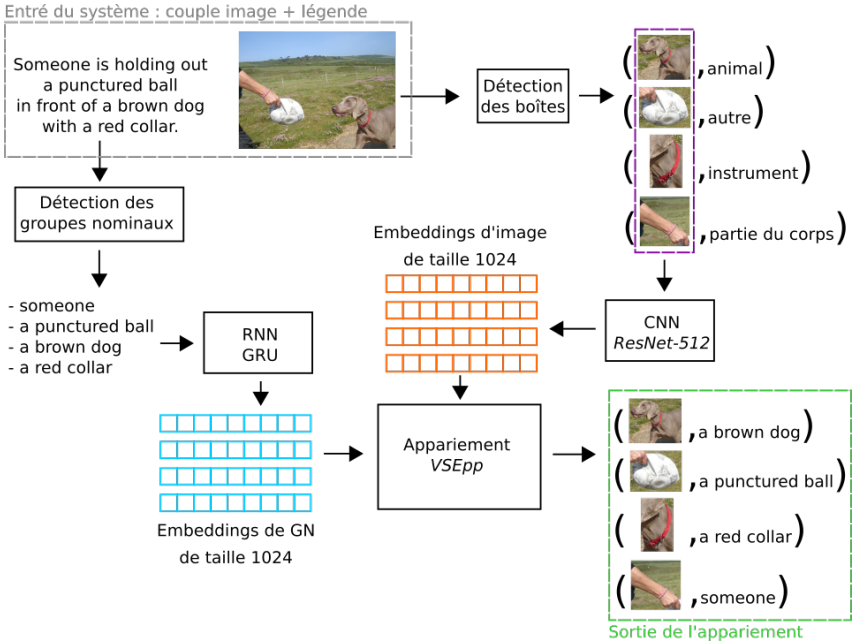

FIGURE 2 – Chaîne de traitement automatique pour l'appariement entre les boîtes des entités d'une image et les groupes nominaux de sa description.

groupes nominaux du corpus *F30kE*.

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 leurs parties de discours. Une évaluation sur notre corpus de test indique un taux d'erreur de 2.2% par mot.

4.3 Appariement

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 2, 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 mots 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'objet 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. 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. Cette projection dans un espace commun est réalisée à l'aide de réseaux de neurones. Les paramètres de ce réseau peuvent être entraînés à partir de paires connues $(\text{image}, \text{texte})$ selon la méthode décrite ci-après.

Cette méthode est fondée sur les plongements visuels sémantiques (*visual semantic embeddings*) de (Faghri *et al.*, 2017), qui tirent parti d'un réseau de neurones convolutionnel pour créer des représentations d'image et de réseaux de neurones récurrents pour créer des représentations des séquences de mots. Côté visuel, le contenu de chaque boîte est redimensionné en 224x224 pixels, puis passé en entrée à un réseau de type *ResNet-512* (He *et al.*, 2016) préentraîné sur la tâche ImageNet [1]. La dernière couche du réseau est remplacée par une couche dense (transformation linéaire) qui projette les représentations vers un vecteur de taille 1024. Côté texte, les mots des groupes nominaux sont d'abord projetés dans une couche d'*embedding* de taille 300 qui fournit des entrées à une couche récurrente de type GRU dont la représentation cachée est de taille 1024. 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. Les activations issues des réseaux de neurones pour les deux modalités sont normalisées par norme L2 et peuvent être comparées à l'aide du produit scalaire, ceci est équivalent au calcul de la similarité *cosine* entre les deux vecteurs.

L'apprentissage [2] des paramètres de ce modèle est effectué en calculant la similarité entre une paire $(\text{image}, \text{texte})$ existant dans les données d'apprentissage et une paire aléatoire avec l'un des deux membres en commun (*triplet ranking*), et en modifiant le modèle pour que le score de la paire valide soit supérieur à celui de la paire invalide (la fonction de coût utilisée est le *hinge loss*).

Méthode	**Entraînement**	**Taux d'erreur**
VSE++	F30k	42.07
VSE++	COCO	38.90
VSE	COCO	37.17
VSE	réentraîné	21.47

TABLE 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 (F30k et 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).

