Construction automatique d'outils et de ressources linguistiques à partir de corpus parallèles

Othman Zennaki

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Othman ZENNAKI

Thèse dirigée par **Laurent BESACIER**, Professeur, UGA
et codirigée par **Nasredine SEMMAR**, Ingénieur Chercheur, CEA LIST

Préparée au sein du **Laboratoire Vision et Ingénierie des Contenus du CEA LIST** et au **Laboratoire d'Informatique de Grenoble**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Construction automatique d'outils et de ressources linguistiques à partir de corpus parallèles

Automatic creation of linguistic tools and resources from parallel corpora

Thèse soutenue publiquement le **« date de soutenance »**, devant le jury composé de :

Mme, Sophie, ROSSET
Directrice de Recherche, LIMSI CNRS, Présidente
Mr, Reinhard, RAPP
Professeur à l'université de Mayence, Rapporteur
Mr, Mounir, ZRIGUI
Professeur à l'université de Monastir, Rapporteur
Mr, Laurent BESACIER
Professeur à l'université Grenoble Alpes, HDR, Directeur de thèse
Mr, Nasredine SEMMAR
Ingénieur chercheur, CEA LIST, Encadrant scientifique
Mr, Zied BOURAOUI
Maître de conférences, Université d'Artois, Invité

Abstract

This thesis focuses on the automatic construction of linguistic tools and resources for analyzing texts of low-resource languages. We propose an approach using Recurrent Neural Networks (RNN) and requiring only a parallel or multi-parallel corpus between a well-resourced language and one or more low-resource languages. This parallel or multi-parallel corpus is used to construct a multilingual representation of words of the source and target languages. We used this multilingual representation to train our neural models and we investigated both uni and bidirectional RNN models. We also proposed a method to include external information (for instance, low-level information from Part-Of-Speech tags) in the RNN to train higher level taggers (for instance, SuperSenses taggers and Syntactic dependency parsers). We demonstrated the validity and genericity of our approach on several languages and we conducted experiments on various NLP tasks : Part-Of-Speech tagging, SuperSenses tagging and Dependency parsing. The obtained results are very satisfactory. Our approach has the following characteristics and advantages : (a) it does not use word alignment information, (b) it does not assume any knowledge about target languages (one requirement is that the two languages (source and target) are not too syntactically divergent), which makes it applicable to a wide range of low-resource languages, (c) it provides authentic multilingual taggers (one tagger for N languages).

Keywords : Natural Language Processing, Multilingualism, Anotation, Cross-lingual transfer, Part-Of-Speech tagging, SuperSenses tagging, Dependency parsing, Recurrent neural networks.

Résumé

Cette thèse porte sur la construction automatique d'outils et de ressources pour l'analyse linguistique de textes des langues peu dotées. Nous proposons une approche utilisant des réseaux de neurones récurrents (RNN - *Recurrent Neural Networks*) et n'ayant besoin que d'un corpus parallèle ou mutli-parallele entre une langue source bien dotée et une ou plusieurs langues cibles moins bien ou peu dotées. Ce corpus parallèle ou mutli-parallele est utilisé pour la construction d'une représentation multilingue des mots des langues source et cible. Nous avons utilisé cette représentation multilingue pour l'apprentissage de nos modèles neuronaux et nous avons exploré deux architectures neuronales : les RNN simples et les RNN bidirectionnels. Nous avons aussi proposé plusieurs variantes des RNN pour la prise en compte d'informations linguistiques de bas niveau (informations morpho-syntaxiques) durant le processus de construction d'annotateurs linguistiques de niveau supérieur (SuperSenses et dépendances syntaxiques). Nous avons démontré la généricité de notre approche sur plusieurs langues ainsi que sur plusieurs tâches d'annotation linguistique. Nous avons construit trois types d'annotateurs linguistiques multilingues : annotateurs morpho-syntaxiques, annotateurs en SuperSenses et annotateurs en dépendances syntaxiques, avec des performances très satisfaisantes. Notre approche a les avantages suivants : (a) elle n'utilise aucune information d'alignement des mots, (b) aucune connaissance concernant les langues cibles traitées n'est requise au préalable (notre seule supposition est que, les langues source et cible n'ont pas une grande divergence syntaxique), ce qui rend notre approche applicable pour le traitement d'un très grand éventail de langues peu dotées, (c) elle permet la construction d'annotateurs multilingues authentiques (un annotateur pour N langages).

Mots clés : Traitement Automatique de la Langue, Multilinguisme, Annotation, Transfert interlingue, Désambiguïsation morpho-syntaxique, Annotation sémantique en gros grain, Analyse en dépendance syntaxique, Réseaux de neurones récurrents.

Remerciements

Je tiens à remercier tout d'abord mon directeur de thèse Laurent Besacier et mon encadrant scientifique Nasredine Semmar pour leur disponibilité, leurs conseils et leur implication dans cette thèse.

Mes remerciements vont également à Reinhard Rapp et Mounir Zrigui pour avoir accepté de rapporter sur cette thèse et à Sophie Rosset pour sa participation au jury.

Je souhaite également adresser mes remerciements aux membres des laboratoires LVIC et LIG.

Enfin, je tiens à exprimer ma gratitude à ma famille notamment mes parents, mes frères et ma femme Nesrine pour leurs encouragements et leur soutien.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Introduction

L'annotation linguistique de ressources consiste à ajouter des informations de nature interprétative aux données brutes originales [Garside et al. 1997]. Ces informations peuvent être d'ordre terminologique, lexical, morphologique, syntaxique ou sémantique et les ressources linguistiques peuvent être des lexiques, dictionnaires, transcriptions de dialogues ou corpus de textes [Véronis 2000].

Les applications utilisant les ressources linguistiques annotées sont nombreuses et diverses : recherche d'information interlingue, fouille de textes, extraction d'informations, aide à la traduction, traduction automatique, etc. C'est la raison pour laquelle, depuis quelques années, la construction automatique de telles ressources est devenue un champ de recherche important en Traitement Automatique de la Langue (TAL) [Hamon et al. 2007] [Viprey and Léthier 2008] [Mazziotta 2010] [Bestgen 2013].

La plupart des approches développées pour la construction de ressources linguistiques annotées ont un objectif commun : minimiser le coût de la production de telles ressources en supprimant l'intervention humaine ou en la limitant à la seule tâche de validation et d'évaluation. Le point commun de ces approches est de trouver et d'explorer des mécanismes non (ou très peu) coûteux pour exploiter des ressources linguistiques annotées déjà disponibles pour certaines langues et des corpus parallèles ou comparables pour produire de nouvelles ressources annotées pour des langues faiblement dotées.

1.1.1 Motivations

Le Traitement Automatique de la Langue (TAL) regroupe à la fois la linguistique et l'informatique. Cette discipline est devenue un axe de recherche indispensable pour analyser la grande masse d'informations disponible et qui évolue sans cesse. Les deux dernières décennies ont vu des progrès significatifs dans tous les domaines de TAL et une multitude d'applications a vu le jour dans des domaines divers et variés tels que le traitement de la parole, la traduction automatique, le résumé automatique, l'indexation et la recherche de documents, l'extraction d'informations, etc. Malheureusement, la plupart des travaux et applications concernent uniquement les langues «riches en ressources linguistiques» et plus particulièrement l'Anglais.

L'intérêt de traiter automatiquement d'autres langues coincidait avec la croissance rapide de l'Internet et de ses millions de pages Web repréesentant tout autant de textes dans différentes langues exploitables et accessibles instantanément. Toutefois la difficulté d'un traitement automatique de ces importantes masses de données réside dans le coût de constitution de ressources linguistiques nécessaires à cette automatisation, en particulier l'annotation de corpus représentatifs. En effet, la constitution de tels corpus annotés est coûteuse aussi bien par la diversité des annotations linguistiques que par le travail nécessaire pour leur filtrage et codification informatique.

1.1.2 Projection interlingue d'annotations linguistiques

La projection interlingue consiste à identifier des équivalences terminologiques, morpho-syntaxiques, syntaxiques ou sémantiques à partir de corpus de textes parallèles ou comparables. Deux processus constituent un prérequis pour les approches de projection interlingue d'annotations à partir de corpus : un alignement au niveau des paragraphes, des phrases ou d'unités lexicales de taille variable, et une analyse linguistique pour l'annotation des textes en langue source.

Ces approches permettent de produire des ressources linguistiques adéquates à moindre coût pour des langues peu dotées mais elles ouvrent également la voie à des recherches sur l'extension multilingue d'outils monolingues. Nous pourrons citer les travaux de Yarowsky

et al. [2001] qui ont utilisé un corpus parallèle pour adapter des outils monolingues (POS Taggers, chunkers et analyseurs morphologiques) à de nouvelles langues. La projection entre langues a été réalisée en utilisant des techniques d'alignement de mots entre les phrases du corpus parallèle. Cette approche a été adaptée par Hwa et al. [2002] aux nivaux grammatical et syntaxique pour faire une projection des informations concernant les dépendances syntaxiques de l'anglais vers le chinois. Feldman et al. [2006] ont expérimenté la projection interlingue à partir de corpus comparables pour transférer des étiquettes morpho-syntaxiques entre le russe, le polonais et le tchèque. L'annotation en allemand de rôles sémantiques par projection interlingue à partir de la paire de langues anglais-allemand a été déjà abordée par Padó and Lapata [2005, 2009]. Pado and Pitel [2007] ont évalué la généricité de cette approche du point de vue des langues en l'appliquant à la paire anglais-français. Les résultats sont proches de ceux obtenus pour l'allemand. Kim et al. [2011] ont utilisé des informations fournies par un aligneur de mots pour transférer les entités nommées et leurs relations de l'anglais vers le coréen en vue de la construction d'un corpus d'apprentissage pour un système d'extraction d'information à partir du Web. Abdulhay [2012] a utilisé les relations sémantiques extraites par transitivité traductionnelle à partir de corpus multilingue aligné pour la constitution d'une ressource sémantique en arabe. Plus récemment, Jabaian [2012] s'est intéressé à la portabilité multilingue d'un système de compréhension de la parole en proposant d'utiliser la traduction automatique afin de minimiser le coût du développement d'un nouveau système de compréhension dans une nouvelle langue [Jabaian et al. 2013].

Les approches de projection interlingue par alignement de mots affichent des résultats satisfaisants en annotations lexicales et morpho-syntaxiques pour les couples de langues voisines, mais les résultats pour les annotations syntaxiques et sémantiques des langues à morphologie riche restent insufisants.

1.1.3 Multilinguisme et traitement automatique de la langue

Les technologies du TAL ont pour objectif d'aider à analyser rapidement de grosses quantités de données textuelles. Cette analyse (linguistique) consiste à déterminer les unités de sens que contient le texte à traiter. Pour réaliser cette tâche, le processus d'analyse

linguistique a besoin d'un ensemble de modules de traitement dont le nombre et la nature varient selon la langue considérée et d'un ensemble de ressources linguistiques adaptées.

Il existe principalement deux types d'approches utilisées pour le développement des outils de TAL : celles à base de règles et de lexiques dites « symboliques » [Fuchs 1993] et celles s'appuyant sur des corpus dites « statistiques » [Cornuéjols and Miclet 2002]. La combinaison de ces deux approches a permis le développement de méthodes hybrides [Jurafsky 2000].

L'analyse linguistique à base de règles repose sur des ressources généralement construites à la main. L'objectif étant le transfert de l'expertise des linguistes pour disposer des lexiques et des règles nécessaires au fonctionnement des outils de TAL. Le principal avantage de cette approche est qu'elle fournit des résultats présentant un minimum de qualité lexicale et grammaticale due à l'utilisation de ces ressources linguistiques.

Les approches statistiques tablent sur la mise en évidence, par des techniques d'apprentissage automatique, des régularités présentes dans des corpus significatifs de textes. Les performances des outils TAL utilisant des modèles neuronaux profonds pour les langues riches en données annotées, s'approchent de plus en plus de la performance humaine. En revanche, les performances de ces outils pour les langues peu dotées sont largement en-dessous des performances des modèles de l'état de l'art basés sur les ressources linguistiques (lexiques et règles grammaticales). Ceci est dû au fait que l'apprentissage profond nécessite des données annotées volumineuses pour fournir des performances élevées.

1.1.3.1 Modules et ressources pour l'analyse linguistique monolingue

Certains modules de l'analyse linguistique sont génériques dans la mesure où ils peuvent assurer le traitement de la majorité des langues traitées. D'autres, plus spécifiques, ne sont utilisés que dans des cas bien précis définis selon la langue à traiter. Une analyse linguistique standard se compose des modules suivants [Besançon et al. 2010] :

1. Tokenisation : Ce module consiste à découper les chaînes de caractères du texte en mots, en prenant en compte le contexte ainsi que les règles de découpage. Ce module utilise généralement des règles de segmentation ainsi que des automates d'états finis.

2. Analyse morphologique : Ce module a pour but de vérifier si le mot (token) appartient à la langue et d'associer à chaque mot des propriétés syntaxiques qui vont servir dans la suite des traitements. Ces propriétés syntaxiques sont décrites en classes appelées catégories grammaticales. La consultation de dictionnaires de formes ou de lemmes permet de récupérer les propriétés syntaxiques concernant les mots à reconnaître.
3. Analyse morpho-syntaxqiue : Après l'analyse morphologique, une partie des mots restent ambigus d'un point de vue grammatical. Par exemple, le mot "car" peut avoir la catégorie grammaticale "Conjonction" ou "Substantif". L'analyse morpho-syntaxique réduit le nombre des ambiguïtés en utilisant soit des règles ou des matrices de désambiguïsation. Les règles sont généralement construites manuellement et les matrices de bi-grams et tri-grams sont obtenues à partir d'un corpus étiqueté et désambiguïsé manuellement.
4. Analyse syntaxique : Ce module consiste à identifier les principaux constituants de la phrase et et les relations qu'ils entretiennent entre eux. Le résultat de l'analyse syntaxique peut être une ou plusieurs structures syntaxiques représentant la phrase en entrées. Ces structures dépendent du formalisme de représentation utilisé : un arbre syntagmatique, un arbre de dépendance ou une structure de traits. L'analyse en dépendance syntaxique consiste à créer un arbre de relations entre les mots de la phrase. Le module d'analyse syntaxique utilise des règles pour l'identification des relations de dépendance ou des corpus annotés en étiquettes morpho-syntaxiques et en relations de dépendance.
5. Analyse sémantique : Ce module a pour objectif de construire une représentation du sens de la phrase à analyser, en associant à chaque concept rencontré un objet ou une action appartenant à un référentiel. Il existe plusieurs niveaux pour l'analyse sémantique : désambiguïsation lexicale, étiquetage en rôles sémantiques, etc.

1.1.3.2 Modules et ressources pour l'analyse multilingue

Les premiers outils de traitement automatique de la langue à base de règles utilisaient des ressources monolingues : lexiques, règles syntaxiques, etc. Mais pour étendre ces outils à de nouvelles langues, il faut construire des ressources pour chaque langue. Cependant, la

construction de ressources linguistiques de qualité est une tâche lente et coûteuse.

1. Corpus parallèles : Un corpus parallèle est un ensemble de traductions et de leurs originaux respectifs, alignés au niveau des paragraphes, phrases ou même des mots.

   (a) Hansard [Roukos et al. 1995] : Ce corpus parallèle anglais-français est composé des débats du Parlement canadien et a été constitué dans les années 1990. Il a été utilisé en traduction automatique et en extraction de terminologie bilingue.

   (b) Arcade : Ce corpus a été constitué dans le cadre du projet ARCADE II [Veronis et al. 2008]. Il est composé de deux corpus multilingues : un corpus extrait du Journal Officiel de la Communauté Européenne (corpus JOCE) et un corpus extrait du journal Le Monde Diplomatique (corpus MD).Le corpus JOCE des langues à écriture latine (français, anglais, allemand, italien et espagnol) contient un million de mots par langues. Il est constitué des questions écrites des parlementaires européens à la Commission et des réponses de celle-ci pour l'année 1993. Le corpus MD est composé de textes parallèles aux contenus homogènes dans une large gamme de langues distantes : arabe, chinois, français, grec, japonais, persan et russe.

   (c) Opus : Opus est base de données composée de plusieurs corpus parallèle et multilingue mis en libre accès sur Internet [1]. Cette base contient entre autres le texte de la Constitution européenne en 21 langues. Ces corpus parallèles ont été particulièrement utilisés en traduction statistique mais aussi en traduction assistée par ordinateur. Plusieurs outils utilisant les techniques d'alignement ont été développés pour exploiter ces corpus parallèles dans le but de produire des bases de données terminologiques (alignement au niveau du mot) ou des mémoires de traduction (alignement au niveau de la phrase).

   Nous avons utilisé lors de nos expérimentations le corpus ARCADE et le corpus Europarl qui fait partie de la base Opus.

2. Plongements lexicaux multilingues : Les word embeddings monolingues sont des représentations continues des mots d'un texte ou d'une séquence, en se basant

1. http ://urd.let.rug.nl/tiedeman/OPUS/

sur la notion de contexte auquel appartient le mot [Mikolov et al. 2013a]. Ces représentations vectorielles permettent de capturer la sémantique exprimée par chacune des occurrences. Ces représentations sont utilisées pour comparer les termes entre eux et exprimer, sous forme de distances, les relations potentielles entre les termes.

(a) MultiVec : MultiVec est un ensemble d'outils permettant de calculer des représentations continues de texte à différents niveaux de granularité (mots ou séquences de mots). Pour comparer les représentations des mots ou des séquences de mots, Multivec propose plusieurs mesures de similarité ou de distance [Bérard et al. 2016].

(b) Barista : BARISTA (Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives) [Gouws et al. 2015] est un modèle fondé sur l'utilisation d'une liste de paires de mots (mot source traduction cible ou bien mot source-mot cible portant la même information linguistique) pivots pour la construction d'une représentation distribuée bilingue. Cette représentation est apprise en utilisant un modèle neuronal sur un corpus bilingue dans lequel les mots pivots d'une langue sont remplacés aléatoirement par les mots correspondants dans l'autre langue.

1.1.4 Structure de la thèse

Cette thèse comprend six chapitres :

Le premier chapitre décrit nos motivations et une partie théorique sur les différentes annotations linguistiques ainsi que les réseaux de neurones utilisés en traitement automatique de la langue.

Dans le chapitre 2 nous passons en revue les différents travaux existants concernant la projection interlingue d'annotation à partir de corpus parallèles.

Le chapitre 3 est consacré à la présentation de notre approche d'annotation multilingue à partir de corpus parallèles. Nous décrivons, dans un premier temps, les différentes

architectures neuronales standard utilisées. Ensuite, nous présentons les améliorations que nous avons apporté à ces architectures.

Dans le chapitre 4, nous présentons un annotateur morpho-syntaxique multilingue contsruit en utilisant des réseaux de neurones récurrents.

Le chapitre 5 est centré sur l'annotation en SuperSenses étendant le modèle neuronal utilisé en désambiguisation morpho-syntaxique.

Dans le chapitre 6, nous étudions l'application de notre modèle neuronal pour l'annotation en dépendance syntaxique.

Enfin, nous présentons les conclusions et les perspectives de cette thèse.

1.2 Jeux d'étiquettes syntaxiques et sémantiques

Nous décrivons dans cette section les jeux d'étiquettes utilisées par les différents outils que nous avons développés pour l'analyse morpho-syntaxique, syntaxique et sémantique.

1.2.1 Les étiquettes morpho-syntaxiques universelles

Pour évaluer notre analyseur morpho-syntaxique utilisant la projection interlingue et fondée sur les réseaux de neurones récurrents, nous avons utilisé des étiquettes universelles [Petrov et al. 2012]. Ceci est dû au fait que notre approche suppose un jeu commun d'étiquettes morpho-syntaxiques entre la langue source et les langues cibles. Le tableau 1.1 met en correspondance les étiquettes du Penn TreeBank et les étiquettes universelles pour l'anglais. Une seule étiquette est attribuée à chaque mot en fonction de son rôle dans la phrase et les mots sont classés à partir de huit catégories grammaticales : le verbe (VB), le nom (NN), le pronom (PR + DT), l'adjectif (JJ), l'adverbe (RB), la préposition (IN), la conjonction (CC), et l'interjection (UH).

1.2.2 Les étiquettes sémantiques à gros grain

L'annotation sémantique à gros grain (SuperSenses) est une tâche qui consiste à annoter chaque unité du texte, avec un jeu d'étiquettes sémantiques générales définies par les catégories lexicographiques de WordNet (SuperSenses). Elle peut être vue comme une

Etiquette Penn TreeBank	Description	Etiquette Universelle
CC	conjunction, coordinating	CONJ
CD	cardinal number	NUM
DT	determiner	DET
EX	existential there	DT
FW	foreign word	X
IN	conjunction, subordinating or preposition	ADP
JJ	adjective	ADJ
JJR	adjective, comparative	ADJ
JJS	adjective, superlative	ADJ
LS	list item marker	X
MD	verb, modal auxillary	VERB
NN	noun, singular or mass	NOUN
NNS	noun, plural	NOUN
NNP	noun, proper singular	NOUN
NNPS	noun, proper plural	NOUN
PDT	predeterminer	DET
POS	possessive ending	PRT
PRP	pronoun, personal	PRON
PRPDOL	pronoun, possessive	PRON
RB	adverb	ADV
RBR	adverb, comparative	ADV
RBS	adverb, superlative	ADV
RP	adverb, particle	PRT
SYM	symbol	X
TO	infinitival to	PRT
UH	interjection	X
VB	verb, base form	VERB
VBZ	verb, 3rd person singular present	VERB
VBP	verb, non-3rd person singular present	VERB
VBD	verb, past tense	VERB
VBN	verb, past participle	VERB
VBG	verb, gerund or present participle	VERB
WDT	wh-determiner	DET
WP	wh-pronoun, personal	PRON
WPDOL	wh-pronoun, possessive	PRON
WRB	wh-adverb	ADV
.	punctuation mark, sentence closer	.
,	punctuation mark, comma	.
:	punctuation mark, colon	.
(	contextual separator, left paren	.
)	contextual separator, right paren	.

TABLE 1.1 – Correspondance entre le jeu d'étiquettes morpho-syntaxiques du Penn treebank et des étiquettes Universelles.

tâche à cheval entre la reconnaissance d'entités nommées et la désambiguïsation lexicale. La reconnaissance d'entités nommées consiste à identifier des objets textuels (i.e. un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms

d'organisations ou d'entreprises,noms de lieux, quantités, distances, valeurs, dates, etc. Le rôle de la désambiguïsation lexicale est de sélectionner les sens corrects d'instances contextualisées de mots ambigus, parmi l'ensemble de leurs sens possibles (ou sens candidats). La complexité à modéliser et à traiter l'ambiguïté lexicale ainsi que les limitations de la reconnaissance d'entités nommées ont fait émerger l'annotation sémantique lexicale (sens lexical) à gros grain. Ce type de d'annotation permet par exemple de lever l'ambiguïté lexicale des mots. Le tableau ci-dessous présente le jeu d'étiquettes SuperSenses [Ciaramita and Johnson 2003]. Ce jeu comprend au total 41 sens, répartis en deux catégories : 26 étiquettes pour représenter les sens des noms et 15 autres pour représenter les sens des verbes, plus une étiquette unique (catch-all) pour les autres unités (adjectifs, adverbes, etc.).

NOUNS			
SuperSens	NOUNS DENOTING	SuperSens	NOUNS DENOTING
act	acts or actions	object	natural objects (not man-made)
animal	animals	quantity	quantities and units of measure
artifact	man-made objects	phenomenon	natural phenomena
attribute	attributes of people and objects	plant	plants
body	body parts	possession	possession and transfer of possession
cognition	cognitive processes and contents	process	natural processes
communication	communicative processes and contents	person	people
event	natural events	relation	relations between people or things or ideas
feeling	feelings and emotions	shape	two and three dimensional shapes
food	foods and drinks	state	stable states of affairs
group	groupings of people or objects	substance	substances
location	spatial position	time	time and temporal relations
motive	goals	Tops	abstract terms for unique beginners
VERBS			
SuperSens	VERBS OF	SuperSens	VERBS OF
body	grooming, dressing and bodily care	emotion	feeling
change	size, temperature change, intensifying	motion	walking, flying, swimming
cognition	thinking, judging, analyzing, doubting	perception	seeing, hearing, feeling
communication	telling, asking, ordering, singing	possession	buying, selling, owning
competition	fighting, athletic activities	social	political and social activities and events
consumption	eating and drinking	stative	being, having, spatial relations
contact	touching, hitting, tying, digging	weather	raining, snowing, thawing, thundering
creation	sewing, baking, painting, performing		

TABLE 1.2 – Le jeu d'étiquettes SuperSenses (annotation sémantique à gros grain).

