Traitement automatique des entités nommées en arabe : détection et traduction

Souhir Gahbiche-Braham — Hélène Bonneau-Maynard — François Yvon

Université Paris Sud & LIMSI-CNRS
BP 133 - 91403 ORSAY Cedex - France
souhir@limsi.fr, hbm@limsi.fr, yvon@limsi.fr

RÉSUMÉ. *La détection des entités nommées (EN) en langue arabe est un prétraitement potentiellement utile pour de nombreuses applications du traitement des langues, en particulier pour la traduction automatique. Cette tâche représente un sérieux défi, compte tenu des spécificités de l'arabe. Dans cet article, nous présentons une étude détaillée des entités nommées en arabe dans le cadre d'une application de traduction automatique statistique. Nous présentons notre système de détection des EN en arabe (NERAr), dans sa configuration de base, puis dans ses diverses évolutions. Dans notre architecture, NERAr est utilisé comme un prétraitement apportant des connaissances externes au système de traduction. Plusieurs stratégies d'intégration de ces connaissances sont considérées; dans la configuration la plus favorable, une évaluation automatique, corroborée par des analyses manuelles, permet d'observer une légère amélioration de la traduction des EN et une réduction des erreurs induites par les mots inconnus.*

ABSTRACT. *The recognition of Arabic Named Entities (NE) is a potentially useful preprocessing step for many Natural Language Processing Applications, such as Statistical Machine Translation (SMT). Due to peculiarities of the written Arabic language, this task is however rather challenging. In this paper, we present a detailed study of Arabic NEs in the context of a SMT system. We present our statistical NE recognition system (NERAr), and its various evolutions. NERAr was then used as a processing step, thus enabling us to incorporate external linguistic knowledge into the SMT system. Several strategies for performing this integration are explored. Automatic evaluations, corroborated by manual inspections, indicate a small improvement of the translation quality of NEs, and a reduction of the errors caused by out-of-vocabulary words.*

MOTS-CLÉS : *entités nommées, traduction automatique, traitement automatique de l'arabe.*

KEYWORDS: *Arabic natural language processing, named entity recognition, machine translation.*

1. Introduction

La détection des entités nommées (EN) est un élément essentiel pour de nombreuses tâches de traitement automatique des langues (TAL), qu'elles soient monolingues ou multilingues, comme la recherche d'information ou la traduction automatique. En témoignent par exemple les ateliers « *Named Entities* » organisés par l'ACL[1] ainsi que, notamment, les campagnes d'évaluation internationales (MUC, CoNLL, ACE)[2] ou nationales (ESTER) organisées au cours des vingt dernières années.

Nous nous intéressons, dans cet article, au traitement des EN dans un contexte de traduction automatique statistique depuis l'arabe vers le français, dans lequel le traitement des EN pose des problèmes particuliers. Un système de traduction statistique apprend à traduire en se fondant sur des exemples de traductions déjà réalisées, extraites de textes parallèles. Dans la mesure où ces corpus d'apprentissage sont relativement limités en taille, se pose alors la question de la traduction des mots qui ne sont pas observés lors de l'entraînement (*mots inconnus*[3]).

Cette question est particulièrement critique dans le cas de l'arabe[4]. L'arabe est en effet une langue morphologiquement complexe (Habash, 2011), ce qui implique que de nombreuses formes possibles sont rarement observées en corpus (Heintz, 2008). Comme on le verra à la section 4, ceci impose, *a minima*, de mettre en œuvre des outils d'analyse morphologique afin de circonscrire l'inventaire des unités sources du système de traduction.

Décomposer les mots inconnus complexes en unités connues n'épuise malheureusement pas le problème, puisque nombreuses de ces formes inconnues correspondent à des noms propres (arabes ou étrangers), qui ne se laissent pas toujours segmenter. Ces cas sont loin d'être marginaux : une étude des mots inconnus réalisée par Habash (2008) sur des corpus journalistiques pour la paire de langues arabe-anglais rapporte qu'environ 40 % des mots inconnus correspondent à des noms propres.

Pour traiter ces mots inconnus, les systèmes de traduction statistique emploient une stratégie par défaut qui consiste à recopier une forme inconnue *verbatim* dans la sortie en langue cible. Cette stratégie est parfois opérante (en particulier pour les noms de personnes) lorsque les langues source et cible utilisent le même alphabet ; dans le cas de l'arabe, elle s'avère inappropriée. Pour améliorer le traitement des mots inconnus, une stratégie courante consistera alors à translittérer les mots inconnus en

1. Association for Computational Linguistics.

2. MUC : Message Understanding Conferences ; CoNLL : Computational Natural Language Learning ; ACE : Automatic Content Extraction.

3. Nous utilisons *mots inconnus* de préférence à *mots hors vocabulaire* pour marquer le fait que nous nous intéressons à traiter les mots du texte à traduire qui ne sont pas connus par le système de traduction.