La Table 1 présente les performances du système d'appariement entre boîtes et groupes nominaux. Le taux d'erreur est calculé de la manière suivante : pour chaque groupe nominal, on considère que l'association est correcte si la boîte dont la similarité avec ce groupe nominal est la plus élevée est bien celle qui lui correspond dans les données de référence. Les résultats sont calculés selon deux méthodes, VSE et VSE++ qui diffèrent par la fonction de coût utilisée pour l'apprentissage (Faghri *et al.*, 2017) et selon le modèle utilisé (modèles fournis avec l'outil *VSEpp* ou modèle ré-entraîné

1. Reconnaissance de 1000 classes de scènes dans des images.

2. Implémentation basée sur `https://github.com/fartashf/vsepp`, mini-batches de taille 48, pendant 30 époques à l'aide de la méthode d'optimisation Adam.

sur les données de notre tâche). 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'erreur d'association, comparé au même modèle ré-entraîné sur les données cible (21% $\rightarrow$ 37%) ce qui démontre l'importance de ré-entraîner le modèle dans des conditions identiques à celles du test.

Une fois le score d'appariement obtenu pour chaque paire $(\text{image}, \text{texte})$, 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). 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 au plus à un groupe nominal. Seules les paires de score supérieur à 0.3 sont considérées (seuil déterminé sur un corpus de développement).

5 Détection d'erreurs de rattachement

La détection d'erreurs de rattachement est réalisée à l'aide du classifieur *Icsiboost* (Favre *et al.*, 2007) qui est fondé sur l'algorithme *Adaboost*. Pour entraîner ce classifier nous avons utilisé trois catégories de caractéristiques : lexicales, conceptuelles et visuelles. 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 = Jean*, *p = avec* et *O = gants*.

En ce qui concerne les caractéristiques lexicales, en partant de l'arbre en dépendance produit par un analyseur nous utilisons : le lemme et la catégorie grammaticale du gouverneur et de l'objet, la distance entre la préposition et son gouverneur. Une description détaillée de ces caractéristiques est présentée dans de précédent travaux (Delecraz *et al.*, 2017).

Les caractéristiques conceptuelles et visuelles 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 celle utilisé dans le corpus *F30kE*, 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, étant donné deux boîtes $b_G = [x_g, y_g, w_g, h_g]$, $b_O = [x_d, y_d, w_d, h_d]$, où (x, y) sont les coordonnées du centre de la boîtes, et (w, h) sont la hauteur et la largeur de la boîte, nous utilisons les caractéristiques proposées par Peyre *et al.* (2017) :

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

Les caractéristiques V_{S1} et V_{S2} représentent respectivement la position relative horizontale et verticale entre les centres deux boîtes. V_{S3} est le rapport entre les tailles des boîtes, V_{S4} l'*IoU* entre les deux boîtes, et V_{S5} et V_{S6} le rapport de forme des deux boîtes. L'annotation du corpus *F30kE* 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 toute les boîtes de l'entité.

En se basant sur toutes ces caractéristiques, le classifieur vérifie si le rattachement proposé par l'analyseur syntaxique est correct ou non. 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 (Delecraz *et al.*, 2017). Lorsqu'un rattachement est détecté comme non correct par le classifieur, nous appliquons un ensemble de 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.

6 Cadre expérimental

Les expériences ont été réalisées à l'aide d'un analyseur en transitions entraîné sur le corpus *Penn Treebank* (Marcus *et al.*, 1993). Le corpus *F30kE*, enrichi des 29068 occurrences de préposition manuellement rattachées à leur gouverneur, a été subdivisé en trois ensembles : apprentissage (23254 prépositions), développement (2907 prépositions) et test (2907 prépositions). L'ensemble d'apprentissage a été utilisé pour entraîner le détecteur d'erreurs.

Les phrases de l'ensemble de test sont d'abord analysées puis les analyses produites sont fournies en entrée au détecteur d'erreurs qui propose éventuellement de modifier certains attachements prépositionnels. La Table 2 présente le taux de bon attachement pour les dix prépositions les plus courantes du test.

6.1 Résultats

Six configurations différentes sont évaluées : *Baseline* 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. Dans la configuration V_S seules les six caractéristiques spatiales sont utilisées et dans V, c'est l'ensemble des caractéristiques visuelles (conceptuelles et spatiales) qui sont utilisées. Pour la configuration L, le correcteur utilise les caractéristiques linguistiques et, finalement, dans VL, c'est l'ensemble des caractéristiques qui sont prises en compte.