1.2.3 Les étiquettes des relations de dépendance syntaxique universelles

Le but de l'analyse en dépendance syntaxique est d'établir des relations entre les mots d'une phrase. Chaque mot de la phrase est gouverné par un unique autre mot. Les mots sont liés par une relation de type gouverneur-dépendant ayant un rôle syntaxique spécifique. L'ensemble de ces relations syntaxiques pour une phrase donnée constitue la structure de dépendances de cette phrase. L'étiquette d'une relation de dépendance syntaxique décrit le

Etiquette Universelle	Description
root	the head of a sentence
nsubj	nominal subject
nsubjpass	passive nominal subject
csubj	clausal subject
csubjpass	clausal passive subject
dobj	direct object
iobj	indirect object
ccomp	clausal complement
xcomp	open clausal complement
nmod	nominal modifier
advmod	adverbial modifier
advcl	adverbial clause modifier
neg	negation
appos	apposition
amod	adjectival modifier
acl	clausal modifier of a noun (adjectival clause)
det	determiner
case	case marking
vocative	addressee
aux	auxiliary verb
auxpass	passive auxiliary
cop	copula verb
mark	subordinating conjunction
expl	expletive
conj	conjunct
cc	coordinating conjunction
discourse	discourse element
compound	relation for marking compound words
name	names
mwe	multiword expressions that are not names
foreign	text in a foreign language
goeswith	two parts of a word that are separated in text
list	used for chains of comparable elements
dislocated	dislocated elements
parataxis	parataxis
remnant	remnant in ellipsis
reparandum	overridden disfluency
punct	punctuation
dep	unspecified dependency

TABLE 1.3 – Le jeu d'étiquettes des relations de dépendance syntaxique universelles.

rôle que les mots gouverneur et dépendant ont l'un pour l'autre. Plusieurs jeux d'étiquettes pour les relations de dépendance syntaxique ont été élaborés mais un jeu d'étiquettes universelles a été développé dans le cadre du projet UD [Nivre et al. 2016]. L'objectif de ce projet est de mettre à la disposition de la communauté du TAL des corpus en dépendances dans différentes langues dont le jeu d'étiquettes est identique. Le tableau ci-dessous présente la liste des étiquettes des relations de dépendance syntaxique universelles [2].

2. http ://universaldependencies.github.io/docs/u/dep/index.html

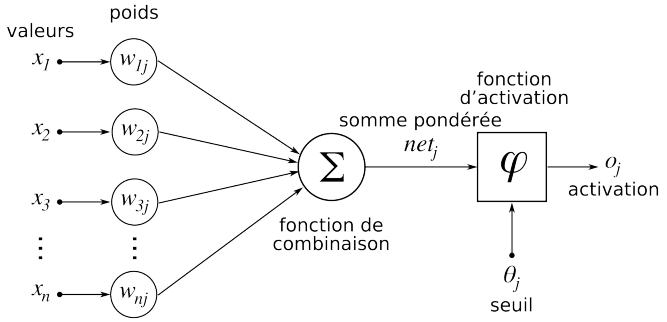

FIGURE 1.1 – Neurone formel.

1.3 Les réseaux de neurones pour l'analyse linguistique

Un réseau de neurones peut être décrit comme un graphe pondéré, composé d'une couche d'entrée, d'une ou plusieurs couches cachées, et d'une couche de sortie. Chaque couche est composée d'un ensemble de neurones formels. Un neurone formel peut être défini par les cinq éléments suivants (voir figure 1.1) :

1. La nature de ses entrées.
2. La fonction d'entrée qui définit le pré-traitement effectué sur les entrées.
3. La fonction d'activation du neurone qui définit son état interne en fonction de son entrée.
4. La fonction de sortie qui calcule la sortie du neurone en fonction de son état d'activation.
5. La nature de la sortie du neurone.

Plusieurs architectures neuronales ont été expérimentées ces dernières années. Parmi ces modèles on trouve les réseaux de neurones feedforward et les réseaux de neurones récurrents.

Le réseau de neurones feedforward est l'un des réseaux multicouches les plus simples [Hornik et al. 1989]. Ce modèle canalise l'entrée directement à travers le réseau, sans jamais toucher deux fois le même nœud. Par conséquent, l'information d'une entrée est perdue après son traitement, et chaque entrée est traitée de façon complètement indépendante des autres [Glorot and Bengio 2010].

Les réseaux de neurones récurrents (on utilise l'acronyme RNN en anglais - pour *Recurrent Neural Networks*) sont théoriquement plus puissants que les réseaux feedforward en modélisation du langage en raison de leur capacité à représenter tous les mots de l'historique plutôt que dans un contexte à longueur fixe [Graves 2012]. Les réseaux de neurones récurrents sont des réseaux de neurones optimisés pour l'apprentissage séquentiel où la sortie précédente influencera la prédiction suivante. Ceci est réalisé par une boucle qui renvoie les sorties au début d'une couche cachée, conservant ainsi l'information séquentielle dans l'état caché du réseau. La nouvelle entrée en conjonction avec l'ancien état caché détermine la sortie suivante.

Les réseaux de neurones récurrents se sont montrés les plus efficaces pour les tâches d'annotation de séquences de mots. C'est à ce type de modèle neuronal que nous nous intéressons principalement dans cette thèse et que nous décrivons en détails dans le chapitre 3.

Notons, enfin, que de plus en plus de travaux de TAL utilisent les réseaux de neurones [Bengio et al. 2003; Collobert et al. 2011; Henderson 2004; Mikolov et al. 2010; Federici and Pirrelli 1993]. Federici and Pirrelli [1993] font partie des premiers à avoir développé des annotateurs morpho-syntaxiques basés sur les réseaux de neurones, Bengio et al. [2003] et Mikolov et al. [2010] ont utlisé les réseaux de neurones pour construire des modèles de langages. Collobert et al. [2011] ont employé les techniques de réseaux de neurones profonds pour l'apprentissage multitâche qui comprend : l'annotation morpho-syntaxique, la reconnaissance d'entités nommées et l'annotation sémantique. Henderson [2004] a aussi proposé des méthodes d'apprentissage d'analyseurs syntaxiques fondées sur les réseaux de neurones.

Chapitre 2

Projection interlingue d'annotations linguistiques - État de l'art

2.1 Introduction

"Comment adapter des outils de TAL conçus pour une langue richement dotée en ressources linguistiques à des langues peu dotées ?", cette question a constitué un sujet de recherche pendant plusieurs années du fait de la disponibilité de plus en plus de corpus de textes parallèles entre les langues richement dotées et les langues peu dotées.

Dans ce chapitre, nous présentons un état de l'art sur les approches de projection interlingue (transfert cross-lingue) d'annotations.

Nous commencerons par un rappel de quelques notions générales sur les annotations linguistiques et les langues peu dotées. Nous présenterons ensuite les principales approches de projection interlingue fondées sur l'alignement de mots. Après avoir passé en revue les principales méthodes d'alignement de mots à partir de corpus de textes parallèles, nous décrirons par la suite les approches non-supervisées et semi-supervisées pour le transfert d'annotations linguistiques. Enfin, nous clôturons ce chapitre par les approches utilisant les plongements lexicaux bilingues.

2.2 Définitions

2.2.1 Annotation linguistique

L'annotation linguistique de ressources consiste à ajouter des informations de nature interprétative aux données brutes originales [Garside et al. 1997]. Ces informations peuvent être d'ordre terminologique, lexical, morphologique, syntaxique ou sémantique et les ressources linguistiques peuvent être des lexiques, dictionnaires, transcriptions de dialogues ou corpus de textes [Véronis 2000; Eshkol-Taravella 2015].

Depuis les premiers travaux du traitement automatique de la langue, trois approches pour l'annotation de ressources linguistiques ont été explorées : manuelle, automatique et semi-automatique. L'approche manuelle sollicite des annotateurs humains pour interpréter et décider de l'annotation à affecter en utilisant un guide d'annotation qui doit être clair, exhaustif et non ambigu [Mélanie-Becquet and Landragin 2014]. Il est, bien entendu, évident que les annotations effectuées par des annotateurs humains dépendent de leur expertise du domaine annoté et de leurs connaissances théoriques. Pour chiffrer la fiabilité (qualité) des annotations, plusieurs mesures ont été développées pour calculer le degré de concordance entre les annotateurs (accord inter- et intra-annotateur). Parmi ces mesures, on trouve l'indice Kappa [Cohen 1960; Carletta 1996]. L'approche automatique consiste à utiliser des outils d'annotation ne nécessitant aucune intervention humaine. Là encore, les annotations produites dépendent des choix théoriques et méthodologiques utilisés lors de la conception de ces outils d'annotation. Les annotations produites automatiquement par des outils peuvent être vérifiées et validées par des annotateurs humains, c'est le rôle de l'annotation semi-automatique.

Nous nous intéressons dans cette thèse au développement d'outils pour l'annotation automatique de corpus. Nous nous focalisons sur les étiquettes morpho-syntaxiques, les étiquettes sémantiques de type SuperSenses et les étiquettes syntaxiques de type relation de dépendance. Nous avons fait le choix d'utiliser des étiquettes universelles pour comparer les résultats de nos outils avec ceux de l'état de l'art [Petrov et al. 2012] [Nivre et al. 2016].

2.2.2 Langues peu dotées

Les langues peu dotées désignent généralement les langues moins bien informatisées que les grandes langues véhiculaires comme l'anglais, l'espagnol ou le français.

Cependant cette définition est vague et non précise. Nous pouvons affiner cette définition en utilisons les travaux de Berment [2004], qui a proposé de catégoriser les langues, selon leur degré d'informatisation en trois catégories : langues très bien dotées, langues moyennement dotées et langues peu ou pas dotées. Le degré d'informatisation d'une langue est évalué par un groupe de locuteurs de cette langue selon une multitude de critères (tels que la disponibilité des outils et ressources : traitement du texte, traitement de l'oral, traduction, etc.). Le degré d'informatisation d'une langue peu dotée est compris entre 0 et $9,99$, on dit qu'une langue est moyennement dotée si son degré d'informatisation est entre 10 et 13.99, et finalement une langue est dite très bien dotée si son degré se situe entre 14 et 20. Cette façon de catégoriser les langues a été critiquée par Prys [2006], qui juge que plusieurs critères, reflétant la disponibilité ou non des ressources plus basiques, ne sont pas pris en compte, tels que la disponibilité ou non des journaux quotidiens sous forme électronique, qui sont utiles dans plusieurs applications du TAL.

Duong [2017] soulève une autre interrogation, il existe une disparité au sein d'une même langue. Selon les tâches traitées, la disponibilité des ressources et les performances des outils traitant ces tâches diffèrent. Par exemple, du point de vue de l'annotation morpho-syntaxique, avec des annotateurs morpho-syntaxiques ayant des performances avoisinant les 97% de précision [Parra Escartín and Martínez Alonso 2015], l'espagnol n'est clairement pas une langue peu dotée. Mais du point de vue de l'analyse automatique des sentiments, l'espagnol, ne possédant pas de corpus annotés, est vu comme une langue peu dotée. Il est donc plus judicieux de caractériser une langue par richement ou peu dotée selon une tâche donnée. Duong [2017] considère donc une langue comme peu dotée pour une tâche donnée, s'il n'existe pas d'algorithme, utilisant les données disponibles, réalisant cette tâche de façon automatique et avec des performances adéquates.

2.2.3 Alignement de mots

L'alignement de mots à partir de corpus de textes parallèles peut se décomposer conceptuellement en deux aspects : il s'agit de repérer les mots du texte source et du texte cible, puis de les mettre en correspondance. Il existe principalement trois approches pour l'alignement de mots à partir de corpus de textes parallèles alignés phrase à phrase :

— Les approches à dominante statistique qui s'appuient par exemple sur les modèles IBM [Brown et al. 1993]. L'outil d'alignement GIZA++ [Och and Ney 2000] implémente notamment ce type d'approche. Cet outil implémente divers modèles de traduction (IBM 1, 2, 3, 4, 5 et HMM). Ces modèles utilisent l'algorithme EM [Dempster et al. 1977] pour l'apprentissage à partir de corpus bilingues. L'alignement des mots est réalisé à l'aide d'un algorithme de recherche de type Viterbi.

— Les approches linguistiques qui utilisent généralement des dictionnaires bilingues déjà disponibles mais aussi les résultats de l'analyse morpho-syntaxique des phrases en langues source et cible [Debili and Sammouda 1992; Debili and Zribi 1996; Bisson 2000]. Les méthodes proposées par [Debili and Zribi 1996] ainsi que [Bisson 2000] utilisent des ressources linguistiques externes (lexiques, règles, etc.) pour apparier les mots des textes parallèles alignés au niveau de la phrase.

— Une combinaison des méthodes statistiques avec différentes sources d'information linguistique [Bourigault 1992; Daille et al. 1994; Gaussier and Langé 1995; Smadja et al. 1996; Blank 2000; Ozdowska 2004; Ozdowska and Claveau 2006; Semmar and Meriama 2010; Bouamor et al. 2012].

2.2.4 Transfert d'annotations

La projection multilingue d'annotations linguistiques a été introduite au début des années 2000 par Yarowsky et al. [2001]. Pour adapter des outils monolingues (analyseurs morphologiques, analyseurs morpho-syntaxiques et analyseurs syntaxiques) à de nouvelles langues, ces auteurs ont proposé de s'appuyer sur l'alignement des mots d'un corpus parallèle pour transférer les annotations (les étiquettes morpho-syntaxiques) d'une langue riche en ressources (anglais) vers des langues peu dotées en ressources (français, chinois, tchèque, espagnol). Cette méthode a été utilisée avec succès dans plusieurs travaux pour la

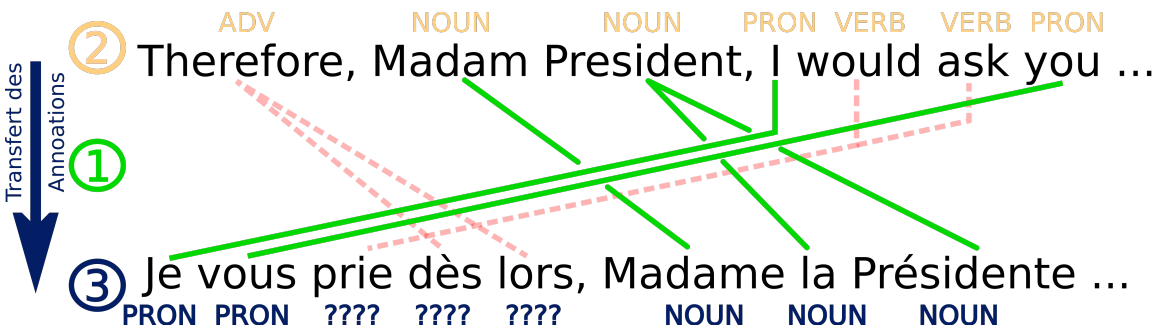

FIGURE 2.1 – Exemple de la projection d'annotation morpho-syntaxique de l'anglais vers le français extrait du corpus parallèle Europarl : l'étape 1 est l'alignement en mots des deux côtés du corpus parallèle, l'étape 2 est l'annotation du côté source et l'étape 3 est la projection des annotations du côté source vers le côté cible. Les mots *"prie, dès, lors"* n'ont pas été annotés car ils n'ont pas été alignés, le mot *"la"* a été mal annoté car aligné par erreur.

projection de plusieurs types d'annotations et de concepts entre une multitude de langues source et cible. On peut citer par exemple : l'apprentissage d'analyseurs morpho-syntaxiques par projection [Das and Petrov 2011; Li et al. 2012; Duong et al. 2013; Täckström et al. 2013a; Wisniewski et al. 2014], le transfert d'entités nommées [Kim et al. 2012], la reconnaissance d'entités nommées bilingues [Wang et al. 2013], l'annotation syntaxique [Jiang et al. 2011; McDonald et al. 2011; Ma and Xia 2014; Rasooli and Collins 2015; Lacroix et al. 2016], la projection d'annotations en sens réalisée par [Bentivogli et al. 2004; van der Plas and Apidianaki 2014], l'annotation en rôles sémantiques sur l'allemand par projection interlingue à partir de la paire de langues anglais-allemand [Padó and Lapata 2005] et dont la généricité a été évaluée dans [Pado and Pitel 2007; Annesi and Basili 2010]. L'approche de transfert d'annotation a été aussi utilisée dans la compréhension automatique de la parole [Jabaian et al. 2013].

Nous présentons dans la figure 2.1 un exemple de projection d'annotation morpho-syntaxique de l'anglais vers le français. L'algorithme commence par une phase d'alignement, puis la langue source est annotée et les annotations sont projetées dans la langue cible. Les annotations qui en résultent, bien qu'elles soient imparfaites (bruitées), sont utilisées en conjonction avec des algorithmes d'apprentissage pour entraîner des systèmes de traitement de la langue cible et construire des outils robustes, peu coûteux et non supervisés. Les

enjeux des approches utilisant la projection interlingue d'annotations, sont les suivants : (a) les alignements sont bruités, ils ne sont pas toujours corrects (par exemple, l'alignement du mot en français *la* avec le mot en anglais *president*, dans l'exemple de la figure 2.1, est erroné et donc l'annotation projetée est fausse), (b) les alignements sont incomplets, le coté cible n'est pas complètement annoté par projection (par exemple, les mots *"prie, dès, lors"* n'ont pas été annotés, dans l'exemple de la figure 2.1, car ils n'ont pas été alignés) et (c) la divergence linguistique entre les différentes langues traitées, ce qui est exprimé dans une langue n'est nécessairement pas présent ou exprimé de la même façon dans une autre langue, et donc l'annotation projetée du côté source n'est pas toujours correcte dans la langue cible, par exemple, le phénomène de recatégorisation entre le français et l'anglais présenté par [Ballard 1991], prenant la phrase en anglais "*He* ***nearly*** *got arrested*" et sa traduction en français "*il* ***faillit*** *se faire arrêter*", la reacatégorisation est au niveau de mot *nearly* (adverbe) est sa traduction *faillit* (verbe), si on projette les annotations de l'anglais, l'annotation du mot *faillit* sera erronée.

Ces raisons font que le résultat bruité de la projection d'annotation (côté cible du corpus parallèle annoté par projection) n'est jamais utilisé dans son état brut, mais post-traité en vue de limiter le bruit engendré par la projection d'annotation, soit avec des méthodes robustes comme présenté par Hwa et al. [2005] ou bien en utilisant des contraintes pour limiter les annotations possibles et admises pour chaque mot du corpus cible [Das and Petrov 2011; Täckström et al. 2013a]. Pour notre part, l'approche que nous proposons [Zennaki et al. 2015b] n'utilise pas d'alignement au niveau des mots, qui sont source d'erreurs. Nous nous basons sur l'alignement des phrases du corpus parallèle pour construire une représentation bilingue des mots (des langues source et cible), cette représentation sera utilisée pour l'apprentissage de nos modèles neuronaux.

Nous avons aussi proposé, en vue de gérer la divergence linguistique entre les langues source et cible, l'utilisation d'un petit corpus annoté en langue cible pour l'adaptation de nos modèles [Zennaki et al. 2015a].

2.3 État de l'art

Il existe principalement deux approches pour la projection interlingue ou le transfert d'annotations linguistiques à partir de langues bien dotées (source) vers les langues faiblement ou peu dotées (cible) : celles utilisant les résultats de l'alignement au niveau des mots et celles s'appuyant sur des représentations interlingues/multilingues entre les langues source et cible.

Dans ce qui suit nous présenterons un état de l'art des travaux traitant de la projection interlingue d'annotation. Nous avons réalisé nos travaux sur trois tâches du TAL : l'annotation morpho-syntaxique, l'annotation sémantique à gros grain (SuperSenses) et l'analyse en dépendance syntaxique. Nous mettons plus l'accent sur les approches de l'état de l'art utilisant la projection interlingue pour réaliser une de ces trois tâches.

2.3.1 Approches basées sur l'alignement de mots

Les approches utilisant la projection interlingue consistent à exploiter des corpus de textes parallèles pour transférer les annotations produites par des outils d'analyse linguistique appliqués à une langue source riche en données annotées vers une langue cible moins bien dotée [Yarowsky et al. 2001].

Plusieurs approches [Yarowsky et al. 2001; Das and Petrov 2011; Duong et al. 2013; Ma and Xia 2014; Lacroix et al. 2016] ont exploité, du fait de leur caractère non-supervisé, les outils d'alignement statistique, ces outils d'alignement de mots sont généralement efficaces pour aligner les mots simples, mais leurs performances ne sont pas toujours satisfaisantes, d'une part, lorsque les langues source et cible ont des morphologies et des structures syntaxiques différentes, et d'autre part, pour aligner les expressions multi-mots [Allauzen and Wisniewski 2010]. Cela a constitué la principale limite pour la projection d'annotations linguistiques. Comme nous verrons dans ce qui suit (section 2.3.1.2) plusieurs études ont proposé [Li et al. 2012; Täckström et al. 2013a; Wisniewski et al. 2014], pour palier aux limites des méthodes d'alignement statistique, de les combiner avec différentes ressources linguistiques.

2.3.1.1 Méthodes non-supervisées

L'aspect non-supervisé de ces méthodes porte sur la langue cible, elles n'utilisent donc pas de ressources en langue cible.

Annotation morpho-syntaxique

Les pionniers dans le domaine de la projection interlingue d'annotations linguistique sont Yarowsky et al. [2001], ils ont proposé d'utiliser un corpus parallèle en vue de projeter les annotations linguistiques d'une langue vers une autre via les alignements automatiques au niveau des mots. Yarowsky et al. [2001] ont pu démontrer l'efficacité de cette approche avec la construction de deux types outils d'analyse linguistique : des analyseurs morpho-syntaxiques et des analyseurs syntaxiques de surface (étape préalable aux analyses syntaxique et sémantique), ces outils ont été construits pour deux langues cibles le Français et le Chinois avec l'Anglais comme langue source.

Tout d'abord, ils ont annoté le côté anglais (source) à l'aide d'un dispositif d'étiquetage supervisé. Puis, à l'aide des alignements au niveau des mots du corpus parallèle, ils ont projeté les annotations du côté source (anglais) vers les langues cibles (le français et le chinois). Ils ont observé que, bien que la projection soit correcte dans de nombreux cas, les étiquettes projetées restent très bruitées. Par conséquent ils proposent d'utiliser une heuristique basée sur le calcul d'un score de confiance, pour chaque couple de phrase aligné, ils attribuent un score d'alignement, et ainsi ils ne gardent que les couples de phrases avec un score d'alignement élevé. Ils n'utilisent, pour la projection d'annotations, que les alignements non bruités. Enfin, le côté cible annoté par projection est utilisé pour l'apprentissage d'outils d'analyse linguistique en langue cible, ces outils seront utilisés pour annoter d'autres textes par la suite. Yarowsky et al. [2001] ont obtenu des performances très satisfaisantes sur l'annotation morpho-syntaxique par projection.

Dans nos premiers travaux [Zennaki et al. 2015b,a], nous avons aussi traité cette tâche, nous avons proposé une méthode basée sur les RNN pour la construction d'annotateurs morpho-syntaxiques multilingues, notre méthode n'utilise que les alignements au niveau des phrases du corpus parallèle et n'utilise pas d'informations d'alignement au niveau

des mots (source d'erreurs et de bruits). Ayant fait le choix d'utiliser dans nos procédés expérimentaux les mêmes ressources (pour l'apprentissage et le test de nos modèles) ainsi que la même métrique d'évaluation que Das and Petrov [2011] et Duong et al. [2013], qui sont des approches de références plus récentes et que nous présentons dans ce qui suit, nous ne pouvons pas comparer les performances de nos modéles à ceux de Yarowsky et al. [2001]. Pour cette raison, nous avons construit un modèle de référence (Projection Simple) inspiré du modèle de Yarowsky et al. [2001] (voir tableau 2.1).