4. Le problème se pose à l'identique pour d'autres langues sources, comme le chinois (Zhang *et al.*, 2011 ; Zhang, 2012) ou encore le russe, et les solutions proposées ici pourront donc s'appliquer dans d'autres contextes.

alphabet latin (Al-Onaizan et Knight, 2002b), à condition de savoir se restreindre à translittérer les formes qui doivent l'être (par exemple les noms de personnnes et de lieux) (Hermjakob *et al.*, 2008), ou bien encore à consulter des dictionnaires bilingues (Hal et Jagarlamudi, 2011).

En résumé, le traitement des formes inconnues dans les textes arabes dans un contexte de traduction impose de distinguer diverses catégories de mots inconnus, afin de leur appliquer des traitements différentiels. Dans ce contexte, le réperage des EN apparaît comme un préalable à la traduction. Ce repérage est pourtant difficile en arabe, du fait notamment de l'absence de distinction entre les majuscules et les minuscules qui est un indicateur précieux pour identifier les noms propres dans les langues utilisant l'alphabet latin. D'autres facteurs se conjuguent pour rendre l' identification et le typage des EN délicats. Mentionnons en particulier l'utilisation de noms communs comme (parties de) prénoms ou noms : ainsi نور (Nour), qui signifie également *lumière*, فرحَان (Farhan) qui est aussi un adjectif (*content*), ou l'utilisation du préfixes comme عبد (Abd) (*serviteur de*) associé à un nom décrivant Dieu[5] ou encore بن (Ben, *fils de*) dans de nombreux noms de personnes d'Afrique du Nord. Mentionnons ensuite la possibilité de préfixer des noms propres également par des articles ou prépositions. L'instabilité orthographique des noms propres dans les diverses régions arabophones, comme celle de leur translittération dans les langues utilisant l'alphabet latin, (ainsi le prénom جميلة traduit par *Jamila* en tunisien, *Djamila* en algérien, *Gamila* en égyptien), constitue une autre source de difficulté, qui tend à réduire l'utilité des répertoires de noms propres (aussi bien pour la détection que pour la traduction), lorsque du moins ceux-ci existent[6].

Nous proposons, dans cet article, une étude complète sur les EN en arabe dans un contexte de traduction automatique. Cette étude présente deux contributions principales : la première est le développement et l'analyse d'un système de détection d'EN pour l'arabe, combiné avec un outil de segmentation morphologique (Gahbiche-Braham *et al.*, 2012). Les EN sur lesquelles nous travaillons correspondent aux trois grandes classes d'entités usuelles : lieu, personne et organisation (Chinchor et Ro-

5. Il existe 99 noms décrivant Allah. Ils peuvent tous être des noms propres, comme par exemple كريم (Karim) traduit par *généreux*, ou aussi عبد الكريم (Abdelkarim).

6. Il n'existe pas, à notre connaissance, de *gazeeter* pour l'arabe et tous les dictionnaires utilisés dans cette étude ont été constitués au cours du projet SAMAR qui est présenté en section 2.

binson, 1997). La seconde contribution correspond à l'implantation et à l'analyse de diverses stratégies pour intégrer ces EN dans un système de traduction statistique.

Le reste de l'article est organisé comme suit. Dans la section 2, la problématique et l'objectif de ces travaux ainsi qu'une description de toutes les ressources utilisées sont présentés. Une analyse de la littérature relative à l'identification et à la traduction des EN en arabe est ensuite exposée (section 3). La section 4 est consacrée à la description de la langue arabe et des outils de traitement automatique que nous avons développés pour la traiter. Nous proposons dans la section 5 notre système de base pour la détection des EN ainsi que l'ensemble des caractéristiques utilisées pour cette tâche. Une comparaison avec l'état de l'art ainsi qu'une adaptation du détecteur d'EN ont été également effectuées. La section 6 est consacrée à la traduction des EN. La méthode que nous avons proposée pour traduire les EN détectées est décrite dans cette section ainsi que l'ensemble des expériences effectuées sur deux corpus de test différents. Une analyse détaillée des résultats a été également effectuée. Finalement, la section 7 conclut ces travaux et discute quelques perspectives.

2. Contexte et motivations

Les travaux que nous présentons dans cet article s'inscrivent dans le cadre du projet SAMAR[7], qui vise à développer une plate-forme de traitement de dépêches en langue arabe. Les données sont principalement des dépêches journalistiques produites par le bureau arabe de l'Agence France-Presse (AFP). Le système de traduction de base utilisé dans ces travaux est entraîné sur un corpus parallèle arabe-français de 265 000 phrases extraites entre décembre 2009 et juillet 2012 avec la méthode présentée dans (Gahbiche-Braham *et al.*, 2011). Le mois de décembre 2010 ne fait pas partie de ces données puisqu'il a servi à extraire les corpus de test et de développement.