Comme nous pouvons le voir dans la Table 2, la précision initiale de l'analyseur syntaxique est de 75% sur les dix prépositions étudiées. Nous pouvons noter que les performances varient beaucoup selon les prépositions, variant de 95% pour la préposition *through*, à 33% pour la préposition *near*. L'utilisation de caractéristiques lexicales permet un gain absolu de 11%. L'apport des caractéristiques visuelles est plus modeste, il est de 3% avec, là aussi, une variabilité importante. Le gain semble surtout toucher des prépositions locatives comme *near* ou des prépositions qui ont un usage très différent du *Penn Treebank* comme *with*.

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

est elle beaucoup plus fiable, elle n'est pas prédite, mais donnée et, dans de nombreux cas, les mots suffisent à lever l'ambiguïté.

Lorsque l'on combine toutes les caractéristiques (visuelles et lexicales), il n'y a pas de gain en moyenne sur les prépositions sélectionnées, mais on observe des améliorations comme par exemple pour *into* qui dénote généralement des relations spatiales et pour laquelle le détecteur arrive à exploiter les caractéristiques visuelles.

Préposition	**#**	**Baseline**	V_C	V_S	V	L	VL
over	111	0.66	0.64	0.66	0.68	0.85	0.84
into	116	0.89	0.89	0.89	0.89	0.92	0.95
next to	137	0.89	0.89	0.89	0.89	0.90	0.89
from	140	0.76	0.76	0.76	0.76	0.86	0.85
on	143	0.85	0.85	0.85	0.85	0.89	0.87
through	145	0.95	0.95	0.95	0.95	0.96	0.96
near	159	0.33	0.53	0.50	0.59	0.84	0.84
for	168	0.73	0.73	0.73	0.72	0.82	0.83
with	310	0.65	0.68	0.68	0.70	0.78	0.78
in	369	0.76	0.76	0.77	0.76	0.85	0.84
TOTAL	1798	0.75	0.77	0.77	0.78	0.86	0.86

TABLE 2 – Taux de rattachements correct sur le test. # indique le nombre d'occurrences de la préposition ; La *baseline* est produite par l'analyseur syntaxique ; V_C représente les concepts visuels, V_S les caractéristiques spatiales, et L sont les caractéristiques lexicales.

6.2 Analyse d'erreurs

Nous présentons ici quelques exemples de couples $(\text{image}, \text{texte})$ pour lesquels l'image a permis, ou pas, de réaliser un rattachement correct en utilisant des ensembles de traits différents. Les Figure 3.a et 3.b montrent des phrases pour lesquelles l'analyseur a effectué un mauvais rattachement que le classifieur a permis de corriger en utilisant uniquement des informations visuelles. Dans la Figure 3.a, 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 3.b la préposition *near* est incorrectement rattaché 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.

Les Figures 4.a et 4.b montrent des phrases pour lesquelles l'utilisation de caractéristiques visuelles uniquement n'a pas permis de corriger un rattachement erroné. Dans la Figure 4.a, la préposition *on* est incorrectement rattachée au mot *building* et dans la Figure 4.b 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 peuvent s'avérer plus performantes. Ainsi les Figures 5.a et 5.b 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 Figures 5.a, la préposition *with* est incorrectement rattaché au mot *jeans* à la place du mot *wearing*. Dans la Figures 5.b, la préposition *in* est rattaché au mot *bike* au lieu du mot *boy*.

a – A boy sitting on a concrete wall with a hat on.

b – Two children are in a grassy area near two horses.

FIGURE 3 – Exemples d'images pour lesquelles le classifieur a utilisé uniquement les caractéristiques visuelles pour corriger le rattachement de la préposition (en rouge).

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 4 – Exemples d'images pour lesquelles le classifieur a mal corrigé le rattachement de la préposition (en rouge) 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 – Exemples d'images avec un rattachement correct en utilisant des caractéristiques visuelles pour les prépositions en rouge alors que l'utilisation de traits lexicaux choisit un mauvais gouverneur.

7 Conclusions et perspectives

Ce travail explore la possibilité de tirer parti d'images pour désambiguïser les rattachements prépositionnels dans des phrases les décrivant, les caractéristiques visuelles améliorant en moyenne de 3 points les performances selon les prépositions, et parfois de manière drastique comme pour *near*. La difficulté du problème réside toutefois au niveau dans la détection et la catégorisation d'objets, ainsi que dans l'alignement entre le texte et les images.

Une meilleure utilisation de l'information provenant de l'image est une piste majeure d'amélioration du système avec notamment l'intégration de l'information issue directement des pixels (comme l'utilisation d'*embeddings* de l'image ou des boîtes englobantes).