Das and Petrov [2011] utilisent aussi un corpus parallèle pour la projection des annotations morpho-syntaxiques de l'anglais vers 8 langues européennes. Ils proposent d'utiliser, en vue de réduire le bruit dû à l'alignement et la projection d'annotations, un graphe de similarité bilingue qui est construit, à partir de liens considérés comme sûrs, entre les deux côtes du corpus parallèle. Les noeuds du graphe sont : des mots individuels du côté source et des trigrammes du côté cible. Les arêtes entre les trigrammes du côté cible sont pondérées avec une fonction de similarité syntaxique basée sur la co-occurrence du mot central du trigramme avec les trigrammes auxquels il est connecté. Pour ce qui est des arêtes entre les deux langues (les mots du côté source et trigrammes du côté cible), elles sont pondérées avec une deuxième fonction de similarité basée sur les scores des alignements automatiques. La projection directe d'annotations, à travers les alignements en mots, ne permet d'annoter qu'une partie du côté cible, le graphe de similarité bilingue est donc utilisé pour propager les annotations des noeuds cibles annotés par projection à d'autres noeuds cibles. Das and Petrov [2011] proposent aussi de construire un dictionnaire des annotations transférées les plus probables pour les trigrammes des mots cibles, puis ce dictionnaire est utilisé pour la réduction du bruit et le filtrage des erreurs de l'annotation par projection. Das and Petrov [2011] ont obtenu ainsi (graphe bilingue et dictionnaire) une précision moyenne de 83.4% sur les 8 langues traitées contre une précision moyenne de 78.8% pour la projection directe (Projection Simple). Duong et al. [2013] ont aussi obtenu des performances similaires à celles de Das and Petrov [2011]. Nous avons évalué nos modèles, en plus du français, [Zennaki et al. 2015b] sur 3 des 8 langues utilisées par [Das and Petrov 2011; Duong et al. 2013], et nous avons obtenu des performances comparables (voir tableau 2.1).

Modèle	**français**	**allemand**	**grec**	**espagnol**
Projection Simple	80.3	78.9	77.5	80.0
(Das, 2011)	—	82.8	**82.5**	84.2
(Duong, 2013)	—	**85.4**	80.4	83.3
(Zennaki, 2015)	**85.6**	82.1	79.9	**84.4**

TABLE 2.1 – Performances en taux d'erreur d'étiquetage de la Projection Simple, Das & Petrov (2011), Duong et al (2013) et Zennaki et al (2015) en étiquettes morpho-syntaxiques universelles.

Contrairement à [Das and Petrov 2011] qui n'utilisent que les alignements avec un dégrée de confiance (score d'alignement) élevé pour réduire le bruit dû aux alignements des mots, Duong et al. [2013] proposent de trier les bi-phrases selon leurs scores d'alignement et de n'utiliser que les n meilleurs pour la projection des annotations morpho-syntaxiques. Les n phrases cibles annotées par projection directe sont utilisées, par la suite, pour l'apprentissage d'un annotateur morpho-syntaxique qu'ils nomment *seed tagger*. Les phrases restantes sont divisées en plusieurs blocs de n bi-phrases, sur ces blocs [Duong et al. 2013] utilisent un processus itératif en 4 étapes : (1) annoter les n phrases cibles du *ième* bloc avec le *seed tagger*, (2) affiner cette annotation avec les annotations obtenues par projection, (3) réapprendre le *seed tagger* sur le *ième* bloc et (4) faire le même procédé avec le bloc suivant jusqu'à l'annotation de tous les blocs.

Annotation en dépendance syntaxique

La projection des dépendances syntaxiques des phrases sources aux phrases cibles à travers les liens d'alignement a été introduite par [Hwa et al. 2005]. Cette méthode a été depuis reprise et améliorée dans de nombreux travaux [McDonald et al. 2011; Ma and Xia 2014; Rasooli and Collins 2015; Lacroix et al. 2016], nous avons évalué nos modèles neuronaux pour l'annotation en dépendance multilingue (voir chapitre 6) sur le même jeu de données que ces approches de l'état de l'art, le tableau 2.2 présente les performances de notre modèle et de ces méthodes d'état de l'art sur 4 langues cibles communes.

Hwa et al. [2005] proposent une méthode semi-supervisée basée sur l'utilisation de plusieurs heuristiques et règles définies manuellement, nous présenterons les principes de cette approche dans la section suivante (section 2.3.1.2).

Modèle	**français**	**allemand**	**espagnol**	**italien**
(McDonald et al., 2011)	73.13	69.77	68.72	70.74
(Ma and Xia, 2014)	67.53	74.30	75.53	77.74
(Rasooli and Collins, 2015)	**79.91**	**74.32**	**78.17**	**79.64**
(Lacroix et al., 2016)	77.92	73.75	76.87	77.82
Notre modèle	75.85	71.98	75.12	72.75

TABLE 2.2 – Performances en pourcentage de mots étant correctement rattachés (–UAS– Unlabeled Attachment Score) d'étiquetage de McDonald et al (2011), Ma and Xia (2014), Rasooli and Collins (2015), Lacroix et al (2016) et notre modèle.

McDonald et al. [2011] proposent, à l'instar de Søgaard [2011], de générer un annotateur délixicalisé en dépendance sur des données annotées en langue source, mais qui sera, dans une étape intermédiaire, re-lexicalisé en utilisant des données cibles annotées en dépendances par projection.

Ma and Xia [2014] utilisent des données parallèles pour transférer des paramètres d'un analyseur en dépendance appris en langue source, ceci afin de construire un analyseur en dépendance pour la langue cible. En utilisant les alignements automatiques des données parallèles, les dépendances sources sont projetées vers le côté cible. Les dépendances cibles sont ainsi pondérées par les poids des dépendances sources correspondantes. Les paramètres de l'analyseur en dépendance cible sont affinés en utilisant des données cibles annotées par un annotateur en dépendance source délexicalisé. L'anglais est utilisé comme langue source, les auteurs appliquent leur approche sur 10 langues cibles.

Les meilleurs résultats pour l'annotation en dépendance par projection interlingue sont ceux obtenus par Rasooli and Collins [2015] (voir tableau 2.2). Rasooli and Collins [2015] proposent une stratégie d'apprentissage qui tire parti des dépendances projetées de haute qualité pour améliorer les résultats des analyseurs en dépendance cibles. Après avoir démontré que la prise en compte des arbres de dépendances incomplets ou partiels (obtenus par projection), durant l'apprentissage des analyseurs en dépendance cibles, dégradait fortement les performances de ces derniers, ils définissent une stratégie d'apprentissage où le modèle est initialement appris exclusivement sur les arbres cibles complets ou denses, puis réappris de façon itérative sur des arbres cibles avec une densité décroissante (mais complétés en utilisant une méthode d'analyse contrainte).

Lacroix et al. [2016] utilisent une méthode similaire à la méthode proposée par Rasooli and Collins [2015]. Ils proposent de filtrer au préalable les arbres de dépendances cibles, obtenus par projection, pour ne garder que les arbres considérés comme fiables. Le corpus cible partiellement annoté (après projection) et filtré est utilisé dans une méthode d'apprentissage partiel pour la construction d'annotateurs en dépendance cibles. Les résultats obtenus par Lacroix et al. [2016] sont décrits dans le tableau 2.2.

Plusieurs autres travaux ont utilisé l'approche de l'apprentissage par transfert (ou transfer learning) afin de proposer des outils de TAL pour les langues peu dotées en données d'apprentissage [Pan et al. 2010]. On peut citer les travaux de Jiang et al. [2015] qui ont appliqué l'apprentissage par transfert pour résoudre le problème d'incompatibilité et de divergence qui peuvent surgir lorsque plusieurs annotations sont exploitées simultanément. Ils ont implémenté plusieurs modèles pour transférer les annotations d'un corpus source au format d'annotation d'un autre corpus cible. L'outil proposé a été évalué sur la segmentation en mots des textes chinois et aussi sur l'analyse en dépendance syntaxique. L'élément clé de leur approche est un classificateur « orienté transfert », qui apprend les régularités de correspondance entre les directives d'annotation à partir d'un corpus parallèle annoté. Ce classificateur comporte deux types d'annotations pour les mêmes données. Dans le même ordre d'idées, Passban et al. [2017] ont mis en oeuvre deux moteurs de traduction, le premier utilise l'approche statistique et le deuxième emploie une approche neuronale, les deux sont entraînés sur une même paire de langues mais ils sont utilisés pour traduire une autre langue. Les deux langues utilisées dans la phase d'entraînement partagent des caractéristique communes (syntaxique et/ou sémantique), ils ont expérimenté le Turc et Azéris/Azerbaïdjanais comme paire de langues. Grâce à l'utilisation de deux moteurs et l'exploitation de la similitude entre la paire de langues, ils ont réussi à développer un modèle fiable pour une langue riche en ressources (Turc), puis ils ont exploité les similitudes interlingues afin d'adapter le modèle pour qu'il fonctionne dans une langue proche avec peu de ressources (Azéris).

2.3.1.2 Méthodes semi-supervisées

Souvent, le transfert entre les langues repose sur l'existence de corpus parallèles et l'utilisation d'outils d'alignement au niveau des mots entre la langue source et la langue cible. Plusieurs outils permettant d'obtenir automatiquement de tels alignements sont disponibles tels que GIZA++ [Och and Ney 2000] qui implémente des modèles statistiques [Brown et al. 1993]. Cependant, les performances des algorithmes d'alignement au niveau des mots ne sont pas toujours satisfaisantes Allauzen and Wisniewski [2010] or cette étape a un impact significatif sur la performance de la projection d'annotations linguistiques [Fraser and Marcu 2007]. Pour pallier ces limitations, des études récentes ont proposé de combiner les étiquettes transférées avec des informations monolingues apprises à l'aide d'une technique partiellement supervisée afin de filtrer les séquences d'étiquettes invalides. Par exemple, Li et al. [2012]; Täckström et al. [2013a]; Wisniewski et al. [2014] ont proposé d'améliorer la performance de la projection en utilisant un dictionnaire d'étiquettes valides pour chaque mot (récupérées à partir de Wiktionary [1]).

Täckström et al. [2013a] ont proposé d'affiner les annotations issues du dictionnaire Wiktionary à l'aide de contraintes de type obtenues par projection interlingue. Le dictionnaire permet de connaître pour chaque mot cible l'ensemble de ses catégories morpho-syntaxiques possibles, ces catégories sont combinées aux annotations obtenues par projection pour réduire l'espace de recherche et donc réduire l'ambiguïté d'annotation. Pour intégrer les deux sources d'information (les étiquettes projetées et le dictionnaire) dans leurs modèles de séquences (HMM et CRF), Täckström et al. [2013a] proposent un ensemble de contraintes de type et une généralisation *ad hoc* de leur critère d'apprentissage.

De même que Täckström et al. [2013a], Wisniewski et al. [2014] proposent de fusionner, dans un premier temps, les annotations obtenues par projection interlingue avec les informations extraites de Wiktionary (associant à chaque mot cible l'ensemble de ses étiquettes possibles) pour annoter automatiquement un corpus d'apprentissage en langue cible, qui sera utilisé, par la suite, avec plusieurs contraintes de type dans un processus d'apprentissage ambiguë à base d'historique.

1. http ://www.wiktionary.org/

Cependant, l'utilisation des contraintes de type lors de l'apprentissage a été vivement critiqué par Pécheux et al. [2015]. Ce détail d'implémentation est donc d'une importance capitale pour obtenir de bonnes performances. Dans leur étude, Pécheux et al. [2015] ont pu démontrer qu'inclure les contraintes de type à l'apprentissage nuit aux performances et donnent comme exemple le finnois et l'indonésien pour lesquels le taux d'erreur est quasiment doublé.

Nous avons proposé dans Zennaki et al. [2015a] d'utiliser un petit corpus annoté en langue cible pour l'adaptation de nos modèles appris par projection. Dans ce scénario semi-supervisé, nous avons utilisé un corpus d'adaptation en allemand de 1000 phrases. Nous avons obtenu une précision sur l'allemand de 90.4% alors que Täckström et al. [2013a] et Wisniewski et al. [2014] ont obtenu respectivement 90.5% et 89.9%.

Hwa et al. [2005] sont les premiers à avoir utilisé la projection interlingue des annotations en dépendance syntaxique. Ils ont proposé une approche basée sur la correspondance directe entre données parallèles, ils estiment que si deux mots en langue source sont respectivement alignés à deux mots en langue cible et s'il existe une dépendance entre les deux mots en langue source, alors cette dépendance existe probablement entre les deux mots en langue cible ; ils définissent donc un ensemble d'actions pour la projection de dépendances selon le type d'alignement (un-à-un, un-à-plusieurs, plusieurs-à-un, un-à-nul ou bien plusieurs-à-plusieurs). En exploitant les alignements au niveau des mots, ils utilisent des opérations et actions prédéfinies combinées à des connaissances capturant les spécificités de la langue source pour projeter les dépendances vers le côté cible. Ils ont évalué leur approche, en utilisant l'anglais comme langue source, sur deux langues cibles l'espagnol et le chinois et ils ont respectivement obtenu 72.1% et 53.9% en UAS (mots étant correctement rattachés). Mais cette approche comporte plusieurs heuristiques et règles qui la rendent difficilement adaptable.

2.3.2 Approches basées sur les plongements lexicaux bilingues de mots

Les plongements lexicaux (word embeddings) sont des représentations vectorielles denses de mots. Ces représentations permettent d'inférer une structure linéaire qui modélise des relations sémantiques et syntaxiques liant les mots en construisant pour chaque mot une

fenêtre de contexte dans laquelle tous les mots sont traités de façon égale. Les plongements lexicaux bilingues sont un moyen très efficace pour le transfert de connaissances d'une langue vers une autre. Cette approche est très prometteuse pour la construction d'outils multilingues pour le TAL [Duong 2017]. La construction de plongements lexicaux bilingues peut être vue comme l'apprentissage de représentations bilingues communes à plusieurs langues.

2.3.2.1 Plongements lexicaux monolingues

Les plongements lexicaux monolingues peuvent être considérés comme une extension des représentations vectorielles standards fondées sur l'hypothèse principale de la sémantique distributionnelle Harris [1954]. Cette hypothèse stipule que "deux mots sont sémantiquement proches s'ils apparaissent dans des contextes similaires". La construction des plongements lexicaux utilise une procédure d'apprentissage dans des modèles prédictifs. Le système apprend à assigner des vecteurs similaires à des mots similaires. Dans la majorité des approches les modèles utilisés sont des réseaux de neurones dont les paramètres sont appris de façon supervisée, à partir de corpus non annotés de très grandes tailles, pour prédire un mot sachant le contexte dans lequel il apparaît [Bengio et al. 2003; Collobert and Weston 2008; Mikolov et al. 2013b; Levy and Goldberg 2014]. Le réseau apprend intrinsèquement à partir de traits latents, en entrée, des représentations compactes, de dimension réduite.

Collobert et al. [2011] ont démontré que les plongements lexicaux permettent de capturer une multitude d'informations d'ordre syntaxique et sémantique très utiles dans de nombreuses tâches du TAL. En conséquence, ces modèles prédictifs ont pu améliorer les performances de plusieurs applications du TAL. On peut citer la compréhension du langage naturel [Collobert and Weston 2008], l'analyse des sentiments [Socher et al. 2013], l'annotation en dépendance syntaxique [Dyer et al. 2015] et la traduction automatique [Bahdanau et al. 2014].

Les méthodes de construction des plongements de mots, qu'on présente ci-dessous, sont fondées sur l'approche pionnière proposée par Bengio et al. [2003]. Cette approche construit les plongements de mots par apprentissage de modèles de langue neuronaux. Collobert and Weston [2008] ont quant à eux utilisé plusieurs tâches en aval, telles que l'annotation

morpho-syntaxique, la reconnaissance des entités nommées, l'analyse syntaxique de surface ainsi que la modélisation du langage, pour la construction de représentations vectorielles de mots (plongements de mots) communes à ces tâches. Plus récemment, Mikolov et al. [2013b] ont proposé l'outil *Word2Vec* permettant d'apprendre des plongements de mots à partir d'une large collection de textes. *Word2Vec* implémente deux modèles : le modèle CBOW (de l'acronyme anglais Continuous Bag-of-Words) et le modèle Skip-Gram. Dans les deux modèles Mikolov et al. [2013b] utilisent une architecture de réseau de neurones simple. Dans CBOW le réseau de neurones construit les plongements de mots en apprenant à prédire un mot sachant son contexte, alors que dans le modèle Skip-Gram le réseau est appris pour prédire le contexte d'un mot en entrée du réseau. La principale critique que l'on peut faire à ces deux modèles est qu'ils n'opèrent que sur un contexte de taille limitée. C'est pour pallier cette limitation que Pennington et al. [2014] ont proposé le modèle GloVe (de l'acronyme anglais Global Vectors for Word Representation) qui utilise directement les statistiques de co-occurrences globales extraites du corpus utilisé pour l'apprentissage.

Plusieurs autres travaux ont proposé d'étendre ces méthodes et techniques à la construction de plongements de phrases, de thésaurus et de documents [Le and Mikolov 2014; Ferret]. Mais la tendance dominante est l'extension des plongements monolingues à la construction de plongements bilingues qu'on présente en détail dans la section suivante.

2.3.2.2 Plongements lexicaux bilingues

D'une autre manière, plusieurs approches récentes fondées sur l'apprentissage d'une représentation interlingue (plongements bilingues) ont été proposées dans le but d'éviter ou de minimiser l'utilisation d'outils d'alignement de mots à partir de corpus parallèles [Durrett et al. 2012; Al-Rfou et al. 2013; Täckström et al. 2013b; Luong et al. 2015; Gouws and Søgaard 2015; Gouws et al. 2015; Ammar et al. 2016]. Ces approches cherchent tout d'abord à extraire des caractéristiques qui soient indépendantes des langues traitées. Ces caractéristiques sont en général matérialisées par des vecteurs représentant le contexte des mots similaires. A partir de ces représentations, le transfert d'annotations en langue source est déduit pour les mots ou fragments cibles les plus proches dans l'espace commun aux deux langues. Afin d'apprendre de telles représentations, plusieurs ressources

peuvent être utilisées telles que des lexiques bilingues [Gouws and Søgaard 2015] et des corpus parallèles [Gouws et al. 2015]. L'utilisation de représentations interlingues a par ailleurs donné lieu à des résultats satisfaisants sur plusieurs tâches du TAL. Par exemple, l'annotation morpho-syntaxique [Gouws and Søgaard 2015], l'annotation sémantique à gros grain [Gouws and Søgaard 2015], la reconnaissance d'entités nommées [Täckström et al. 2012], la classification de documents [Gouws et al. 2015], l'annotation syntaxique [Xiao and Guo 2015; Ammar et al. 2016], l'annotation en rôles sémantiques [Titov and Klementiev 2012] et la traduction automatique [Zou et al. 2013].

Luong et al. [2015] ont proposé le modèle BiSkip (Bilingual Skip-Gram Model) pour la construction de plongements bilingues à partir de corpus parallèles. BiSkip est une extension du modèle monolingue Skip-Gram [Mikolov et al. 2013b] où le contexte d'un mot est étendu pour comporter des liens bilingues obtenus par alignement de mots à partir du corpus parallèle, le modèle BiSkip est entraîné pour prédire un mot sachant son contexte bilingue. Ce modèle a été implémenté dans l'outil BiVec [Luong et al. 2015] et enrichi par la suite dans l'outil MultiVec [Bérard et al. 2016]. Nous avons utilisé MultiVec pour générer les plongements bilingues externes à nos modèles neuronaux (voir section 4.2.3), dans le but d'évaluer leur impact par rapport aux plongements bilingues internes à nos modèles construits à partir de notre représentation multilingue des mots (section 3.1).

Nous proposons d'aborder dans cette thèse la problématique de la projection interlingue d'annotations à partir de corpus parallèles selon les axes suivants dans le prolongement des travaux utilisant des représentations interlingues/multilingues plutôt que des alignements bruités :

1. Nous apprenons en utilisant un réseau de neurones récurrent un espace de représentation multilingue commun à plusieurs langues. En entrée du réseau, notre représentation commune est fondée simplement sur la co-occurrence des mots des langues source et cible d'un corpus parallèle.
2. Nous utilisons cette représentation pour apprendre (à partir d'une annotation linguistique côté source) un annotateur pour des textes en langue cible.

Chapitre 3

Réseaux de neurones récurrents pour l'annotation multilingue

Dans ce chapitre, nous expliquons comment, à partir des limitations des méthodes de l'état de l'art pour la projection interlingue d'annotations linguistiques (présentées dans le chapitre précédent) relatives à l'étape d'alignement mot à mot des phrases du corpus parallèle, nous en sommes venus, dans nos travaux, à nous intéresser à la non prise en compte des informations bruitées issues de cet alignement, mais de représenter ces informations de façon intrinsèque dans l'architecture du réseau de neurones. Nous présentons tout d'abord notre approche pour la construction d'annotateurs multilingues basés sur les réseaux de neurones récurrents. Ensuite, nous proposons plusieurs améliorations de nos modèles neuronaux afin de permettre une meilleure prise en compte des mots hors-vocabulaire ainsi que l'ajout d'informations externes.

Notre contexte expérimental est la construction d'annotateurs multilingues, comprenant des langues faiblement (ou peu) dotées. Pour ce faire, nous supposons disposer de corpus parallèle/multi-parallèle, dont la langue source est richement dotée (disposant d'un annotateur supervisé). Notre idée est d'apprendre un annotateur multilingue en utilisant les réseaux de neurones récurrents sur le côté source pré-annoté du corpus parallèle, puis d'utiliser le réseau appris pour annoter les textes en langue(s) cible(s) du corpus parallèle/multi-parallèle.

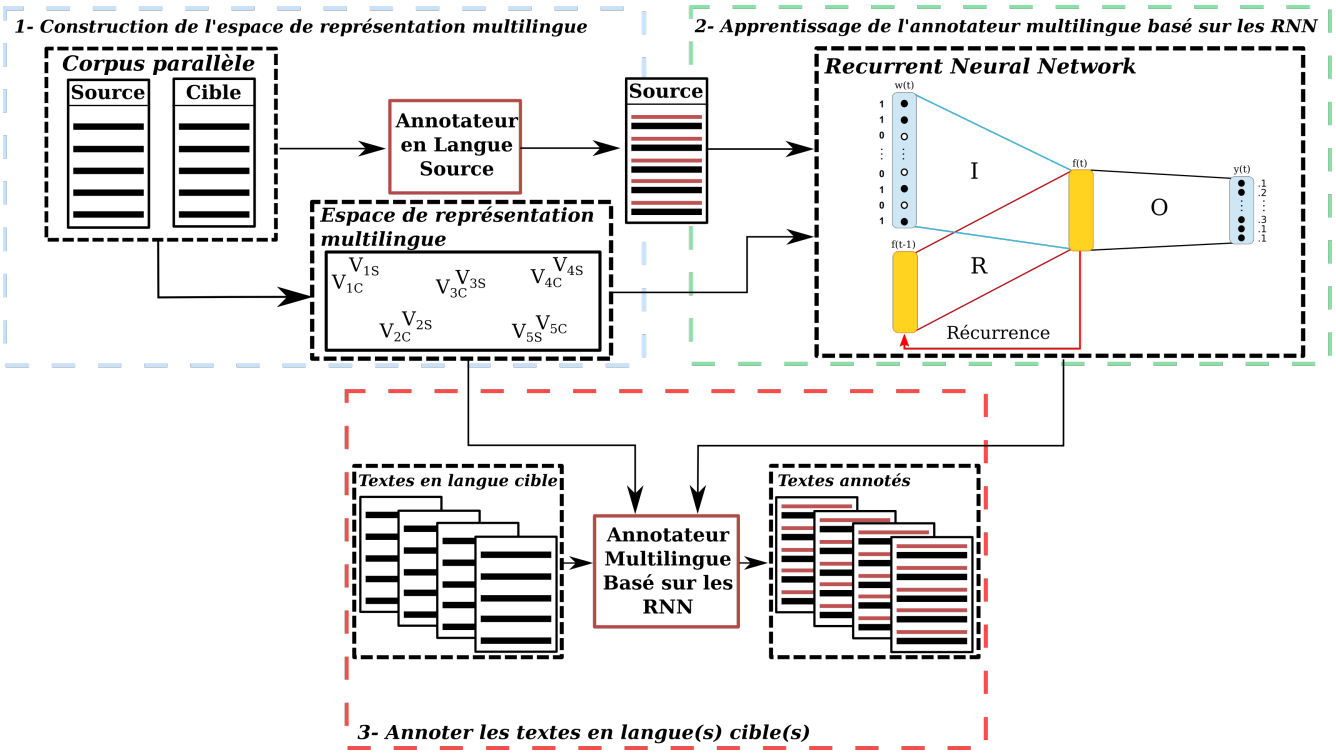

FIGURE 3.1 – Vue d'ensemble de notre approche pour la construction d'annotateurs multilingues basés sur les réseaux de neurones récurrents. Tout d'abord notre espace de représentation multilingue des mots est construit (1). Ces représentations sont ensuite utilisées avec le côté source annoté de corpus parallèle pour l'apprentissage du RNN (2). Le modèle appris est finalement utilisé pour annoter les textes en langue cible (3).

3.1 Approche proposée

Comme décrit dans la section 1.3, les réseaux de neurones sont généralement classés dans deux grandes catégories : les réseaux de neurones *Feed-forward* [Bengio et al. 2006] et les réseaux de neurones Récurrents (on utilise l'acronyme RNN en anglais - pour *Recurrent Neural Networks*) [Mikolov et al. 2010]. [Sundermeyer et al. 2013] ont montré que les modèles de langue statistiques basés sur une architecture récurrente présentent de meilleures performances que les modèles basés sur une architecture *Feed-forward*. Cela vient du fait que les réseaux de neurones récurrents utilisent un contexte de taille non limitée, contrairement aux réseaux *Feed-forward* dont la topologie limite la taille du contexte pris en compte. Cette propriété a motivé notre choix d'utiliser, dans nos expériences, un réseau de neurones

de type récurrent [Elman 1990].

Dans cette section, nous décrivons en détail l'approche proposée pour la construction d'un annotateur multilingue basé sur les RNNs. Notre approche, qui ne nécessite aucune ressource externe, requiert simplement un corpus parallèle et un annotateur pré-existant dans la langue source.

3.1.1 Description de notre approche

Comme le montre la figure 3.1, notre approche comporte trois étapes. Dans la première étape, nous utilisons un corpus parallèle/multi-parallèle entre une langue source bien dotée (riche en corpus annotés) et une ou plusieurs langues cibles moins bien dotées pour construire un espace de représentation multilingue pour les mots (en langues source et cible(s)), dans lequel un mot source et sa traduction cible possèdent des représentations vectorielles proches. La deuxième étape consiste à utiliser cette représentation multilingue combinée avec le côté source du corpus parallèle/multi-parallèle pour apprendre un réseau de neurones récurrent comme annotateur en langue source. Ceci permettrait alors d'utiliser, dans la troisième étape, l'annotateur de type RNN (apprit initialement sur le côté source) combiné avec notre représentation multilingue pour annoter les textes en langue(s) cible(s).

Représentation multilingue des mots : Notre méthode pour la construction de la représentation multilingue des mots est indépendante de toute ressource externe, et est basée uniquement sur un corpus parallèle/multi-parallèle aligné au niveau des phrases seulement et n'applique aucun pré-traitement du type *alignement automatique en mots* qui est, comme dit précédemment, une source d'erreurs et de bruit.

Notre espace de représentation multilingue est construit en associant à chaque mot (source, cible) son empreinte distributionnelle $V_{wi}, i = 1, ..., N$, où N est le nombre de bi-phrases dans le corpus parallèle. Si w apparaît dans la $i^{\text{ème}}$ bi-phrase alors $V_{wi} = 1$; par conséquent, les neurones de la couche d'entrée représentant le mot courant w, sont mis à 0 sauf ceux correspondants aux bi-phrases contenant le mot w, qui sont mis à 1.

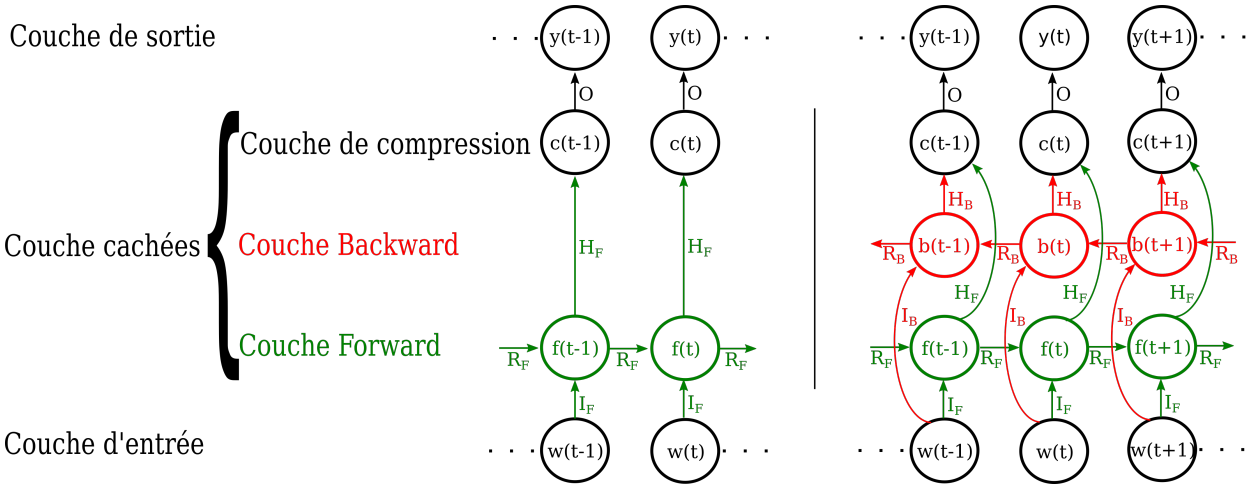

FIGURE 3.2 – Architectures RNN Simple et Bidirectionnelle utilisées dans nos travaux.

Généralement, un mot source et sa traduction cible cooccurrent dans un nombre important de bi-phrases du corpus parallèle, par conséquent leurs empreintes distributionnelles sont proches. Nous pouvons donc utiliser l'annotateur de type RNN, appris initialement sur le côté source, pour annoter les textes en langue(s) cible(s) (grâce à notre *représentation vectorielle multilingue*).

3.1.2 Architectures neuronales utilisées

Dans cette section, nous décrivons les deux architectures de RNN de type Elman, que nous avons utilisées dans nos travaux (voir figure 3.2) : l'architecture *Simple* (SRNN) et l'architecture *bidirectionnelle* (BRNN).

Le caractère récurrent des modèles RNN d'Elman réside dans le fait que, la connexion de récurrence se situe au niveau de la couche cachée.

3.1.2.1 RNN simple

Dans un RNN *simple* de type Elman (SRNN), à chaque temps t la couche cachée au temps *t-1* est présentée en plus de l'entrée courante du réseau. Cette connexion permet au SRNN de capturer l'historique précédent (contexte passé), et non pas juste les $n-1$

précédentes entrées, ce modèle peut, théoriquement, représenter des contextes de tailles importantes.

L'architecture du SRNN que nous avons considérée dans nos travaux est illustrée dans la figure 3.2. Nous proposons un modèle composé d'une succession de quatre couches de neurones : une couche d'entrée au temps t, une première couche cachée notée $f(t)$ (on utilise l'appellation Forward en anglais - pour couche du contexte passé), une deuxième couche cachée $c(t)$ (nommée aussi couche de compression), et une couche de sortie $y(t)$. Chaque neurone de la couche d'entrée est relié à tous les neurones de la couche cachée par les matrices des poids I_F et R_F. Les neurones des deux couches cachées sont connectés entre eux par la matrice des poids H_F. La matrice des poids O connecte tout neurone de la deuxième couche cachée à chaque neurone de la couche de sortie.

Dans notre modèle, la couche d'entrée est formée par la concaténation de la représentation vectorielle $w(t)$ du mot courant dans notre espace de représentation multilingue (les neurones de la couche d'entrée représentant le mot courant w, sont mis à 0 sauf ceux correspondants aux bi-phrases contenant le mot w, qui sont mis à 1), et de la première couche cachée (couche Forward) au temps précédent $f(t-1)$, ce qui confère au réseau neuronal son aspect récurrent.

Par ailleurs, nous utilisons deux couches cachées (des expériences préliminaires ont montré que ceci permet d'obtenir de meilleures performances), avec des tailles variables (de 80 à 1024 neurones). Pour la fonction d'activation, nous utilisons la fonction *sigmoïde*. Nous pensons que ces couches cachées devraient permettre de capturer intrinsèquement des informations d'alignement au niveau des mots.

Les valeurs en sortie de notre modèle, sachant l'entrée w et $f(t-1)$, sont calculées comme suit :

— Propagation en avant de l'entrée pour calculer la valeur de la couche Forward :

$$f(t) = \Sigma(w(t).I_F(t) + f(t-1).R_F(t)) \tag{3.1}$$

— Calculer la valeur de la couche de compression :

$$c(t) = \Sigma(f(t).H_F(t)) \tag{3.2}$$

— Calculer la sortie du réseau (prédiction du modèle) y(t) :

$$y(t) = \Gamma(c(t).O(t)) \tag{3.3}$$

où Σ est une fonction d'activation de type *sigmoïde* :

$$\Sigma(z) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

et Γ est une fonction de type *softmax* :

$$\Gamma(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \tag{3.5}$$

Pour une meilleure annotation linguistique, il serait bénéfique de considérer, en plus du contexte passé (contenu dans la couche *Forward*), le contexte futur. Nous pouvons donc affirmer que notre SRNN n'est pas optimal pour l'annotation linguistique, puisque le réseau ignore l'information future, et tente à optimiser la prédiction en sortie sachant uniquement le mot courant et le contexte passé. De ce fait, le réseau SRNN est pénalisé par rapport à l'approche de projection standard (notre référence - voir section 4.1) fondée sur l'étiqueteur TnT [Brants 2000] qui prend en compte les contextes gauche et droit du mot courant. L'extension de l'architecture SRNN pour la prise en compte du contexte droit est nommée architecture bidirectionnelle (BRNN acronyme en anglais - pour Bidirectional Recurrent Neural Network) [Schuster and Paliwal 1997].

3.1.2.2 RNN bidirectionnelle

L'architecture bidirectionnelle utilise en même temps les contextes passé et futur mémorisés par les deux couches forward et backward, comme indique dans la figure 3.2.

Le fonctionnement du BRNN, décrit dans l'article de référence sur les réseaux neuronaux bidirectionnels [Schuster and Paliwal 1997], est donc comme suit :

— Le sous-réseau backward est utilisé pour prédire les états futurs de la couche backward.

— Les états de la couche backward sont initialisés avec les valeurs obtenues précédemment pour la phase forward.

— Le modèle global parcourt la séquence en avant pour calculer les valeurs des couches de compression c(t) eq(3.6) et de sortie y(t) eq(3.3) :

$$c(t) = \Sigma(f(t).H_F(t) + b(t).R_B(t)) \tag{3.6}$$

3.1.3 Construction de nos modèles neuronaux -Algorithme d'apprentissage-

Tout d'abord, avant l'apprentissage du modèle, quelques étapes de pré-traitement sont nécessaires. Celles-ci sont appliquées sur notre corpus d'apprentissage (corpus parallèle source/cible) et sur notre corpus de validation en langue source :

— Étiqueter le côté source du corpus parallèle/multi-parallèle et le corpus de validation (avec l'annotateur supervisé disponible en langue source).

— Construire notre espace de représentation multilingue (représentations vectorielles communes basées sur les empreintes distributionnelles des mots source et cible) à partir du corpus parallèle/multi-parallèle [1].

L'apprentissage de notre modèle neuronal est un processus itératif sur le corpus d'apprentissage (côté *source* annoté) par descente de gradient stochastique basée sur l'algorithme de **R**étro-**P**ropagation (**RP**) du gradient de l'erreur [Rumelhart et al. 1985] (voir section 3.1.3.1). Cependant, pour un apprentissage plus efficace, l'algorithme de **R**étro-**P**ropagation du gradient de l'erreur à **T**ravers le **T**emps (**RPTT**) [Rumelhart et al. 1985] (voir section 3.1.3.2) peut être utilisé pour rétro-propager le gradient de l'erreur dans le temps à travers la matrice de récurrence R_F, ce qui permet au modèle d'optimiser l'information stockée dans la couche cachée, de façon à ce que cette information soit la plus explicite et utile à l'avenir. Dans [Mikolov et al. 2011], les auteurs ont pu démontrer une dégradation des performances des réseaux récurrents appris juste avec la **RP**.

L'algorithme 1 présenté ci-dessous décrit une époque d'entraînement du réseau.

1. Il est important de noter que si ce corpus parallèle change - par exemple si de nouvelles données sont disponibles - les représentations vectorielles pourront être soit conservées à l'identique soit mises à jour (en augmentant la taille du vecteur) avant le ré-apprentissage du RNN.

Algorithme 1 : Apprentissage d'un annotateur multilingue basé sur un RNN Simple

1 : Initialiser les matrices des poids du réseau avec une distribution normale $\mathcal{N}(0, 0.1)$.

2 : Initialiser le compteur du temps $t = 0$, et initialiser l'état des neurones de la couche cachée $f(t)$ à 1.

3 : Incrémenter le compteur du temps t de 1.

4 : Présenter le vecteur représentant le mot $w(t)$ dans la couche d'entrée.

5 : Recopier l'état de la couche cachée $f(t-1)$ dans la couche d'entrée.

6 : Calculer les valeurs des couches cachées $f(t)$ et $c(t)$ puis la valeur de la couche de sortie $y(t)$.

7 : Calculer l'erreur de prédiction $e_0(t) = d(t) - y(t)$ (différence entre la sortie attendue et la sortie prédite).

8 : Mettre à jour les matrices des poids O, H_f et I_f avec l'algorithme de rétropropagation (**RP**) du gradient de l'erreur.

9 : Si t est multiple d'une période P alors mettre à jour la matrice des poids de récurrence R_f avec l'algorithme de la rétro-propagation du gradient de l'erreur à travers le temps (RPTT).

Sinon ne rien faire.

10 : Si le corpus d'apprentissage comporte encore des exemples, alors revenir à 3.

Les matrices des poids du réseau sont mises à jour en utilisant l'erreur de prédiction pondérée par un pas d'apprentissage α, initialement fixé à 0.1.

Après chaque itération (époque), le corpus de validation est annoté en utilisant le réseau de neurones appris jusque-là. Les sorties sont comparées aux sorties de l'annotateur supervisé, pour calculer le taux d'erreur d'annotation du réseau. Si le taux d'erreur diminue d'une époque à une autre, le pas d'apprentissage reste inchangé et l'apprentissage continue durant une nouvelle époque. Sinon, le pas d'apprentissage est diminué de moitié au début de

la nouvelle époque. La convergence du modèle est obtenue si le taux d'erreur d'annotation ne diminue plus d'une époque à une autre. Généralement le réseau converge en 5 à 10 époques.

L'apprentissage a pour but de maximiser la fonction objectif suivante :

$$f(\lambda) = \sum_{t=1}^{T} log\ y_{\ell t}(t) \tag{3.7}$$

Tels que les exemples d'apprentissage (mots du corpus d'apprentissage), sont étiquetés avec $t = 1 \ldots T$, et ℓt représente l'index de la prédiction correcte sachant les t mots précédents, finalement λ représente les paramètres du réseau, dont fait partie les matrices des poids qu'on cherche à apprendre en utilisant la RP et la RPTT, qu'on présente dans ce qui suit.

Comme présenté précédemment, le modèle neuronal ainsi appris est utilisé pour annoter les textes *cible*, via l'utilisation de notre espace de représentation multilingue. Il est important de noter que dans le cas d'un corpus parallèle multilingue (multi-parallèle), le même modèle neuronal pourra annoter toutes les langues *cibles* sans être re-entraîné. On dispose donc d'un véritable annotateur multilingue.

3.1.3.1 Rétro-Propagation du gradient de l'erreur

Connaissant la sortie attendue (sortie désirée) et si toutes les fonctions d'activation utilisées sont dérivables, tout réseau de neurones peut être entraîné par Rétro-Propagation du gradient de l'erreur. Pour chaque mot $\{w(t), t = 1, ..., T\}$, faisant partie du corpus d'apprentissage (côté source du corpus parallèle annoté), l'inférence du réseau est lancée, on obtient en sortie du réseau $y(t)$ l'annotation prédite du mot en entrée ($w(t)$). Le corpus d'apprentissage étant préalablement annoté, on dispose de la sortie attendue du réseau $d(t)$ (annotation exacte du mot $w(t)$). La mise à jour des matrices des poids du réseau par la rétro-propagation du gradient de l'erreur s'effectue comme suit :

— Calcule de l'erreur de prédiction (erreur en sortie du réseau) $e_O(t)$:

$$e_O(t) = d(t) - y(t) \tag{3.8}$$

— Mise à jour de la matrice des poids O entre la seconde couche cachée $c(t)$ et la couche de sortie $y(t)$:

$$O(t+1) = O(t) + c(t)e_O(t)^T\alpha \tag{3.9}$$

— Pour pouvoir mettre à jour la matrice des poids H_f entre les deux couches cachées, l'erreur en sortie $e_O(t)$ est rétro-propagée vers la seconde couche cachée :

$$e_c(t) = e_O(t)^T V(t)c(t)(1-c(t)) \tag{3.10}$$

— Mise à jour de la matrice des poids H_f avec l'erreur rétro-propagée $e_c(t)$:

$$H_f(t+1) = H_f(t) + f(t)e_c(t)^T\alpha \tag{3.11}$$

— Pour pouvoir mettre à jour les matrices des poids I_f et R_f entre la couche d'entrée et la première couche cachée, l'erreur rétro-propagée $e_c(t)$ est à nouveau propagée vers la première couche cachée :

$$e_f(t) = e_f(t)^T H_f(t)f(t)(1-f(t)) \tag{3.12}$$

— Mise à jour des matrices des poids I_f et R_f avec l'erreur rétro-propagée $e_f(t)$:

$$I_f(t+1) = I_f(t) + w(t)e_f(t)^T\alpha \tag{3.13}$$

$$R_f(t+1) = R_f(t) + f(t-1)e_f(t)^T\alpha \tag{3.14}$$

où α est le pas d'apprentissage, initialement fixé à 0.1.

La rétro-propagation du gradient de l'erreur (RP) tente à optimiser la prédiction en sortie sachant le mot courant ($w(t)$) et la couche de récurrence (première couche cachée) à l'état précèdent ($f(t)$), mais aucun effort n'est fait pour optimiser l'information du contexte (historique) stockée dans la couche de récurrence, de façon à ce que cette information du contexte soit la plus explicite et utile à l'avenir. Pour cela, la RP est conjuguée avec la rétro-propagation du gradient de l'erreur à travers le temps (RPTT).

3.1.3.2 Rétro-Propagation du gradient de l'erreur à travers le temps

La RPTT consiste à rétro-propager récursivement l'erreur $e_f(t)$ (obtenue au niveau de la première couche cachée $f(t)$ au temps t) vers les couches de récurrences aux temps précédents $\{f(t-\tau), \tau = 1, ..., P\}$, où P est le pas de récurrence (fixé dans nos expérimentations entre 5 et 10 pas de temps). Il est important de noter que la RPTT nécessite de sauvegarder l'état de la première couche cachée et des matrices de récurrence R_f des P précédentes étapes. P peut donc s'apparenter à la taille de l'historique (contexte gauche du mot $w(t)$) qu'on souhaite prendre en compte.

L'erreur $e_f(t)$ est récursivement rétro-propagée par :

$$e_f(t-\tau-1) = e_f(t-\tau)^T R_f(t-\tau) f(t-\tau)(1-f(t-\tau)) \tag{3.15}$$

La matrice des poids I_f et la matrice des poids de récurrence R_f (voir figure 3.2), sont mises à jour comme suit :

$$I_f(t+1) = I_f(t) + \sum_{\tau=0}^{P} w(t-\tau) e_f(t-\tau)^T \alpha \tag{3.16}$$

$$R_f(t+1) = R_f(t) + \sum_{\tau=0}^{P} f(t-\tau-1) e_f(t-\tau)^T \alpha \tag{3.17}$$

3.2 Améliorations de nos modèles neuronaux

3.2.1 Traitement des mots hors-vocabulaire

Dans notre approche la gestion des mots inconnus (OOV en anglais – pour *out-of-vocabulary*) est pour l'instant quasi-inexistante. En effet, comme les mots inconnus ne figurent pas dans le corpus d'apprentissage (corpus parallèle/multi-parallèle utilisé pour la construction de notre espace de représentation multilingue) leurs représentations vectorielles sont nulles (ne contiennent que des O), et pour les annoter durant la phase de test, le réseau de neurones n'utilise que l'information de récurrence ce qui est une information insuffisante pour une bonne annotation.

Afin de mieux annoter les mots inconnus, nous utilisons le modèle CBOW (Continuous Bag Of Word), modèle décrit dans [Mikolov et al. 2013b]. Le modèle CBOW : a pour objectif, en se basant sur une fenêtre de mots adjacents (contexte), de prédire un mot selon son contexte. Nous utilisons CBOW pour remplacer chacun des mots inconnus dans un contexte donné par le mot connu qui lui est le plus proche dans ce même contexte, puis nous utilisons, en entrée du RNN, comme représentation vectorielle (au lieu de la représentation vide) du mot inconnu la représentation vectorielle du mot connu le plus proche.

3.2.2 Nouvelles variantes de RNN pour l'ajout d'informations externes

Dans la section précédente, nous avons défini la façon dont notre modèle neuronal est construit pour l'annotation linguistique. Nous avons néanmoins réalisé une modification spécifique de notre modèle neuronal, avec la prise en compte d'informations linguistiques externes de bas niveau pour la construction d'annotateurs plus complexes. En particulier, nous proposons d'intégrer des annotations morpho-syntaxiques dans notre architecture neuronale pour l'apprentissage non supervisé d'annotateurs sémantiques multilingues à gros grain (annotation en *SuperSenses*). Cette tâche d'annotation en *SuperSenses* prend une importance grandissante dans plusieurs applications du TAL. Cependant, à notre connaissance, il n'y a pas de travaux sur cette tâche pour le français.

L'idée sous-jacente à cette modification (intégration des traits morpho-syntaxiques dans notre modèle neuronal) est d'utiliser les annotations morpho-syntaxiques (annotations basiques) pour lever une partie de l'ambiguïté relative à l'annotation en *SuperSenses* en vue d'une meilleure adaptation de notre modèle à la tâche d'annotation en *SuperSenses*. Nous pensons que l'intégration des annotations morpho-syntaxiques devrait permettre d'améliorer les performances de notre modèle neuronal.