Remerciements

Ces travaux ont été réalisés grâce au soutien financier apporté par la Direction Générale de l'Armement (DGA) en partenariat avec Aix-Marseille Université dans le cadre du *Club des partenaires Défense*. Les travaux de Leonor Becerra-Bonache ont été réalisés dans le cadre de sa délégation CNRS au Laboratoire d'Informatique et Système d'Aix-Marseille Université.

Références

AGIRRE E., BALDWIN T. & MARTINEZ D. (2008). Improving parsing and pp attachment performance with sense information. In *ACL*, p. 317–325.

BELINKOV Y., LEI T., BARZILAY R. & GLOBERSON A. (2014). Exploring compositional architectures and word vector representations for prepositional phrase attachment. *Transactions of the Association for Computational Linguistics*, **2**, 561–572.

CHANG A. X., MONROE W., SAVVA M., POTTS C. & MANNING C. D. (2015). Text to 3d scene generation with rich lexical grounding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, p. 53–62.

CHRISTIE G., LADDHA A., AGRAWAL A., ANTOL S., GOYAL Y., KOCHERSBERGER K. & BATRA D. (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.

COCO M. I. & KELLER F. (2015). The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, **68**(1), 46 – 74.

COYNE R. & SPROAT R. (2001). Wordseye : an automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001*, p. 487–496.

DASIGI P., AMMAR W., DYER C. & HOVY E. (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)*, volume 1, p. 2089–2098.

DE KOK D., MA J., DIMA C. & HINRICHS E. (2017). Pp attachment : Where do we stand ? *EACL 2017*, p. 311.

DELECRAZ S., NASR A., BECHET F. & FAVRE B. (2017). Correcting prepositional phrase attachments using multimodal corpora. In *Proceedings of the 15th International Conference on Parsing Technologies*, p. 72–77.

FAGHRI F., FLEET D. J., KIROS J. R. & FIDLER S. (2017). Vse++ : Improved visual-semantic embeddings. *arXiv preprint arXiv :1707.05612*.

FANG H., GUPTA S., IANDOLA F. N., SRIVASTAVA R. K., DENG L., DOLLÁR P., GAO J., HE X., MITCHELL M., PLATT J. C., ZITNICK C. L. & ZWEIG G. (2015). From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, p. 1473–1482.

FAVRE B., HAKKANI-TÜR D. & CUENDET S. (2007). Icsiboost. `http://code.google.com/p/icsiboost`.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778.

KARPATHY A. & FEI-FEI L. (2015). Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, p. 3128–3137.

MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, **19**(2), 313–330.

MIRROSHANDEL S. A. & NASR A. (2016). Integrating selectional constraints and subcategorization frames in a dependency parser. *Computational Linguistics*.

PEYRE J., LAPTEV I., SCHMID C. & SIVIC J. (2017). Weakly-supervised learning of visual relations. In *IEEE International Conference on Computer Vision, ICCV 2017*, p. 5189–5198.

PLUMMER B. A., WANG L., CERVANTES C. M., CAICEDO J. C., HOCKENMAIER J. & LAZEBNIK S. (2017). Flickr30k entities : Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, **123**(1), 74–93.

RAKSHIT G., SONTAKKE S., BHATTACHARYYA P. & HAFFARI G. (2016). Prepositional attachment disambiguation using bilingual parsing and alignments. *arXiv preprint arXiv :1603.08594*.

REDMON J. & FARHADI A. (2017). Yolo9000 : Better, faster, stronger. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, p. 6517–6525 : IEEE.

SHAERLAEKENS A. (1973). *The Two-Word Sentence in Child Language Development : A Study Based on Evidence Provided by Dutch-Speaking Triplets*. The Hague : Mouton.

SNOW C. E. (1972). Mothers' speech to children learning language. *Child Development*, **43**(2), 549 – 565.

SPIVEY M. J., TANENHAUS M. K., EBERHARD K. M. & SEDIVY J. C. (2002). Eye movements and spoken language comprehension : Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, **45**(4), 447 – 481.

VINYALS O., TOSHEV A., BENGIO S. & ERHAN D. (2015). Show and tell : A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, p. 3156–3164.

YOUNG P., LAI A., HODOSH M. & HOCKENMAIER J. (2014). From image descriptions to visual denotations : New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, **2**, 67–78.