Nous proposons trois architectures neuronales pour l'intégration des annotations morpho-syntaxiques à différents niveaux de représentation. Comme spécifié dans la figure 3.3, l'intégration peut se faire soit au niveau de la couche d'entrée soit au niveau de la première couche cachée ou bien au niveau de la deuxième couche cachée. Dans ces trois modèles,

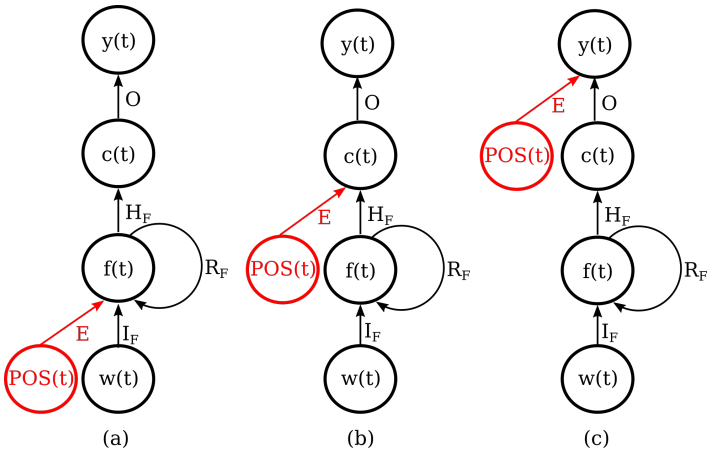

FIGURE 3.3 – Architectures SRNN avec prise en compte de l'annotation morpho-syntaxique (POS) du mot courant w(t) sur trois niveaux ; de gauche à droite : (a) au niveau de la couche d'entrée, (b) au niveau de la première couche cachée, (c) au niveau de la deuxième couche cachée.

l'annotation morpho-syntaxique du mot en cours $w(t)$ est représentée par un vecteur $(POS(t))$ de même taille que l'ensemble des annotations morpho-syntaxiques considérées (nous utilisons le jeu de 12 étiquettes morpho-syntaxiques universelles définis dans [Petrov et al. 2012] - voir section 1.2.1), toutes les composantes du vecteur sont égales à zéro, sauf la composante correspondant à l'indice de l'annotation morpho-syntaxique du mot en cours qui est égal à 1.

3.2.3 Combinaison des modèles basées RNN et projection interlingue standard

Notre approche s'appuie sur une hypothèse forte qui n'est pas toujours vérifiée : que l'ordre des mots entre langues source et cible est conservé. Selon le couple de langues considérées, les mots traduits en langue cible peuvent avoir un ordre différent de l'ordre des mots de la phrase source. Par exemple, pour le couple de langues anglais-français, l'ordre des mots d'une phrase nominale est le plus souvent inversé, car en français, un complément qui définit un mot le suit plus souvent qu'il ne le précède, par ex. *the European Commission* se traduirait par *la Commission européenne*.

Dans le but de lever la contrainte liée à la non-conservation de l'ordre des mots d'une

langue à l'autre, nous proposons la combinaison de notre modèle RNN avec l'approche de projection standard (notre référence - voir section 4.1).

Le modèle *M1*, basé sur l'approche de projection standard, et le modèle *M2*, basé RNN, utilisent des stratégies différentes pour l'annotation des mots (TnT [Brants 2000] est fondé sur les modèles de Markov alors que le RNN est un modèle neuronal), Il serait donc judicieux de combiner les deux approches, du fait de leur complémentarité, afin d'allier les avantages de chaque approche, en vue d'obtenir les meilleures performances possibles.

La probabilité d'annoter le mot courant w est obtenue par :

$$P_{M12}(t|w) = (\mu P_{M1}(t|w, C_{M1}) + (1 - \mu)P_{M2}(t|w, C_{M2})) \tag{3.18}$$

Où, C_{M1} et C_{M2} sont, respectivement les contextes de w considérés par *M1* et *M2*. Le paramètre d'interpolation μ (importance de chaque modèle) est ajusté par validation croisée sur le corpus de test. Le mot courant w est annoté avec l'étiquette la plus probable, en utilisant la fonction f :

$$f(w) = \arg\max_{t}(P_{M12}(t|w)) \tag{3.19}$$

3.3 Conclusion

Dans ce chapitre nous avons présenté notre approche pour la construction d'annotateurs multilingues. Nous avons choisi d'utiliser, du fait qu'ils peuvent exploiter un contexte de taille non limitée, des réseaux de neurones récurrents simple et bidirectionnelle.

Nous avons proposé plusieurs améliorations de notre approche pour une meilleure prise en compte des mots inconnus et pour l'intégration d'informations externes dans nos modèles.

Nous présentons dans les chapitres suivants l'application de notre approche, dans un cadre multilingue, sur trois tâches du TAL :

— Annotation morpho-syntaxique.

— Annotation sémantique à gros grain (SuperSense).

— Annotation en dépendances syntaxiques.

Chapitre 4

Annotateur morpho-syntaxique multilingue fondé sur les réseaux de neurones récurrents

L'annotation morpho-syntaxique consiste à attribuer à chaque unité lexicale du corpus une étiquette apportant certaines informations linguistiques (catégorie gramaticale, genre, nombre, etc.). Avant de présenter l'évaluation de notre approche pour la création d'annotateurs morpho-syntaxiques multilingues basés sur les réseaux de neurones récurrents (RNN), nous décrivons tout d'abord l'approche par projection simple à laquelle nous allons nous comparer (et qui sera aussi combinée — au cours des expériences qui vont suivre — avec la méthode que nous proposons).

4.1 Annotateur morpho-syntaxique non supervisé par projection simple - notre référence

L'approche pour construire notre étiqueteur morpho-syntaxique non supervisé par projection simple (décrit par l'algorithme 2) est très proche de celle introduite par [Yarowsky et al. 2001]. Cette approche, qui a été réutilisée plus récemment par [Duong et al. 2013], correspond à l'état de l'art des annotateurs morpho-syntaxiques non supervisés. Ces auteurs utilisent l'alignement automatique en mots (obtenu à partir d'un corpus parallèle) pour projeter les annotations de la langue source vers la langue cible, en vue de construire des annotateurs morpho-syntaxiques pour la langue cible.

L'algorithme 2 est décrit dans l'encadré ci-dessous :

Algorithme 2 : Méthode de référence par projection d'annotations selon un alignement automatique en mots

---

1 : Annoter le côté source du corpus parallèle.

2 : Aligner automatiquement le corpus parallèle en utilisant GIZA++ ou un autre outil d'alignement en mots.

3 : Projeter les annotations directement pour les alignements 1-1.

4 : Pour les correspondances N-1, projeter l'annotation du mot se trouvant à la position N/2 arrondi à l'entier supérieur.

5 : Annoter les mots non-alignés avec l'étiquette la plus fréquente qui leur est associée dans le corpus.

6 : Apprendre un analyseur morpho-syntaxique à partir de la partie cible du corpus désormais annotée (par exemple, dans notre cas, nous utilisons TnT tagger [Brants 2000].

4.2 Évaluation de notre approche pour la construction l'annotateur morpho-syntaxique multilingue

4.2.1 Corpus et outils

Initialement, nous avons expérimenté notre approche sur le couple de langues anglais-français, où le français est considéré comme langue cible. Le français n'est certainement pas une langue faiblement dotée, mais le fait qu'il dispose d'un annotateur morpho-syntaxique supervisé (TreeTagger [Schmid 1995]), nous a permis de construire une *pseudo vérité terrain* (sur le corpus de test) pour évaluer notre approche. Nous avons utilisé un corpus d'apprentissage de 10000 bi-phrases, extrait du corpus parallèle (anglais-français) ARCADE II [Veronis et al. 2008], dont le côté source a été annoté par l'outil *TreeTagger* [Schmid 1995] pour l'anglais. Notre corpus de validation (en anglais - pour le règlage du RNN) contient 1000 phrases (non présentes dans le corpus d'apprentissage), et est aussi extrait du corpus ARCADE II puis annotées par le toolkit *TreeTagger* pour l'anglais. Nous avons

construit notre corpus test (français) à partir de 1000 phrases extraites du corpus ARCADE II, et annoté par le toolkit *TreeTagger* pour le français, puis corrigées manuellement.

Ayant obtenu des résultats intéressants sur le couple de langues anglais-français, nous nous sommes ensuite intéressés à la généralisation de notre approche pour d'autres langues : l'allemand, le grec et l'espagnol. Afin de pouvoir rendre nos résultats comparables avec ceux de [Das and Petrov 2011; Duong et al. 2013; Gouws and Søgaard 2015], nous suivons leur protocole : nous partons de l'anglais comme langue source et utilisons un corpus parallèle contenant 65, 000 bi-phrases et un corpus de validation (anglais) de 10, 000 phrases extraits d'Europarl [Koehn 2005]. Nous évaluons les résultats de nos approches sur les mêmes corpus de test, qui sont ceux des campagnes d'évaluation d'analyse en dépendances CoNLL [Buchholz and Marsi 2006]. Ces corpus ont été annotés manuellement par des experts linguistes. Nous utilisons aussi la même métrique d'évaluation (le taux d'erreur d'étiquetage) et le même jeu d'étiquettes (*Universal Tagset* [Petrov et al. 2012]).

Pour l'apprentissage de nos modèles basés sur les RNN, les parties source (anglais) du corpus d'apprentissage (ARCADE II pour le couples de langue anglais–français et Europarl pour les autres couples de langues) ainsi que le corpus de validation sont annotés par le toolkit TreeTagger pour l'anglais.

Afin de construire nos modèles par projection simple (Algorithme 2 - voir section 4.1), la partie cible des corpus d'apprentissage est étiquetée par projection des annotations du côté source (annoté par le toolkit TreeTagger pour l'anglais) en utilisant les alignements obtenus par GIZA++ [Och and Ney 2000], la partie cible du corpus d'apprentissage ainsi annoté est utilisée pour l'apprentissage d'un annotateur morpho-syntaxique en langue cible basé sur TnT [Brants 2000]. Par souci d'uniformité, nous avons aussi transformé les étiquettes morpho-syntaxiques fines (de TreeTagger et de CoNLL) en leurs équivalents dans le jeu étiquettes universelles via les règles de correspondances proposées dans Petrov et al. [2012].

Pour tirer parti des avantages de chacun de ces deux modèles *M1* (Projection Simple) et *M2* (RNN), il est intéressant d'explorer leur combinaison. La combinaison des deux modèles pour chacune des langues considérées (équation 3.18) est estimée par validation croisée sur le corpus de test. Le mot w est annoté avec l'étiquette t_w la plus probable, en

utilisant la fonction f donnée par l'équation 3.19

Afin de déterminer la pertinence de notre représentation multilingue de mots décrite dans la section 3.1.1, nous avons aussi exploré l'impact qu'aurait l'utilisation d'une représentation basée sur les plongements bilingues/multilingues de mots (multi-lingual words embeddings) en entrée de notre modèle RNN. Pour ce faire, nous avons utilisé une approche de l'état de l'art, implémentée dans l'outil MultiVec [Bérard et al. 2016], pour la construction de plongements bilingues de mots de façon générique (ne sont pas spécifiques à la tâche traitée). Nous avons réalisé l'apprentissage de l'outil MultiVec, pour la construction des plongements bilingues de mots, en utilisant le corpus parallèle Europarl entre l'anglais et chacune des langues cibles que nous avons considérées dans nos expérimentations.

4.2.2 Adaptation du modèle neuronal sur la langue cible

L'un des atouts important des méthodes utilisant les réseaux de neurones réside dans leur possibilité d'adaptation. Notre modèle RNN multilingue (décrit dans la section 3.1) est appris uniquement sur le côté source du corpus parallèle, il est par conséquent non supervisé pour les langues cibles. Dans la section 3.2.3 nous avons proposé, en vue de lever la contrainte liée à la divergence inter-langue, de combiner notre approche RNN avec l'approche de projection standard (notre référence - voir section 4.1), dans cette section nous proposons d'adapter notre modèle RNN multilingue sur un corpus d'adaptation en langue cible (de taille modeste et manuellement étiqueté). Le principe de ce processus d'adaptation est de construire, à partir de notre modèle non supervisé, un modèle légèrement supervisé capable de capturer les spécificités linguistiques relatives à la langue cible. Le processus d'adaptation est résumé comme suite (les étapes 1 et 2 correspondent à la construction de modèle non supervisé) :

1. Associer à chaque mot w (appartenant aux vocabulaires des langues source et cible) une représentation vectorielle dans notre espace de représentation multilingue.
2. Le côté source du corpus parallèle est annoté (avec l'annotateur morpho-syntaxique supervisé disponible) et associer à notre espace de représentation multilingue pour l'apprentissage de notre annotateur morpho-syntaxique multilingue basé RNN (Algorithme 1 - voir section 3.1.3).

Modèle \ Langue		français		allemand		grec		espagnol	
		All	OOV	All	OOV	All	OOV	All	OOV
Réf.	Projection Simple	80.3	77.1	78.9	73.0	77.5	72.8	80.0	79.7
	SRNN MultiVec	75.0	65.4	70.3	68.8	71.1	65.4	73.4	62.4
Notres	SRNN	78.5	70.0	76.1	76.4	75.7	70.7	78.8	72.6
	BRNN	80.6	70.9	77.5	76.6	77.2	71.0	80.5	73.1
	BRNN - OOV	81.4	77.8	77.6	77.8	77.9	75.3	80.6	74.7
Comb.	Proj. + SRNN	84.5	78.8	81.5	77.0	78.3	74.6	83.6	81.2
	Proj. + BRNN	85.2	79.0	81.9	77.1	79.2	75.0	84.4	81.7
	Proj. + BRNN - OOV	**85.6**	**80.4**	82.1	**78.7**	79.9	**78.5**	**84.4**	**81.9**
SOTA	(Das, 2011)	—	—	82.8	—	**82.5**	—	84.2	—
	(Duong, 2013)	—	—	**85.4**	—	80.4	—	83.3	—
	(Gouws, 2015)	—	—	84.8	—	—	—	82.6	—

TABLE 4.1 – Performances en taux d'étiquetage correct (mots hors vocabulaire) (Projection Simple, SRNN avec utilisation des plongements bilingues MultiVec en entrée, SRNN, BRNN, BRNN-OOV avec traitement des OOV et Projection+RNN) - et comparaison avec les approches de l'état de l'art (SOTA) Das & Petrov (2011), Duong et al (2013) et Gouws & Søgaard (2015) sur All (tous les mots du vocabulaire) et sur OOV (mots hors-vocabulaire).

3. Poursuivre l'apprentissage du modèle RNN (adaptation avec un apprentissage légèrement supervisé) sur un petit corpus en langue source (manuellement annoté).

Cette approche est particulièrement souhaitable dans un scénario itérative, le modèle non supervisé est utilisé pour annoter des textes en langue cible qui seront post-édités pour produire rapidement un corpus d'adaptation de bonne qualité pour l'apprentissage du modèle légèrement supervisé.

4.2.3 Analyse des résultats

Nous reportons dans le tableau 4.1 les performances obtenues pour l'annotation morpho-syntaxique non-supervisée en langue cible. On note que les performances des modèles basés sur les RNN bidirectionnels (BRNN) sont meilleurs comparés aux performances des modèles basés sur les RNN simples (SRNN), ce qui signifie que la prise en compte des contextes passé et futur dans les BRNN permet une meilleure annotation par rapport à la prise en compte du contexte passé seul dans les SRNN.

Le tableau 4.1 présente aussi les performances de nos modèles neuronaux avant et après

la prise en compte des mots hors vocabulaire (OOV), méthode que nous avons présentée dans la section 3.2.1. De façon plus précise, il apparaît que si on remplace le mot inconnu dans un contexte donné par le mot connu qui lui est le plus proche dans le même contexte, améliore significativement les performances de nos modèles.

Nous avons évalué plusieurs topologies de réseaux de neurones récurrents simple et bidirectionnel, avec une ou deux couches cachées, et avec différentes tailles. Les meilleures performances sont celles des annotateurs basés sur des réseaux de neurones à deux couches cachées, contenant respectivement 640 et 160 neurones (RNN-640-160). Ces performances sont proches des annotateurs par projection simple. Nous avons également combiné l'approche classique avec notre méthode par réseaux récurrents (voir section 3.2.3). Les résultats expérimentaux de notre combinaison (Projection+RNN) sont proches de ceux de l'état de l'art des annotateurs morpho-syntaxiques non supervisés [Das and Petrov 2011; Duong et al. 2013] (qui ont utilisé la totalité du corpus Europarl, alors que nous avons utilisé un sous ensemble de 65,000 bi-phrases) et [Gouws and Søgaard 2015] (qui ont utilisé des ressources externes telles que Wiktionary et Wikipedia). Les résultats obtenus montrent une bonne complémentarité entre projection simple et RNN. Il est intéressant de noter que les modèles RNN sont plus performants en utilisant notre représentation multilingue des mots (3.1.1) que en utilisant les plongements bilingues MultiVec.

Il est aussi important de noter qu'un seul modèle SRNN ou BRNN est appliqué sur l'allemand, le grec et l'espagnol ; nous avons vraiment construit des modèles multilingues !

Anglais	*Financial* situation of the European Parliament.
Français	*Finances* de la commission européenne.
SRNN	*Finances* / VERB ...
BRNN	*Finances* / **NOUN** ...

TABLE 4.2 – Effet de l'architecture bidirectionnelle.

Afin de connaître dans quelle mesure la prise en compte du contexte droit (architecture bidirectionnelle) améliore les performances des modèles RNN, nous avons analysé le corpus de test français. Dans l'exemple fourni dans le tableau 4.2, l'ambiguïté liée à l'annotation

du mot *Finances* est levée par la prise en compte du contexte droit (les annotations des mots – de la commission européenne –) dans le modèle BRNN.

Modèle \ Taille Corpus Allemand	0	100	500	1k	2k	5k	7k	10k
RNN Non-supervisé + Adaptation sur l'allemand	76.1	**82.1**	**87.3**	**90.4**	90.7	91.2	91.4	92.4
RNN supervisé pour l'allemand	—	71.0	76.4	82.1	90.6	93.0	94.2	95.2
TnT supervisé pour l'allemand	—	80.5	86.5	89.0	92.2	94.1	95.3	95.7
RNN supervisé + TnT supervisé pour l'allemand	—	81.0	86.7	90.1	**94.2**	**95.3**	**95.7**	**96.0**

TABLE 4.3 – Modèle légèrement supervisé sur la langue cible (Allemand) : Effet de la taille du corpus d'adaptation (annoté manuellement) sur notre méthode décrite dans la Section 4.2.2 (RNN Non-supervisé + Adaptation appris sur le côté Anglais du corpus Europarl et adapté sur l'allemand). 0 signifie la non utilisation de corpus allemand durant l'apprentissage.

Nous reportons dans le tableau 4.3 les résultats obtenus après l'adaptation de notre modèle RNN Non-supervisé, l'adaptation est réalisée sur un corpus de taille croissante (allons de 100 à 10000 phrases). Nous nous concentrons uniquement sur l'allemand comme langue cible. Nous avons comparé ces résultats avec deux modèles supervisés basés sur TnT et RNN. Les modèles supervisés sont appris uniquement sur le corpus d'adaptation.

Les résultats montrent que l'adaptation de notre modèle RNN non supervisé, est efficace dans des contextes de ressources très faibles (< 1000 phrases en langage cible). Lorsque plus de données sont disponibles (> 1000 phrases), les approches supervisées commencent à être plus performantes. On remarque que les modèles RNN et TnT restent complémentaires, leur combinaison a montré une amélioration de la précision de l'annotation.

La figure 4.1 représente les résultats obtenus, avec les mêmes méthodes sur les mots inconnus (OOV) seulement. L'examen de ces résultats montre clairement la limitation de notre modèle RNN Non-supervisé + Adaptation pour le traitement des mots inconnus. Cela vient du fait que les vecteurs de représentation des mots en entrée du RNN restent inchangés durant la phase d'adaptation du modèle (provient du corpus d'apprentissage initial)

En cas de divergence d'ordre des mots, nous avons observé que notre modèle peut encore gérer une certaine divergence, notamment dans les cas suivants :

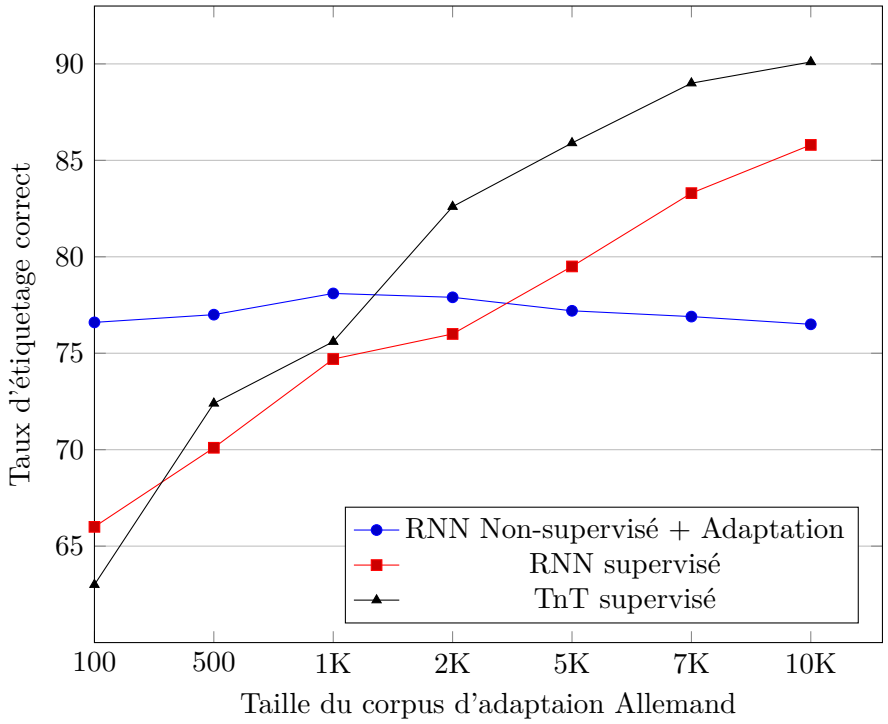

FIGURE 4.1 – Précision sur le mots inconnus par rapport à la taille du corpus d'adaptation/apprentissage en allemand des modéles RNN Non-supervisé + Adaptation, RNN supervisé et TnT supervisé.

— De toute évidence, si l'étiquette du mot courant à annoter est non ambiguë (dans le cas de l'ordre ADJ et NOUN de l'anglais vers le français - voir tableau 4.4), alors l'information de contexte (historique RNN) n'a aucun effet.

— Lorsque le contexte est erroné (du fait que l'ordre des mots pour le corpus de test en langue cible est différent du corpus d'apprentissage en langue source), l'étiquette correcte du mot peut être récupérée en utilisant la combinaison (RNN + Projection interlingue - voir tableau 4.5)

Treetagger Supervisé Anglais	... other/ADJ specific/ADJ groups/NOUN ...
RNN Non-supervisé Français	... autres/ADJ groupes/NOUN spécifiques/ADJ ...

TABLE 4.4 – Divergence dans l'ordre des mot de l'anglais vers le français -étiquette non ambiguë-.

Treetagger Supervisé Anglais	... two/NUM local/ADJ groups/NOUN ...
RNN Non-supervisé Français	... deux/NUM groupes/NOUN locaux/**NOUN** ...
Projection + RNN Français	... deux/NUM groupes/NOUN locaux/ **ADJ** ...

TABLE 4.5 – Divergence dans l'ordre des mot de l'anglais vers le français -étiquette ambiguë-.

4.2.4 Bilan

Pour résumer, les expériences décrites dans ce chapitre nous permettent de valider notre approche pour la construction d'annotateurs morpho-syntaxiques multilingues. Les résultats obtenus montrent l'efficacité de la prise en compte du contexte droit (BRNN) ainsi que la prise en compte des mots hors vocabulaires (OOV), ce qui permet d'améliorer les résultats de nos modèles. L'utilisation des plongements de mots (obtenus avec MultiVec) externes à nos modèles dégrade leurs performances. Il nous semble que la raison principale de cette baisse est le fait que les plongements sont appris indépendamment de la tâche traitée. Le scénario d'apprentissage semi-supervisé que nous avons proposé permet à notre modèle d'être plus efficace, cela vient du fait que l'adaptation sur la langue cible du modèle multilingue, appris sur la langue source, permet une meilleure prise en compte de la divergence linguistique entre les langues source et cible. Les résultats obtenus sur la tâche d'annotation morpho-syntaxique sont encourageants et nous amènent, afin de valider notre approche, à explorer d'autres tâches du TAL. Nous présentons dans les chapitres suivants nos travaux pour la construction d'annotateurs multilingues en SuperSenses (voir chapitre 5) et d'analyseurs multilingues en dépendances syntaxiques (voir chapitre 6).

Chapitre 5

Annotateur en SuperSenses multilingue fondé sur les réseaux de neurones récurrents

Nous proposons dans ce chapitre une amélioration de notre approche (basée sur les réseaux de neurones récurrents) avec la prise en compte d'informations linguistiques de bas niveau pour la construction d'annotateurs linguistiques plus complexes. Ainsi, nous démontrons que l'intégration des annotations morpho-syntaxiques dans notre modèle neuronal est utile pour la construction d'annotateurs sémantiques à gros grains (*SuperSenses*) multilingues. Cette tâche d'annotation en *SuperSenses* prend une importance grandissante dans plusieurs applications du TAL. Cependant, à notre connaissance, il n'y a pas de travaux sur cette tâche pour le français. La réalisation d'un tel système pour le français est donc un autre contribution. Nos expériences portent sur la projection d'annotations de l'anglais vers le français et l'italien.

5.1 Annotation en SuperSenses

L'annotation en SuperSenses (SuperSenses Tagging – SST –) est une tâche du TAL qui consiste à annoter chaque unité du texte, avec un jeu d'étiquettes sémantiques générales définies par les catégories lexicographiques de WordNet (*SuperSenses*). Elle peut être vue comme une tâche à cheval entre la reconnaissance d'entités nommées (REN) et la désambiguïsation lexicale (Word Sense Disambiguation – WSD) : tout en étant une extension

de la REN, elle est une simplification de la WSD.

Reconnaissance d'entités nommées : Issue de la recherche d'information, la reconnaissance d'entités nommées (REN) consiste à rechercher des objets textuels (i.e. un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc. La REN fait sans doute partie des tâches les plus étudiées du TAL et apparaît en effet comme fondamentale pour diverses applications telles que l'analyse de contenu, la recherche et l'extraction d'information, les systèmes de question-réponse, le résumé automatique, etc. [Ehrmann 2008]. Néanmoins, cette tâche reste limitée à la reconnaissance de catégories générales en omettant des unités informatives (catégories plus fines) potentiellement importantes dans certains contextes applicatifs.

Désambiguïsation lexicale : Le processus de désambiguïsation lexicale consiste à sélectionner les sens corrects d'instances contextualisées de mots ambigus, parmi l'ensemble de leurs sens possibles (ou sens candidats). Cette tâche, est primordiale pour tout système du TAL, telles que : la traduction automatique, la recherche d'informations, la reconnaissance de la parole ou l'analyse grammaticale. La REN peut prendre part à un processus de désambiguïasation lexicale en tant qu'information sémantique. Concrètement, les entités nommées sont utilisées comme « filtre » au niveau des restrictions de sélection (ou sous-catégorisation sémantique) des sens d'une unité lexicale parmi plusieurs dizaines de milliers de sens très spécifiques. Ces sens sont inventoriés dans des dictionnaires, incluant les entités nommées. Un tel degré de granularité (très fin), rend la distinction entre les sens trop subtile, pour être capturée automatiquement de façon robuste. De plus, l'apprentissage automatique de modèles nécessite des données manuellement annotées en sens, difficiles à obtenir.

WordNet : Le *Princeton WordNet* [Fellbaum 1998], est probalement la resource la plus utilisée pour la désambiguïsation lexicale (Word Sense Disambiguation – WSD). Il est organisé autour de la notion d'ensemble de synonymes (synsets) décrits par une partie du discours (nom, verbe, adjectif, adverbe), une définition et leurs liens (hyperonyme, hyponyme, antonyme, etc.). Chaque sens d'un item lexical (entrée) correspond à un synset. La version courante du Princeton WordNet (version 3.0) comprend 155 287 items lexicaux

pour un total de 117 659 synsets.

Annotation sémantique à gros grain (*SuperSenses Tagging*) : La complexité à modéliser et à traiter l'ambiguïté lexicale ainsi que les limitations de la REN ont fait émerger des tâches d'annotation lexico-sémantique (sens lexical) à gros grain. Ce type de tâche présente plusieurs avantages. Par exemple, l'intérêt d'utiliser les annotations en sens à gros grain, afin de lever l'ambiguïté lexicale des mots, a été souligné dès l'apparition des premiers réseaux lexico-sémantiques [Peters et al. 1998]. [Kohomban and Lee 2005] ont proposé d'utiliser les catégories lexicographiques du *Princeton WordNet* comme amorce à l'annotation avec des sens plus spécifiques (à grain fin).

L'annotation en *SuperSenses* (SuperSenses Tagging – SST –) est l'une de ces tâches alternatives, elle a été introduite par [Ciaramita and Johnson 2003], comme une étape de désambiguïsation intermédiaire pour la réalisation de la WSD. Ces mêmes auteurs ont proposé d'utiliser un perceptron structuré entrainé et évalué sur le SemCor Corpus [Miller et al. 1993]. [Ciaramita and Johnson 2003] ont appelé « SuperSenses » les 41 catégories lexicographiques du *Princeton WordNet* [Fellbaum 1998]. Les catégories WordNet sont vues comme la principale base pour la construction d'un ensemble de catégories plus exhaustives en vue de mieux annoter les concepts occurant dans une phrase. Un autre avantage de l'utilisation de ces catégories est leur universalité (elles sont communes à plusieurs langues), ce qui permet leur utilisation pour des tâches telles que la traduction automatique ou la recherche et l'extraction d'informations multilingues. L'annotation en *SuperSenses* a été aussi utilisée comme première étape pour plusieurs tâches, telles que la désambiguïsation en sens [Ye and Baldwin 2007], le réordonnancement des hypothèses d'un analyseur syntaxique [Collins and Koo 2005], l'annotation de tweets [Johannsen et al. 2014], et la détection de métaphores [Tsvetkov et al. 2013].

En dehors de l'anglais, [Schneider et al. 2012] ont construit un corpus annoté en SuperSense extrait de Wikipedia Arabe.

Les étiquettes *SuperSenses* : Le jeu d'étiquettes *SuperSenses* original comprend au total 41 sens, répartis en deux catégories : 26 étiquettes pour représenter les sens des noms et 15 autres pour représenter les sens des verbes, plus une étiquette unique (catch-all) pour les autres unités (adjectifs, adverbes, etc.). [Ciaramita and Johnson 2003] présentent

ces étiquettes en détail (voir tableau 1.3).

5.2 Annotateur en SuperSenses multilingue fondé sur les réseaux de neurones récurrents

5.2.1 Corpus et outils

Notre méthode ne nécessite qu'un corpus parallèle (ou multi-parallèle) entre une langue source bien dotée (riche en corpus annotés en sens) et une ou plusieurs langues cibles moins bien dotées. Un corpus parallèle peut être obtenu soit par construction manuelle (corpus non bruité) ou en utilisant un système de traduction automatique (corpus bruité). Nous étudions l'impact de la qualité du corpus parallèle sur notre approche en utilisant des traductions manuelles ou automatiques du corpus *SemCor* [Miller et al. 1993] (corpus anglais annoté en sens).

Pour tester notre approche dans un cadre multilingue, nous l'évaluons sur l'italien et le français considérés comme langues cibles. La seule source de données utilisée dans les deux cas est le corpus parallèle multilingue MultiSemCor (MSC), résultant des traductions du [Miller et al. 1993] (corpus anglais annoté en sens) de l'anglais vers l'italien (manuellement et automatiquement) et vers le français (automatiquement).

SemCor : Le SemCor [Miller et al. 1993] [1] est un sous-ensemble du Corpus de Brown [Kucera and Francis 1979]. Sur les 700 000 mots de ce dernier, environ 230 000 sont annotés avec des *synsets* du Princeton *WordNet*.

MultiSemCor : Nous disposons du MultiSemCor Anglais/Italien (MSC-IT-1) construit par traduction manuelle du SemCor Anglais vers l'italien [Bentivogli and Pianta 2005] [2]. Afin d'étudier l'influence de la qualité du corpus parallèle (traduction manuelle/automatique) sur notre système, nous traduisons aussi automatiquement le SemCor Anglais vers l'italien pour obtenir un MutliSemCor Anglais/Italien (MSC-IT-2) construit automatiquement. Ne disposant pas de MultiSemcor Anglais/Français construit manuellement, nous utilisons deux traductions du SemCor vers le français provenant de deux systèmes de

1. http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor
2. http://multisemcor.fbk.eu/

traduction distincts : - MultiSemcor Anglais/Français (MSC-FR-1) [3] obtenu avec le système de traduction statistique anglais-français basé sur la boîte à outils *Moses* [Hoang and Koehn 2008] mis au point par le Laboratoire d'Informatique de Grenoble [Besacier et al. 2012] ; - MultiSemcor Anglais/Français (MSC-FR-2) obtenu avec le système de traduction statistique anglais-français en ligne (Google Traduction [4]).

Corpus d'apprentissage : Puisque le SemCor est annoté avec les *synsets* de Princeton *WordNet* et que l'apprentissage de nos modèles se fait sur des *SuperSenses*, nous avons réalisé une conversion des annotations du SemCor des synsets WordNet vers les *SuperSenses*. Nous avons ensuite appris nos systèmes sur les différentes versions du MultiSemCor (décrites précédemment), avec du côté source le SemCor annoté en *SuperSenses*.

Corpus d'évaluation : Pour estimer les performances de nos modèles et pouvoir les comparer avec des modèles existants, nous utilisons le corpus d'évaluation de la tâche 12 (désambiguïsation lexicale multilingue) de Semeval 2013 [Navigli et al. 2013], qui est un corpus d'évaluation traduit en 5 langues (anglais, français, allemand, italien et espagnol) pour lequel nous utilisons les textes italien et français. Cependant, ces textes de la campagne d'évaluation Semeval 2013 (tâche 12) ont été annotés en sens issus de *BabelNet* [Navigli and Ponzetto 2012], nous avons donc tout d'abord réalisé une conversion des sens BabelNet vers les synsets WordNet, puis les synsets WordNet ont été convertis vers les *SuperSenses*.

5.2.2 Systèmes évalués

Les objectifs de nos expérimentations sont les suivants : d'une part, nous souhaitons valider notre modèle neuronal sur une nouvelle tâche (l'annotation en *SuperSenses* RNN-SST), et d'autre part, nous souhaitons évaluer l'apport de la prise en compte des annotations morpho-syntaxiques dans notre modèle neuronal pour l'annotation en *SuperSenses* (RNN-SST-POS) ; enfin, nous voulons évaluer l'influence de la qualité des corpus parallèles (traduction manuelle ou automatique) sur les performances de nos modèles neuronaux (RNN-SST et RNN-SST-POS) et du modèle basé sur la projection interlingue d'annotations.

En résumé :

3. `http://getalp.imag.fr/static/wsd/TALN2015/ressources/frenchSemcor3.0.zip`
4. `https://translate.google.com/`

— nous avons construit quatre annotateurs en *SuperSenses* basés sur la projection interlingue (Projection Simple), méthode que nous avons présentée dans la section 4.1, en utilisant les corpus MultiSemCor (MSC-IT-1, MSC-IT-2, MSC-Fr-1, et MSC-FR-2) décrits dans la section 5.2.1.

— nous avons aussi utilisé les corpus MultiSemCor pour construire des espaces de représentations multilingues anglais-italien-français, dans ces espaces, chaque mot de la langue source (anglais) possède une représentation unique, du fait qu'on utilise le même côté source commun aux quatre MultiSemCor (SemCor), alors que chacun des mots cibles (italien et français) possède deux représentations distinctes.

— nous avons utilisé le SemCor anglais annoté en *SuperSenses* pour l'apprentissage de notre modèle neuronal multilingue RNN-SST, via nos espaces de représentation multilingues.

— nous avons annoté le SemCor anglais en utilisant *TreeTagger* (analyseur morpho-syntaxique proposé par [Schmid 1995]), puis avons utilisé le SemCor ainsi annoté, pour l'apprentissage de trois modèles neuronaux RNN-SST-POS, correspondants aux trois architectures proposées dans la section 3.2.2 :

  — RNN-SST-POS-In : ajout de l'annotation morpho-syntaxique au niveau de la couche d'entrée.

  — RNN-SST-POS-H1 : ajout de l'annotation morpho-syntaxique au niveau de la première couche cachée.

  — RNN-SST-POS-H2 : ajout de l'annotation morpho-syntaxique au niveau de la deuxième couche cachée.

5.2.3 Analyse des résultats

Comme mentionné précédemment, nos modèles présentés ci-dessus ont été évalués sur les corpus d'évaluations de Semeval 2013 (tâche 12 - désambiguïsation lexicale - WSD), ce qui nous permet une comparaison avec les systèmes ayant participé à la tâche 12 de Semeval 2013, mais cette comparaison est indirecte du fait que ces systèmes n'ont pas été construits spécialement pour l'annotation en *SuperSenses* mais pour la désambiguïsation lexicale en utilisant les sens BabelNet (nous avons donc opéré une conversion des sens

	Modèle \ Lang.	Italien		Français	
		MSC-IT-1 trad man.	**MSC-IT-2 trad. auto**	**MSC-FR-1 trad. auto**	**MSC-FR-2 trad auto.**
Réf.	Projection Simple	61.3	45.6	42.6	44.5
Nos SST Basés RNN	SRNN	59.4	46.2	46.2	47.0
	BRNN	59.7	46.2	46.0	47.2
	SRNN-POS-In	61.0	47.0	46.5	47.3
	SRNN-POS-H1	59.8	46.5	46.8	47.4
	SRNN-POS-H2	63.1	48.7	47.7	49.8
	BRNN-POS-In	61.2	47.0	46.4	47.3
	BRNN-POS-H1	60.1	46.5	46.8	47.5
	BRNN-POS-H2	63.2	48.8	47.7	50.0
	BRNN-POS-H2 - OOV	64.6	49.5	48.4	50.7
Combinaison	Proj. + SRNN	62.0	46.7	46.5	47.4
	Proj. + BRNN	62.2	46.8	46.4	47.5
	Proj. + SRNN-POS-In	62.9	47.4	46.9	47.7
	Proj. + SRNN-POS-H1	62.5	47.0	47.1	48.0
	Proj. + SRNN-POS-H2	63.5	49.2	48.0	50.1
	Proj. + BRNN-POS-In	62.9	47.5	46.9	47.8
	Proj. + BRNN-POS-H1	62.7	47.0	47.0	48.0
	Proj. + BRNN-POS-H2	63.6	49.3	48.0	50.3
	Proj. + BRNN-POS-H2 - OOV	**64.7**	49.8	48.6	51.0
S-E	MFS Semeval 2013	60.7		**52.4**	
	GETALP (Schwab *et al.* 2012)	40.2		34.6	

TABLE 5.1 – Performances en taux d'étiquetage correct (Projection Simple, RNN-SST, RNN-SST-POS, Combinaisons Projection+RNN) - et une comparaison directe avec deux modèles Semeval 2013 et une autre indirecte avec le Modèle BARISTA sur le Danois.

BabelNet vers les *SuperSenses*). Nous avons identifié deux systèmes provenant de cette tâche, se rapprochant le plus de notre cadre expérimental :

— **MFS Semeval 2013** : Système de référence fourni par SemEval 2013 pour la tâche de WSD, ce système utilise l'heuristique du sens WordNet le plus fréquent (MFC - *Most Frequent Sense*) et constitue une *baseline* solide (car elle fait appel aux sens issus d'une base lexicale multilingue, BabelNet, tandis que nos méthodes tentent d'induire ces sens à partir de l'anglais).

— **GETALP** : Système non-supervisé pour la WSD proposé par [Schwab et al. 2012] basé sur un algorithme à colonies de fourmis.

Nous précisons également que deux autres systèmes ont participé à cette tâche, le système DAEBAK ! [Navigli and Lapata 2010] et le système UMCC-DLSI [Gutiérrez Vázquez et al. 2011]. Contrairement à notre système et aux systèmes GETALP et MFS Semeval 2013, les systèmes DAEBAK ! et UMCC-DLSI bénéficiaient de ressources externes et de la richesse des informations de BabelNet. Notre cadre expérimental non supervisé se rapprochant plus des systèmes GETALP et MFS Semeval 2013, nous pensons qu'il est plus judicieux de nous comparer aux performances de ces deux systèmes [5].

Le modèle BARISTA (*Bilingual Adaptive Reshuffling with Individual Stochastic Alternatives*) [Gouws and Søgaard 2015] est un modèle fondé sur l'utilisation d'une liste de paires de mots (mot source-traduction cible ou bien mot source-mot cible portant la même information linguistique) pivots pour la construction d'une représentation distribuée bilingue. Cette représentation est apprise en utilisant le modèle neuronal CBOW [6] sur un corpus bilingue dans lequel les mots pivots d'une langue sont remplacés aléatoirement par les mots correspondants dans l'autre langue.

Par souci d'exhaustivité, nous avons choisi de reporter, dans notre tableau des résultats, les performances sur l'annotation en *SuperSenses* en danois du modèle BARISTA, malheureusement, nous n'avons pas pu évaluer nos modèles sur le danois (corpus d'évaluation non disponible).

Le tableau 5.1 présente les performances en taux d'étiquetage correct de nos annotateurs en *SuperSenses* neuronaux (avec ou sans prise en compte des annotations morpho-syntaxiques), des annotateurs en *SuperSenses* fondés sur une projection interlingue, de la référence MFS Semeval 2013, du système GETALP de Semeval 2013 et du modèle BARISTA. De ces résultats, nous pouvons faire les observations suivantes :

— Nos modèles obtiennent les meilleurs résultats sur l'italien en utilisant le corpus non bruité MSC-IT-1 (traduction manuelle du SemCor vers l'italien).

— La qualité du corpus parallèle influe significativement sur les performances de nos modèles : en utilisant des corpus bruités (traduction automatique du SemCor vers l'italien et le français) les performances de la totalité de nos modèles se dégradent ;

5. DAEBAK ! et UMCC-DLSI ont obtenus pour la tâche d'annotation en SuperSenses respectivement 68.1% et 72.5% sur l'italien, et 59.8% et 67.6 % sur le français

6. `https://code.google.com/p/word2vec/`

cela se vérifie également, à moindre mesure, d'une traduction automatique à l'autre (le corpus MSC-FR-1 obtenu par Google Traduction est moins bruité que le corpus MSC-FR-2 obtenu par le système Moses).

— Les meilleurs performances sont celles des modèles résultant de la combinaison entre projection simple et RNN, ce qui montre la complémentarité de ces deux approches.

— Dans le cas des corpus bruités, l'approche neuronale semble plus robuste que l'approche référence par projection simple (*baseline*).

— Le modèle neuronal le plus performant est RNN-SST-POS-H2, ce qui démontre que la prise en compte des annotations morpho-syntaxiques au niveau de la deuxième couche est la plus pertinente ; une prise en compte tardive des informations morpho-syntaxiques semble donc préférable pour un annotateur sémantique à gros grain neuronal.

Le tableau 5.2 illustre un exemple où la prise en compte de l'annotation morpho-syntaxique au niveau de la deuxième couche cachée de notre annotateur neuronal en SupserSenses (RNN-SST-POS-H2) a permis de lever l'ambiguïté lié au mot *conseiller*, annoté par le RNN-SST comme *verb.communication* alors que l'annotation exacte est *noun.person.*

Français	... qui est également conseiller sur le climat pour le Mexique.
RNN-SST	... conseiller/**verb.communication** ...
RNN-SST-POS-H2	... conseiller/**noun.person** ...

TABLE 5.2 – Effet de la prise en compte des annotations morpho-syntaxiques.

5.2.4 Bilan

Nous avons démontré que la prise en compte des annotations morpho-syntaxiques dans notre architecture neuronale améliore les performances de notre modèle sur l'annotation en *SuperSenses*. Nous avons ainsi montré la faisabilité et la généricité de l'approche sur deux langues cibles : l'italien et le français. Par ailleurs, nous avons aussi pu évaluer l'impact de la qualité du corpus parallèle sur notre approche (corpus obtenu par traductions manuelles ou automatiques).

Chapitre 6

Analyseur multilingue en dépendances syntaxique fondé sur les réseaux de neurones récurrents

L'analyse en dépendances syntaxique est une étape importante dans de nombreux processus du traitement automatique de la langue tels que la traduction automatique, l'analyse sémantique, etc. La construction d'analyseur multiligue en dépendances syntaxique constitue, en termes de complexité de réalisation par rapport à l'annotation morpho-syntaxique et à l'annotation en SuperSenses, la suite logique de nos travaux.

Tandis que les annotations morpho-syntaxiques et SuperSenses opèrent au niveau des mots en fournissant des informations sur la catégorie syntaxique ou sémantique de chaque mot de la phrase, l'annotation en dépendance syntaxique est utilisée en vue de prédire la structure sous-jacente d'une phrase.

Dans cette section, nous commençons par présenter le principe de l'analyse en dépendances syntaxique par transition (*Arc-Eager*) [Nivre 2003] puis nous présentons l'adaptation de notre modèle pour l'analyse syntaxique par transition en expliquant en quoi le système *Arc-Eager* est adapté à notre approche.

6.1 Analyse syntaxique en dépendances par transition

L'analyse syntaxique en dépendances par transition vise à établir des relations binaires entre les mots d'une phrase. Formellement, une dépendance est une relation du type

gouverneur-dépendant ayant un rôle syntaxique entre deux mots de la phrase. Chaque mot de la phrase est gouverné par un unique autre mot (excepté la racine de la phrase). L'ensemble de ces relations entre les mots d'une phrase donnée forme une structure de dépendances.

D'un autre côté, l'analyse en dépendances syntaxique par transition consiste donc, à partir d'une phrase donnée, à construire une structure de dépendances correcte pour cette phrase, basée sur une bonne séquence de transitions.

Le principe est de parcourir la phrase, en appliquant des transitions (actions) à des configurations (états), pour extraire les liens syntaxiques (dépendances) existant entre les mots de cette phrase, tels que par exemple les relations sujet-verbe ou verbe-objet.

Un système par transition est donc équivalant à un automate dans lequel les configurations décrivent l'état de l'analyse (position dans la phrase et avancement dans la construction de la structure de dépendances) et les transitions permettent de passer d'une configuration à une autre. Procéder à une analyse équivaut à chercher un chemin dans l'automate à partir de la configuration initiale jusqu'à une configuration finale. Chercher la bonne analyse revient alors à chercher le chemin permettant d'atteindre la configuration finale contenant la bonne structure de dépendances [Lacroix 2014].

La structure de dépendances, nommée arbre en dépendances, est construite de façon incrémentale, l'analyseur parcourt la phrase de gauche à droite et ajoute des dépendances par application de transitions (actions). Les actions sont propres au système par transition employé pour l'analyse.

Un système par transition est défini par un quadruplet $\langle T, C, c_0, Ct \rangle$ où :

— T est un ensemble de transitions ;

— C est un ensemble de configurations ;

— $c_0 \in C$ est la configuration initiale ;

— $C_t \subset C$ est l'ensemble des configurations finales.

Une configuration est définie par un triplet $\langle \sigma, \beta, A \rangle$ où :

— σ est une pile de mots (déjà traités ou partiellement traités) ;

— β est une mémoire tampon contenant les mots en attente de traitement ;

— A est un ensemble d'arcs correspondant aux dépendances de la structure finale.

Un arc (dépendance) est représenté par un triplet de la forme $(j; l; i)$ où j est l'indice du gouverneur, i est l'indice du dépendant et l est l'étiquette de la dépendance.

Dans le système par transition Arc-Eager, que nous avons utilisé dans nos travaux et que nous présentons ici, l'ensemble des transitions T contient quatre actions possibles :

— **LEFT-ARC** (arc gauche) ajoute une dépendance gauche entre le mot du haut de la pile σ et le premier mot de la mémoire tampon β et supprime le mot du haut de la pile ;

— **RIGHT-ARC** (arc droit) ajoute une dépendance droite entre le mot du haut de la pile σ et le premier mot de la mémoire tampon β et ajoute le premier mot de la mémoire tampon en haut de la pile ;

— **SHIFT** (décalage) dépile un mot de la mémoire tampon β pour l'empiler sur la pile σ ;

— **REDUCE** (reduction) supprime le mot du haut de la pile σ si celui-ci est rattaché par une dépendance.

Le tableau 6.1 présente formellement les conditions d'application et les effets des transitions du modèle Arc-Eager sur les configurations. Les transitions LEFT-ARC et RIGHT-ARC permettent d'ajouter des dépendances gauches et droites si les mots dépendants ne sont pas déjà rattachés par d'autres dépendances. Lors de l'application de la transition LEFT-ARC (voir tableau 6.1), une dépendance gauche, ayant comme gouverneur le mot w_j (premier mot de la mémoire tampon β) et comme dépendant le mot w_i (mot en haut de la pile σ), est ajoutée dans l'ensemble des dépendances A, le mot w_i est supprimé et le mot w_j est déplacé en haut de la pile σ. Lorsqu'une dépendance droite est ajoutée, ayant comme gouverneur le mot w_i et comme dépendant le mot w_j (premier mot de β), le mot w_j est déplacé en haut de la pile σ. Le mot n'est pas supprimé lors de l'application de la transition Right-Arc car il peut encore avoir des dépendants à droite. La transition Reduce permet de supprimer le mot du haut de la pile si celui-ci est rattaché par une dépendance et Shift fait passer le premier mot de β sur σ.

Comme présenté ci-dessus, une configuration inclut une pile σ, une mémoire tampon β permettant de sauvegarder les mots à traiter et une structure de donnée A afin de

Actions	**Effets sur les configurations**	**Conditions**
LEFT-ARC	$(\sigma \mid w_i, w_j \mid \beta, A) \Rightarrow (\sigma, w_j \mid \beta, A \cup \{(j,i)\})$	$i \neq 0 \wedge \neg\exists k(k,i) \in A$
RIGHT-ARC	$(\sigma \mid w_i, w_j \mid \beta, A) \Rightarrow (\sigma \mid w_i w_j, \beta, A \cup \{(i,j)\})$	$\neg\exists k(k,j) \in A$
REDUCE	$(\sigma \mid w_i, \beta, A) \Rightarrow (\sigma, \beta, A)$	$\exists k(k,i) \in A$
SHIFT	$(\sigma, w_i \mid \beta, A) \Rightarrow (\sigma \mid w_i, \beta, A)$	

TABLE 6.1 – Définition des transitions du système *arc-eager*.

mémoriser les dépendances trouvées, elle est de la forme : $< \sigma, \beta, A >$. L'analyse d'une phrase $W = w_1...w_n$ donnée commence par l'affectation de cette phrase à la configuration initiale du système $c_0 =< [w_0], [w_1...w_n], \emptyset >$ où w_0 est la racine artificielle (dans la structure de dépendances résultante tous les mots sont rattachés directement ou indirectement à w_0), le passage d'une configuration à une autre se fait par l'application d'une des transitions possibles du système (voir tableau 6.1), le but étant de prédire les transitions qui permettront de passer d'une configuration à une autre jusqu'à obtenir une configuration finale dans laquelle est contenue la structure de dépendances correcte pour la phrase à analyser. Une configuration finale sera alors de la forme $< [w_0], [], A >$. L'analyse est alors l'application d'une suite de transitions de la configuration initiale à la configuration finale.

Nous présentons dans le tableau 6.2 un exemple d'application de ce système. La structure de dépendances pour la phrase «Le nouveau député écologiste défendra nos droits au parlement.» est donnée dans la figure 6.1. La séquence de transitions permettant d'obtenir cette structure par la méthode d'analyse par transition *arc-eager* est donnée dans le tableau 6.2.

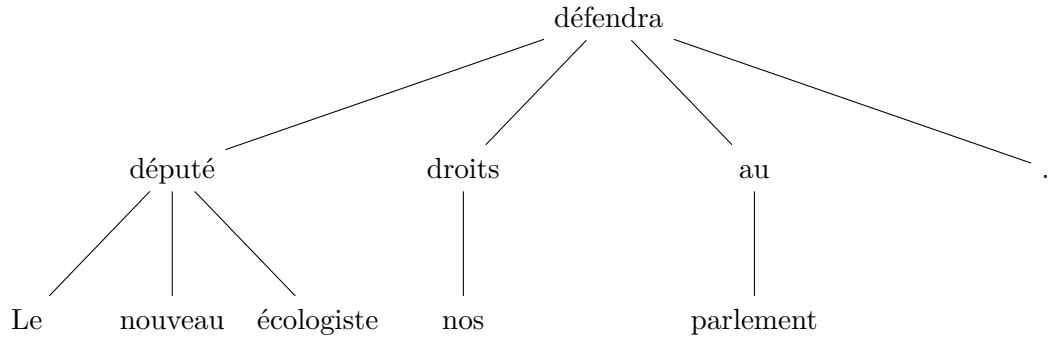

FIGURE 6.1 – Arbre de dépendances pour la phrase «Le nouveau député écologiste défendra nos droits au parlement.».

Transition		**Configuration**		
		$([w_0]$,	$[Le, nouveau, ..., parlement, .]$,	$\emptyset)$
SHIFT	$\Rightarrow$	$([w_0, Le]$,	$[nouveau, ..., parlement, .]$,	$\emptyset)$
SHIFT	$\Rightarrow$	$([w_0, Le, nouveau]$,	$[député, ..., parlement, .]$,	$\emptyset)$
LEFT-ARC	$\Rightarrow$	$([w_0, Le]$,	$[député, ..., parlement, .]$,	$A = \{(3, 2)\})$
LEFT-ARC	$\Rightarrow$	$([w_0]$,	$[député, ..., parlement, .]$,	$A_1 = A \cup \{(3, 1)\})$
SHIFT	$\Rightarrow$	$([w_0, député]$,	$[écologiste, ..., parlement, .]$,	$A_1)$
RIGHT-ARC	$\Rightarrow$	$([w_0, député, écologiste]$,	$[défendra, ..., parlement, .]$,	$A_2 = A_1 \cup \{(3, 4)\})$
REDUCE	$\Rightarrow$	$([w_0, député,]$,	$[défendra, ..., parlement, .]$,	$A_2)$
LEFT-ARC	$\Rightarrow$	$([w_0]$,	$[défendra, ..., parlement, .]$,	$A_3 = A_2 \cup \{(5, 3)\})$
SHIFT	$\Rightarrow$	$([w_0, défendra]$,	$[nos, ..., parlement, .]$,	$A_3)$
SHIFT	$\Rightarrow$	$([w_0, défendra, nos]$,	$[droits, ..., parlement, .]$,	$A_3)$
LEFT-ARC	$\Rightarrow$	$([w_0, défendra]$,	$[droits, ..., parlement, .]$,	$A_4 = A_3 \cup \{(7, 6)\})$
RIGHT-ARC	$\Rightarrow$	$([w_0, défendra, droits]$,	$[au, parlement, .]$,	$A_5 = A_4 \cup \{(5, 7)\})$
REDUCE	$\Rightarrow$	$([w_0, défendra]$,	$[au, parlement, .]$,	$A_5)$
RIGHT-ARC	$\Rightarrow$	$([w_0, défendra, au]$,	$[parlement, .]$,	$A_6 = A_5 \cup \{(5, 8)\})$
RIGHT-ARC	$\Rightarrow$	$([w_0, défendra, au, parlement]$,	$[.]$,	$A_7 = A_6 \cup \{(8, 9)\})$
REDUCE	$\Rightarrow$	$([w_0, défendra, au]$,	$[.]$,	$A_7)$
REDUCE	$\Rightarrow$	$([w_0, défendra]$,	$[.]$,	$A_7)$
RIGHT-ARC	$\Rightarrow$	$([w_0, défendra, .]$,	$[]$,	$A_8 = A_7 \cup \{(5, 10)\})$
REDUCE	$\Rightarrow$	$([w_0, défendra]$,	$[]$,	$A_8)$
REDUCE	$\Rightarrow$	$([w_0]$,	$[]$,	$A_8)$

TABLE 6.2 – Séquence de transitions permettant d'obtenir la structure de dépendances de la phrase «Le nouveau député écologiste défendra nos droits au parlement.».

6.2 Analyseur multilingue en dépendances syntaxique fondé sur les réseaux de neurones récurrents

Dans cette section, nous allons présenter tout d'abord l'adaptation de notre modèle pour la construction d'analyseurs multilingues en dépendances syntaxiques fondé sur les réseaux de neurones récurrents, avant de présenter les corpus utilisés et les résultats obtenus sur l'anglais, le français, l'allemand, l'espagnol et l'italien.

6.2.1 Adaptation du modèle neuronal à l'analyse syntaxique en dépendances par transition

Comme présenté dans la section précédente, une analyse par transition d'une phrase donnée consiste à trouver la séquence de transitions permettant de construire la structure de dépendances correcte pour la phrase. À partir d'une configuration initiale contenant la phrase, il s'agit de prédire et d'appliquer les bonnes transitions pour obtenir une configuration finale contenant la structure de dépendances. Il y a donc, d'une part, le système par transition (description formelle des opérations précédemment présentées), et

d'autre part, le système de prédiction qui détermine les transitions à appliquer.

Le principe d'une analyse par transition (telle qu'employée par les algorithmes du MaltParser [Nivre et al. 2006] par exemple) réside alors dans la bonne prédiction des transitions. La prédiction des transitions est couramment effectuée à l'aide de modèles statistiques. Il s'agit, lors du parcours d'une analyse par transition et pour chaque nouvelle configuration engendrée, de prédire la transition la plus probable sachant cette configuration. Cette propriété de l'analyse syntaxique par transition fait que notre modèle est parfaitement adapté à la prédiction des transitions.

Une étape préliminaire est nécessaire pour la mise en place de notre analyseur multilingue en dépendances syntaxique fondé sur les réseaux de neurones récurrents. Elle consiste à extraire le corpus d'apprentissage du corpus en dépendances. Pour ce faire, chaque structure de dépendances correctement annotée est convertie en une séquence de transitions permettant de reconstruire cette structure. L'ensemble des séquences de transition constitue notre corpus d'apprentissage. Ce corpus est utilisé pour entraîner notre modèle à prédire les transitions en fonction des configurations.

Pour la conversion des structures de dépendances en séquences de transitions, nous avons utilisé une méthode inspirée de la méthode proposée par Lacroix et al. [2016].

6.2.1.1 Notre modèle pour l'analyse en dépendances syntaxique

L'architecture de notre modèle pour l'analyse en dépendances syntaxique basé sur les RNN est donné dans la figure 6.2. Comme le montre la figure 6.2 nous utilisons un SRNN composé d'une succession de quatre couches de neurones (voir section 3.1.2.1). Ayant la configuration C_t au temps t, la couche d'entrée est maintenant constituée de la concaténation des éléments suivants $[w_i, w_j, f(t-1), Action(t-1)]$, où :

- w_i et w_j : sont respectivement les représentations vectorielles (extraites de notre représentation multilingue des mots, voir section 3.1.1) du mot se trouvant en haut de la pile σ et du premier mot de la mémoire tampon β.
- f(t-1) : est la première couche (couche Forward) au temps précédent.
- Action (t-1) : présente l'une des quatre actions possibles, prédite en sortie du réseau au temps précédent.

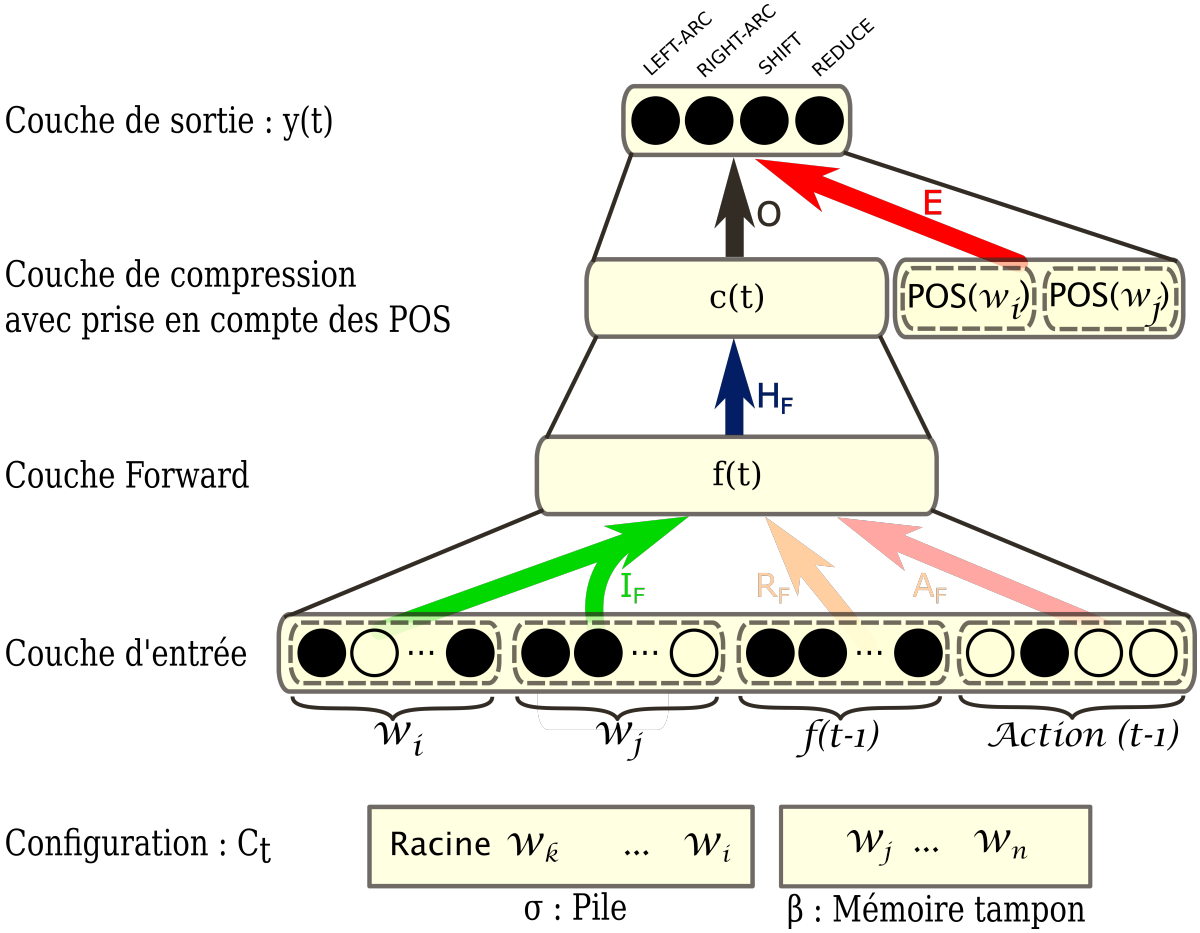

FIGURE 6.2 – Analyseur multilingue en dépendances syntaxique par transition fondé sur les réseaux de neurones récurrents.

Comme présenté dans la section 3.1.2.1, nous avons fait le choix d'utiliser deux couches cachées dans nos modèles neuronaux, une couche Forward et une couche de compression, et comme fonction d'activation, la fonction *sigmoïde* (voir eq. 3.4). Nous avons démontré que l'intégration des traits morpho-syntaxiques dans notre modèle neuronal, permet d'améliorer les performances de notre modèle pour l'annotation en SuperSenses (voir Chapitre 5). Nous avons aussi démontré que la prise en compte des annotations morpho-syntaxiques au niveau de la deuxième couche est la plus pertinente (voir section 5.2.3). De ce fait, nous avons fait le choix d'intégrer cette information (les annotations morpho-syntaxiques des mots en entrée du réseau w_i et w_j) au niveau de la couche de compression (deuxième couche cachée) de notre analyseur multilingue en dépendances syntaxique fondé sur les RNN.

Les valeurs en sortie de notre modèle, sachant l'entrée $[w_i, w_j, f(t-1), Action(t-1)]$, sont calculées comme suit :

— Calculer la valeur de la couche Forward f(t), avec la prise en compte des modifications opérées sur la couche d'entrée, l'équation 3.1 devient :

$$f(t) = \Sigma(w_i w_j . I_F(t) + f(t-1).R_F(t) + Action(t-1).A_F(t)) \tag{6.1}$$

— Calculer la valeur de la couche de compression en utilisant l'équation 3.2.

— Calculer la sortie du réseau (prédiction du modèle) y(t), avec la prise en compte des annotations morpho-syntaxiques intégrées dans la couche de compression, l'équation 3.3 devient :

$$y(t) = \Gamma(c(t).O(t) + POS(w_i)POS(w_j).E(t)) \tag{6.2}$$

où Γ est une fonction de type *softmax* (voir eq. (3.5))

Comme présenté dans la section 3.1.3.2, nous avons utilisé, pour l'apprentissage de nos modèles, l'algorithme de la rétro-propagation du gradient de l'erreur à travers le temps [Rumelhart et al. 1985].

6.2.2 Corpus et outils

Notre but est d'évaluer les performances de notre modèle neuronal pour l'analyse syntaxique par transition (voir figure 6.2) et aussi pour évaluer les bénéfices de la prise en compte des informations morpho-syntaxiques, des mots en entrée du réseau, au niveau de la deuxième couche cachée. Nous réalisons nos expérimentations sur 5 langues du corpus Universal Dependency Treebank [1] (UDT, v2.0 standard) [McDonald et al. 2013] : anglais, français, allemand, espagnol et italien.

Nous comparons les performances de nos modèles avec quatre méthodes d'état de l'art basées sur le transfert cross-lingue des dépendances :

— McDonald et al. [2011] : ont proposé une méthode basée sur la construction d'un annotateur en dépendances délexicalisé, appris sur une langue bien dotée (langue

1. https ://github.com/ryanmcd/uni-dep-tb

source), qui sera par la suite re-lexicalisé sur des données de la langue cible.

— Ma and Xia [2014] : cette deuxième méthode est basée sur le transfert de connaissances cross-lingue en utilisant une méthode de régularisation d'entropie.

— Lacroix et al. [2016] : ont proposé une méthode se rapprochant de la méthode de Rasooli and Collins [2015], aussi basée sur l'apprentissage d'analyseur en dépendances cross-lingue par projection partielle de dépendances.

Nous utilisons le même corpus parallèle (Europarl) que ces méthodes de l'état de l'art, ainsi que les mêmes corpus d'apprentissage et d'évaluation (UDT v2.0 std). Nous évaluons aussi nos modèles avec la même métrique d'évaluation : le UAS [2] (excluant la ponctuation).

L'apprentissage de nos modèles neuronaux est fait sur plusieurs sous-ensembles extraits du corpus Europarl [Koehn 2005] constitués des phrases communes aux 5 langues que nous étudions ici. Le côté anglais (côté source) du corpus Europarl est annoté en dépendances syntaxiques en utilisant deux analyseurs syntaxiques différents : le MaltParser [Nivre et al. 2006] (basé sur l'analyse en dépendances par transitions) et le Syntaxnet [3] (plus précisément le modèle prés-entraîné pour l'anglais Parsey McParseface) [Andor et al. 2016].

6.2.3 Systèmes évalués

Pour l'apprentissage de nos modèles neuronaux nous avons extrait trois corpus multi-parallèles contenant respectivement $60K$, $100K$ et $150K$ phrases communes aux 5 langues que nous avons choisies de traiter, la construction de ces trois corpus d'apprentissage est motivée par notre souhait, d'étudier l'impact de la taille de notre représentation multilingue et l'impact des données d'apprentissage sur les performances de nos modèles neuronaux. Nous considérons, dans nos expérimentations, le côté anglais comme côté source (c'est la langue la mieux dotée). La seconde étape de notre processus expérimental réside dans le fait d'annoter en dépendance le côté source de nos trois corpus d'apprentissage. Pour ce faire, nous utilisons trois analyseurs syntaxiques : l'outil MaltParser - le *MaltParser-UD-Treebank* appris sur le corpus d'entraînement du UDT (v2.0 std), et le *MaltParser-Penn-Treebank*

2. UAS (Unlabeled Attachment Score) correspond au pourcentage de mots étant correctement rattachés sur l'ensemble du corpus, à l'exception des ponctuations.

3. http ://github.com/tensorflow/models/tree/master/synt axnet

appris sur le *Penn Treebank II* [Marcus et al. 1994] ; le troisième analyseur est le modèle pré-entraîné pour l'anglais de SyntaxNet (Parsey McParseface). Nous disposons à cette étape de 9 corpus d'apprentissage (chacun de nos trois corpus étant annoté par chacun des trois annotateurs), que nous avons utilisé pour l'apprentissage de nos modèles neuronaux. Pour nommer nos modèles nous avons adopté le formalisme suivant : RNN-nom de l'analyseur en dépendances utilisé pour annoter le corpus d'apprentissage-Taille du corpus d'apprentissage. Afin d'évaluer l'effet de la prise en compte des annotations morpho-syntaxiques pour l'analyse en dépendances, nous avons appris des modèles avec (comportant la mention POS-H2 dans leurs noms) et sans prise en compte de cette information.

6.2.4 Analyse des résultats

Les performances de nos analyseurs en dépendances basés sur les RNN et méthodes de l'état de l'art basées sur le transfert cross-lingue sont présentées dans le tableau 6.3. Les scores obtenus par nos analyseurs sont proches des scores obtenus par les méthodes de l'état de l'art. En nous concentrant sur les scores obtenus par nos annotateurs, nous constatons que les modèles avec prise en compte des annotations morpho-syntaxiques obtiennent les meilleurs résultats. Notons aussi que ce sont les modèles appris sur les corpus de taille *100K* qui sont les plus performants de nos modèles et que si on augmente la taille des corpus à 150K les performances baissent. Ces baisses des scores s'expliquent entre autres par le fait que l'augmentation du corpus d'apprentissage fait augmenter aussi la taille des représentations vectorielles des mots et par la suite l'augmentation de la taille de la couche d'entrée du réseau, cette augmentation crée un déséquilibre puisqu'on n'augmente pas la taille des autres couches du réseau.

Notons également que les performances de nos modèles restent éloignés des performances des annotateurs supervisés. Cela peut être dû au fait que nous utilisons pour l'apprentissage de nos modèles des corpus d'apprentissage annotés par ces annotateurs supervisés.

	Modèle \ Lang.	Anglais	Français	Allemand	Espagnol	Italien
Modèles Sup.	MaltParser-UD-Treebank	87.17	–	–	–	–
	MaltParser-Penn-Treebank	91.83	–	–	–	–
	SyntaxNet	89.42	79.41	–	–	–
Nos Parseurs Basés RNN	RNN-MaltP.-UD-Treebank-60k	74.06	64.94	64.02	65.79	64.61
	RNN-MaltP.-UD-Treebank-100k	81.32	73.25	69.80	72.39	70.90
	RNN-MaltP.-UD-Treebank-100k-POS-H2	81.90	73.57	70.13	72.60	71.15
	RNN-MaltP.-UD-Treebank-150k	79.34	71.44	68.87	71.48	69.70
	RNN-MaltP.-Penn-Treebank-60k	78.92	70.87	67.12	70.22	66.79
	RNN-MaltP.-Penn-Treebank-100k	83.42	75.76	71.81	74.93	72.50
	RNN-MaltP.-Penn-Treebank-100k-POS-H2	**84.07**	**75.85**	**71.98**	**75.12**	**72.75**
	RNN-MaltP.-Penn-Treebank-150k	81.65	75.02	70.26	74.46	71.00
	RNN-SyntaxNet-60k	75.36	66.07	64.21	65.98	64.95
	RNN-SyntaxNet-100k	82.72	74.00	69.86	72.72	71.30
	RNN-SyntaxNet-100k-POS-H2	83.16	74.46	70.02	72.95	71.56
	RNN-SyntaxNet-150k	80.35	72.23	69.15	71.93	69.65
SOTA	(McDonald *et al.*, 2011)	–	73.13	69.77	68.72	70.74
	(Ma and Xia, 2014)	–	76.53	74.30	75.53	77.74
	(Rasooli and Collins, 2015)	–	**79.91**	**74.32**	**78.17**	**79.46**
	(Lacroix *et al.*, 2016)	–	77.92	73.75	76.87	77.82

TABLE 6.3 – Performances en UAS de nos parseurs RNN avec prise en compte des POS et comparaison avec quatre méthodes de l'état de l'art.

6.3 Bilan

Nous avons démontré qu'il était possible d'apprendre un analyseur en dépendances basé sur les RNN. Nous avons aussi validé notre constat du chapitre 5, que la prise en compte des informations linguistiques dans les RNN permet d'améliorer les performances de ces modèles.

Conclusion

Conclusion

Au cours de cette thèse nous nous sommes intéressés à la problématique de la construction rapide d'outils et de ressources pour les langues peu dotées. D'une part, on constate que parmi les 6000 langues existantes dans le monde, un petit nombre de langues richement dotées, disposent d'outils d'analyse linguistique performants et de ressources annotées de bonne qualité. La construction de ces outils et ressources et un processus très coûteux qui a nécessité des efforts considérables et qui a mobilisé des milliers de chercheurs durant plusieurs années. Il est donc impossible de refaire les mêmes efforts, déployés pour les langues richement dotées, pour le traitement des langues peu dotées.

D'autre part, la disponibilité, ces dernières années, de large corpus parallèles et d'une multitude de sites internet multilingues, a poussé les chercheurs à utiliser ces corpus parallèles (préalablement alignés automatiquement au niveau des mots) pour transférer des annotations d'une langue source (richement dotée) vers une autre langue cible (faiblement dotée). En outre, cette méthode a permis d'exploiter, pour la construction automatique d'outils et de ressources pour des langues peu dotées, les ressources disponibles pour des langues richement dotées ainsi que les annotations (ou systèmes) déjà disponibles pour ces mêmes langues. Cependant, les performances des outils non-supervisés (basés sur la projection interlingue d'annotations) restent relativement faibles, en comparaison avec les performances des outils supervisés. Cela est dû en grande partie au fait que les performances des algorithmes d'alignement au niveau des mots ne sont pas toujours satisfaisantes (du point de vue de la qualité des alignements prédits) et l'étape d'alignement au niveau des mots (un alignement n'est pas toujours 1-1, il peut être 1-N, N-N, etc.) constitue

aujourd'hui un facteur limitant la projection d'annotations linguistiques [Fraser and Marcu 2007]. Il est donc nécessaire de travailler à lever cette limitation pour l'amélioration des performances des outils d'analyse des langues peu dotées.

Le chapitre 3 de cette thèse présente les travaux que nous avons menés dans le cadre de la mise en place d'un modèle basé sur les réseaux de neurones récurrents pour la construction d'annotateurs linguistiques multilingues. Notre approche utilise un corpus parallèle aligné au niveau des phrases seulement et n'applique aucun pré-traitement du type alignement automatique en mots qui peut être source d'erreurs et de bruit. Notre approche n'utilise pas d'autres sources d'information, ce qui la rend applicable à un large éventail de langues peu dotées. Dans le cadre de nos travaux, nous avons proposé une représentation multilingue des mots. Représentation que nous avons utilisée pour l'apprentissage de nos modèles neuronaux. Nous avons utilisé pour nos expérimentations deux architectures de réseaux de neurones récurrents : l'architecture simple et l'architecture bidirectionnelle. Nous nous basons sur le corpus parallèle pour la construction de notre représentation multilingue des mots. Étant donné que les mots inconnus (mots hors vocabulaire) ne figurent pas dans le corpus parallèle, leur prise en compte était quasi-inexistante dans notre modèle. Nous avons alors proposé un processus de pré-traitement des mots hors vocabulaire. Ce pré-traitement s'est révélé efficace, nos expérimentations ont montré une amélioration des performances de nos modèles sur les mots inconnus et par conséquent une amélioration des performances globales.

Nos modèles neuronaux ont été utilisés, dans un cadre multilingue, pour le traitement de trois tâches relevant du traitement automatique de la langue : l'annotation morpho-syntaxique, l'annotation en SuperSenses et l'annotation en dépendances syntaxique. Hiérarchiquement dans une analyse linguistique standard, l'annotation morpho-syntaxique fait partie des tâches qui sont relativement de bas niveau, les résultats de cette tâche sont utilisés pour la réalisation des tâches plus complexe (annotation en SuperSenses et annotation en dépendances syntaxiques). Afin de pouvoir prendre en compte le résultat de l'analyse morpho-syntaxique dans nos annotateurs en SuperSenses et en dépendances syntaxiques nous avons proposé plusieurs variantes de RNN, l'ajout d'informations morpho-syntaxiques nous a permis d'améliorer les performances de nos modèles neuronaux pour les

annotations en SuperSenses et en dépendances.

Nous présentons, dans le chapitre 4, l'utilisation de notre approche pour la construction d'annotateurs morpho-syntaxiques multilingues (non supervisés pour les langues cibles). Nous avons obtenu des résultats très satisfaisants, équivalents aux résultats des approches de l'état de l'art. De plus, en utilisant un corpus d'adaptation en langue cible, de petite taille, nous avons démontré l'efficacité de notre approche dans un cadre semi-supervisé.

Dans le chapitre 5, nous proposons d'adapter notre approche neuronale pour la construction d'annotateurs en SuperSenses. Nous avons obtenu des résultats proches du système standard. Nous avons aussi montré la généricité de notre approche.

Le chapitre 6, présente l'adaptation de notre approche pour la construction d'analyseurs multilingues en dépendances syntaxique. Nous nous sommes inspirés du système par transition (Arc-Eager) pour l'adaptation de notre approche à cette tâche. Nous avons évalué l'efficacité de cette adaptation en utilisant l'Universal Dependency Treebank. Les résultats que nous avons obtenus sont proches des résultats des approches de référence.

À travers nos travaux, nous avons d'une part proposé une approche pour la construction d'annotateurs multilingues et d'autre part démontré sa généricité sur plusieurs langues et plusieurs tâches d'annotation.

Perspectives

Les travaux que nous avons menés dans le cadre de cette thèse nous ont permis de mettre en évidence que la prise en compte des mots inconnus ainsi que l'intégration des informations linguistiques dans nos modèles neuronaux permet une amélioration significative de leurs performances. Cependant, plusieurs pistes d'amélioration restent à explorer. Nos envisageons dans nos travaux futurs d'expérimenter deux stratégies d'apprentissages des réseaux de neurones : (a) l'apprentissage multi-tâche, dont l'efficacité a été démontré par Collobert and Weston [2008]; Collobert et al. [2011], afin de construire un modèle qui réalise simultanément les analyses syntaxique et sémantique, (b) l'apprentissage multi-source, utilisé avec succès par Duong [2017], pour exploiter diverses ressources annotées provenant de plusieurs langues sources.

Enfin, dans le cadre de la prise en compte de la divergence linguistique entre les langues traitées, et en se basant sur les travaux de Aufrant et al. [2015], nous envisageons d'utiliser des connaissances linguistiques extraites par exemple de la base du World Atlas of Language Structures (WALS) [4].

4. http ://wals.info

Publications

Les idées et les résultats présentés dans cette thèse ont été publié dans les articles suivants :

— Othman Zennaki, Nasredine Semmar, Laurent Besacier, « A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages ». *Natural Language Engineering, August 2018, n.d., 1–25.*

— Othman Zennaki, Nasredine Semmar, Laurent Besacier, « Inducing Multilingual Text Analysis Tools Using Bidirectional Recurrent Neural Networks ». *26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japan : 2016.

— Othman Zennaki, Nasredine Semmar, Laurent Besacier, « Projection Interlingue d'Étiquettes pour l'Annotation Sémantique Non Supervisée ». *23ème Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN-RECITAL 2016)*, Paris, France : 2016.

— Othman Zennaki, Nasredine Semmar, Laurent Besacier, « Unsupervised and Lightly Supervised Part-of-Speech Tagging Using Recurrent Neural Networks ». *The 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, Shanghai, Chine : 2015.

— Othman Zennaki, Nasredine Semmar, Laurent Besacier, « Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle ». *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL 2015)*, Caen, France : 2015.

Les idées exposées dans cette thèse ont aussi données lieu à plusieurs collaborations publiées dans les articles suivants :

— Sara Meftah, Nasredine Semmar, Othman Zennaki, Fatiha Sadat, « Using Transfer

Learning in Part-Of-Speech Tagging of English Tweets ». *8th Language Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2017)*, Poznań, Pologne : 2017.

— Nasredine Semmar, Othman Zennaki, Meriama Laib, « Etude de l'impact d'un lexique bilingue spécialisé sur la performance d'un moteur de traduction à base d'exemples ». *23ème Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN-RECITAL 2016)*, Paris, France : 2016.

— Nasredine Semmar, Othman Zennaki, Meriama Laib, « Improving the Performance of an Example-Based Machine Translation System Using a Domain-specific Bilingual Lexicon ». *The 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, Shanghai, Chine : 2015.

— Nasredine Semmar, Othman Zennaki, Meriama Laib, « Evaluating the Impact of Using a Domain-specific Bilingual Lexicon on the Performance of a Hybrid Machine Translation Approach ». *Recent Advances in Natural Language Processing (RANLP 2015)*.

— Nasredine Semmar, Othman Zennaki, Meriama Laib, « Using Cross-Language Information Retrieval and Statistical Language Modelling in Example-Based Machine Translation ». *In Proceed-ings of the 36-th Translating and the Computer conference*, England : 2014 .

Bibliographie

Authoul Abdulhay. *Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné*. PhD thesis, Université de Grenoble, France, 2012.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot : Distributed word representations for multilingual nlp. *CoNLL-2013*, pages 183–192, 2013.

Alexandre Allauzen and Guillaume Wisniewski. Modèles discriminants pour l'alignement mot à mot. *Traitement Automatique des Langues*, 50(3) :173–203, 2010.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. Massively multilingual word embeddings. *arXiv preprint arXiv :1602.01925*, 2016.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *arXiv preprint arXiv :1603.06042*, 2016.

Paolo Annesi and Roberto Basili. Cross-lingual alignment of framenet annotations through hidden markov models. In *Computational Linguistics and Intelligent Text Processing*, pages 12–25. 2010.

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. Zero-resource dependency parsing : Boosting delexicalized cross-lingual transfer with linguistic knowledge. In *COLING 2016, the 26th International Conference on Computational Linguistics*, pages 119–130. The COLING 2016 Organizing Committee, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by

jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September 2014. URL `https://arxiv.org/abs/1409.0473`.

Michel Ballard. *Eléments pour une didactique de la traduction*. PhD thesis, Université Sorbonne nouvelle – Paris 3, France, 1991.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3 :1137–1155, 2003.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

Luisa Bentivogli and Emanuele Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources : the multisemcor corpus. *Natural Language Engineering*, 11(03) :247–261, 2005.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of the 20th CoNLL*, page 364, 2004.

Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. Multivec : a multilingual and multilevel representation learning toolkit for nlp. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pages 4188–4192, 2016.

Vincent Berment. *Méthodes pour informatiser les langues et les groupes de langues «peu dotées»*. PhD thesis, Mémoire de thèse de doctorat, Université Joseph-Fourier-Grenoble I, 2004.

Laurent Besacier, Benjamin Lecouteux, Marwen Azouzi, and Ngoc-Quang Luong. The lig english to french machine translation system for iwslt 2012. In *IWSLT*, pages 102–108, 2012.

Romaric Besançon, Gaël De Chalendar, Olivier Ferret, Faïza Gara, Olivier Mesnard, Meriama Laïb, and Nasredine Semmar. Lima : A multilingual framework for linguistic

analysis and linguistic resources development and evaluation. In *Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta*, 2010.

Yves Bestgen. Construction automatique de ressources lexicales pour la fouille d'opinion : extension aux n-grammes. *Dixième édition de la Conférence en Recherche d'Information et Applications (CORIA 2016)*, 2013.

F Bisson. *U Méthodes et outils pour l'appariement de textes bilingues*. PhD thesis, Thèse de Doctorat en Informatique. Université Paris VII, 2000.

Ingeborg Blank. Terminology extraction from parallel technical texts. In *Parallel Text Processing*, pages 237–252. Springer, 2000.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. Identifying bilingual multi-word expressions for statistical machine translation. In *LREC*, pages 674–679, 2012.

Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 977–981. Association for Computational Linguistics, 1992.

Thorsten Brants. Tnt : a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231, 2000.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation : Parameter estimation. *Computational linguistics*, 19(2) :263–311, 1993.

Sabine Buchholz and Erwin Marsi. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164, 2006.

Jean Carletta. Assessing agreement on classification tasks : the kappa statistic. *Computational linguistics*, 22(2) :249–254, 1996.

Massimiliano Ciaramita and Mark Johnson. Supersense tagging of unknown nouns in wordnet. In *EMNLP*, 2003.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1) :37–46, 1960.

Michael Collins and Terry Koo. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1) :25–70, 2005.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing from scratch. *The Journal of Machine Learning Research*, 12 :2493–2537, 2011.

Antoine Cornuéjols and Laurent Miclet. *Apprentissage artificiel : concepts et algorithmes*. Editions Eyrolles, Paris, 2002.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 515–521. Association for Computational Linguistics, 1994.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, 1 :600–609, 2011.

Fathi Debili and Elyes Sammouda. Appariement des phrases de textes bilingues. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 517–538, 1992.

Fathi Debili and Adnane Zribi. Les dépendances syntaxiques au service de l'appariement des mots. *Actes du 10ème Congrès RFIA*, 1996.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Long Duong. *Natural language processing for resource-poor languages*. PhD thesis, The University of Melbourne, Australia, 2017.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. Simpler unsupervised pos tagging with bilingual projections. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 634–639, 2013.

Greg Durrett, Adam Pauls, and Dan Klein. Syntactic transfer using a bilingual lexicon. In *The Joint Conference on EMNLP and CoNLL*, pages 1–11, 2012.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, pages 334–343, 2015.

Maud Ehrmann. *Les entités nommées, de la linguistique au TAL*. PhD thesis, Univ. Paris 7, 2008.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2) :179–211, 1990.

Iris Eshkol-Taravella. *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral*. PhD thesis, Mémoire d'Habilitation à Diriger des Recherches, Université d'Orléans, 2015.

S Federici and V Pirrelli. Analogical modelling of text tagging. *unpublished report, Istituto diLinguistica Computazionale, Italy*, 1993.

Anna Feldman, Jirka Hana, and Chris Brew. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC*, pages 549–554, 2006.

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Olivier Ferret. Construire des représentations denses à partir de thésaurus distributionnels. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 93.

Alexander Fraser and Daniel Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3) :293–303, 2007.

Catherine Fuchs. Linguistique et traitement automatiques des langues. *Hachette Education, Paris*, 1993.

Roger Garside, Geoffrey N Leech, and Tony McEnery. *Corpus annotation : linguistic information from computer text corpora*, chapter Introduction corpus annotation. Taylor & Francis, 1997.

Eric Gaussier and J-M Langé. Modèles statistiques pour l'extraction de lexiques bilingues. *TAL. Traitement automatique des langues*, 36(1-2) :133–155, 1995.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, pages 1386–1390, 2015.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa : Fast bilingual distributed representations without word alignments. *ICML 2015*, 2015.

Alex Graves. Supervised sequence labelling with recurrent neural networks. 2012. *Springer*, 385, 2012.

Yoan Gutiérrez Vázquez, Antonio Fernández Orquín, Andrés Montoyo Guijarro, Sonia Vázquez Pérez, et al. Enriching the integration of semantic resources based on wordnet. 2011.

Thierry Hamon, Julien Derivière, and Adeline Nazarenko. Ogmios : une plate-forme d'annotation linguistique de collection de documents issus du web. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 103–112. Association pour le Traitement Automatique des Langues, 2007.

Zellig S Harris. Distributional structure. *Word*, 10(2-3) :146–162, 1954.

James Henderson. Discriminative training of a neural network statistical parser. In *Proceedings of the 42nd Annual Meeting on ACL*, page 95, 2004.

Hieu Hoang and Philipp Koehn. Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 58–65, 2008.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5) :359–366, 1989.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 392–399. Association for Computational Linguistics, 2002.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3) :311–325, 2005.

Bassam Jabaian. *Systèmes de compréhension et de traduction de la parole : vers une approche unifiée dans le cadre de la portabilité multilingue des systèmes de dialogue*. PhD thesis, Université d'Avignon et des Pays de Vaucluse, France, 2012.

Bassam Jabaian, Laurent Besacier, and Fabrice Lefevre. Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE TASLP*, 21(3) :636–648, 2013. URL `https://hal.archives-ouvertes.fr/hal-00953651`.

Wenbin Jiang, Qun Liu, and Yajuan Lü. Relaxed cross-lingual projection of constituent syntax. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201. Association for Computational Linguistics, 2011.

Wenbin Jiang, Yajuan Lü, Liang Huang, and Qun Liu. Automatic adaptation of annotations. *Computational Linguistics*, 41(1) :119–147, 2015.

Anders Johannsen, Dirk Hovy, Héctor Martınez Alonso, Barbara Plank, and Anders Søgaard. More or less supervised supersense tagging of twitter. *SEM (2014)*, page 1, 2014.

Daniel Jurafsky. Speech and language processing : An introduction to natural language processing. *Computational linguistics, and speech recognition*, 2000.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. A cross-lingual annotation projection-based self-supervision approach for open information extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 741–748, 2011.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the ACL*, volume 1, pages 694–702, 2012.

Philipp Koehn. Europarl : A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Upali S Kohomban and Wee Sun Lee. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting on ACL*, pages 34–41, 2005.

H Kucera and W Francis. A standard corpus of present-day edited american english, for use with digital computers (revised and amplified from 1967 version), 1979.

Ophélie Lacroix. *From syntactic tagging for categorial dependency grammars to transition-based parsing in the domain of non-projective dependency parsing.* Theses, Université de Nantes, December 2014. URL `https://hal.archives-ouvertes.fr/tel-01112072`.

Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. Apprentissage d'analyseur en dépendances cross-lingue par projection partielle de dépendances. In *Actes de la 23e conférence sur le Traitement Automatique des Langues Naturelles*, pages 1–14, 2016.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, pages 302–308, 2014.

Shen Li, Joao V Graça, and Ben Taskar. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on EMNLP and CoNLL*, pages 1389–1398. Association for Computational Linguistics, 2012.

Thang Luong, Hieu Pham, and Christopher D. Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, 2015.

Xuezhe Ma and Fei Xia. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, pages 1337–1348, 2014.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank : annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, 1994.

Nicolas Mazziotta. Logiciel notabene pour l'annotation linguistique. annotations et conceptualisations multiples. *Recherches qualitatives. Hors-série*, 9 :83–94, 2010.

Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics, 2011.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, pages 92–97, 2013.

Frédérique Mélanie-Becquet and Frédéric Landragin. Linguistique outillée pourl'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages*, (3) : 117–137, 2014.

Tomáš Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1045–1048, 2010.

Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE, 2011.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *In The Workshop Proceedings of ICLR*, 2013a.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In *Proceedings of the workshop on HLT*, pages 303–308, 1993.

Roberto Navigli and Mirella Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence*, 32 (4) :678–692, 2010.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193 :217–250, 2012.

Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 : Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics*, volume 2, pages 222–231, 2013.

Joakim Nivre. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT*. Citeseer, 2003.

Joakim Nivre, Johan Hall, and Jens Nilsson. Maltparser : A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219, 2006.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of LREC*, 2016.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, 2000.

Sylwia Ozdowska. Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés. In *Proceedings of 11ème conférence TALNRECITAL*, 2004.

Sylwia Ozdowska and Vincent Claveau. Inférence de règles de propagation syntaxique pour l'alignement de mots. *TAL*, 47(1) :167–186, 2006.

Sebastian Padó and Mirella Lapata. Cross-linguistic projection of role-semantic information. In *HLT and EMNLP*, 2005.

Sebastian Padó and Mirella Lapata. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36 :307–340, 2009.

Sebastian Pado and Guillaume Pitel. Annotation précise du français en sémantique de rôles par projection cross-linguistique. *TALN-07*, 2007.

Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10) :1345–1359, 2010.

Carla Parra Escartín and Héctor Martínez Alonso. Choosing a spanish part-of-speech tagger for a lexically sensitive task. *Procesamiento del Lenguaje Natural*, (54), 2015.

Peyman Passban, Qun Liu, and Andy Way. Providing morphological information for smt using neural networks. *The Prague Bulletin of Mathematical Linguistics*, 108(1) :271–282, 2017.

Nicolas Pécheux, Alexandre Allauzen, Thomas Lavergne, Guillaume Wisniewski, and François Yvon. Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles crf. In *Conférence sur le Traitement Automatique des Langues Naturelles*, 2015.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Wim Peters, Ivonne Peters, and Piek Vossen. Automatic sense clustering in eurowordnet. 1998.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey*, pages 2089–2096, 2012.

Delyth Prys. The blark matrix and its relation to the language resources situation for the celtic languages. *Strategies for developing machine translation for minority languages*, page 31, 2006.

Mohammad Sadegh Rasooli and Michael Collins. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, 2015.

Salim Roukos, David Graff, and Dan Melamed. Hansard french/english. *Linguistic Data Consortium*, 1995.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

Helmut Schmid. Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43 :28, 1995.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A Smith. Coarse lexical semantic annotation with supersenses : an arabic case study. In *Proceedings of the 50th Annual Meeting of the ACL*, volume 2, pages 253–258, 2012.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11) :2673–2681, 1997.

Didier Schwab, Jérôme Goulian, Andon Tchechmedjiev, and Hervé Blanchon. Ant colony algorithm for the unsupervised word sense disambiguation of texts : Comparison and evaluation. In *COLING*, pages 2389–2404, 2012.

Nasredine Semmar and Laib Meriama. Using a hybrid word alignment approach for automatic construction and updating of arabic to french lexicons. In *Editors & Workshop Chairs*, page 114, 2010.

Frank Smadja, Kathleen R McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons : A statistical approach. *Computational linguistics*, 22(1) : 1–38, 1996.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Anders Søgaard. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, pages 682–686. Association for Computational Linguistics, 2011.

Martin Sundermeyer, Ilya Oparin, J-L Gauvain, Ben Freiberg, Ralf Schluter, and Hermann Ney. Comparison of feedforward and recurrent neural network language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8430–8434. IEEE, 2013.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters

for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the NAACL-HLT*, pages 477–487, 2012.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the ACL*, 1 :1–12, 2013a.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. Target language adaptation of discriminative transfer parsers. 2013b.

Ivan Titov and Alexandre Klementiev. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the ACL*, volume 1, pages 647–656, 2012.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. Cross-lingual metaphor detection using common semantic features. *Meta4NLP*, page 45, 2013.

Lonneke van der Plas and Marianna Apidianaki. Cross-lingual word sense disambiguation for predicate labelling of french. *Proceedings of TALN 2014*, page 46, 2014.

J Veronis, O Hamon, C Ayache, R Belmouhoub, O Kraif, D Laurent, TMH Nguyen, N Semmar, F Stuck, and W Zaghouani. Arcade ii action de recherche concertée sur l'alignement de documents et son évaluation, 2008.

Jean Véronis. Annotation automatique de corpus : panorama et état de la technique. *Ingénierie des langues*, 4, 2000.

Jean-Marie Viprey and Virginie Léthier. Annotation linguistique de corpus : vers l'exhaustivité par la convivialité. *JADT'09, 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, 2008.

Mengqiu Wang, Wanxiang Che, and Christopher D Manning. Joint word alignment and bilingual named entity recognition using dual decomposition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, pages 1073–1082, 2013.

Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. Cross-lingual part-of-speech tagging through ambiguous learning. In *EMNLP*, volume 14, pages 1779–1785, 2014.

Min Xiao and Yuhong Guo. Annotation projection-based representation learning for cross-lingual dependency parsing. *CoNLL*, page 73, 2015.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on HLT*, pages 1–8, 2001.

Patrick Ye and Timothy Baldwin. Melb-yb : Preposition sense disambiguation using rich semantic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 241–244, 2007.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks. In *The 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015), Shanghai, Chine*, 2015a.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL 2015), Caen, France*, 2015b. URL `http://www.atala.org/taln_archives/TALN/TALN-2015/taln-2015-court-032`.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, 2013.