2.1. *Problématique*

Comme mentionné ci-dessus, il est fréquent qu'une EN à traduire ne soit pas présente dans les corpus parallèles qui servent à entraîner les systèmes de traduction. Dans Gahbiche-Braham *et al.* (2011), nous avons en effet observé, pour notre application, qu'environ 25 % des formes inconnues correspondent à des EN.

Une première analyse menée sur une extraction automatique des EN[8] sur le corpus AFP (250 dépêches par jour) permet d'illustrer ce phénomène, particulièrement présent dans des textes traitant de l'actualité. La figure 1 montre l'évolution du nombre d'EN nouvelles apparaissant chaque semaine pour la période de mars à juillet 2012 dans les dépêches AFP pour les trois catégories : lieu, personne et organisation.

7. SAMAR est l'acronyme de station d'analyse multimédia en langue arabe, `http://www.samar.fr`.

8. Les EN ont été extraites avec notre propre détecteur d'EN, qui sera présenté à la section 5.

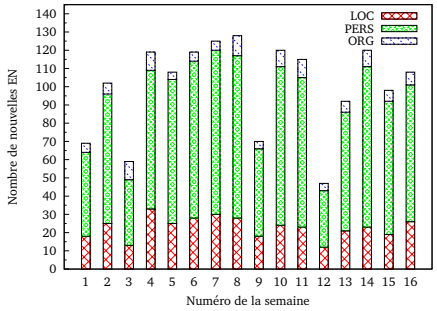

Figure 1. *Évolution du nombre d'EN nouvelles chaque semaine, de mars à juillet 2012*

À partir d'un nombre initial de 37 446 EN distinctes (extraites pendant la période de décembre 2009 à mars 2012), on observe que chaque semaine, de façon assez régulière, une centaine d'EN nouvelles apparaissent. Sur une période de 16 semaines, au total 1 599 nouvelles EN sont apparues : 1 122 personnes, 366 lieux, et 111 organisations. Non connues d'un système de traduction qui aurait été entraîné sur le corpus correspondant à la période initiale, ces EN seront nécessairement mal (ou pas) traduites.

Nous nous sommes donc intéressés à mettre en place, pour la traduction des noms propres, des stratégies ne nécessitant pas de devoir régulièrement réentraîner le système de traduction sur des corpus qui couvriraient toujours plus de données.

Afin d'avoir une idée sur le pourcentage d'erreurs de traduction des EN, nous avons utilisé le corpus parallèle Arcade II (Chiao *et al.*, 2006), qui montre la particularité d'être annoté en EN en arabe comme en français (voir ci-après une description de ce corpus), nous fournissant ainsi un moyen très simple d'évaluer la qualité de traduction des EN. Le corpus Arcade II a été traduit avec le système de traduction AFP, qui sera présenté au début de la section 2. Le tableau 1 présente le pourcentage des EN traduites correctement.

	LOC	PERS	ORG	**Total**
Système de base	70,38 %	51,24 %	44,63 %	55,41 %

Tableau 1. *Pourcentage d'EN bien traduites par le système de traduction AFP (calculé par rapport à la référence)*

Une vérification automatique des EN a été réalisée pour voir le taux d'EN traduites correctement. Pour chaque ligne, toutes les EN se trouvant dans la référence ont été recherchées dans la ligne correspondante de la sortie de traduction automatique. D'après le tableau 1, presque la moitié (45 %) des EN n'ont pas été traduites correctement.

2.2. *Objectifs*

Ce travail se focalise principalement sur deux objectifs. Le premier est d'étudier et de comparer différentes versions d'un modèle de base de détection d'EN : une première version du modèle de base, qui est ensuite améliorée en augmentant le nombre de descripteurs, puis une deuxième version, qui exploite une adaptation non supervisée.

Le deuxième objectif consiste à tester différentes méthodes de traduction des EN à l'aide de dictionnaires bilingues de noms propres pour améliorer la traduction automatique des EN et réduire le nombre de mots inconnus qui sont mal traduits.

2.3. *Description des données*

Dans nos travaux, deux types de corpus ont été utilisés : des corpus monolingues pour la tâche de détection des EN et des corpus bilingues pour la tâche de traduction.

2.3.1. *Corpus monolingues en arabe*

Trois corpus monolingues ont été utilisés : le corpus ANER, le corpus AFP et le corpus *Gold AFP*.

Corpus ANER : les expériences de détection des EN ont été réalisées sur le corpus ANER[9] (Benajiba *et al.*, 2007) composé de plus de 150 000 occurrences de mots (4 871 phrases). Il peut être considéré comme le corpus de référence pour la tâche de détection des EN.

Le corpus distingue quatre types d'EN : lieu (LOC : 40 % des EN observées), personne (PERS : 32 %), organisation (ORG : 18 %) et une classe « divers » regroupant tous les autres types (MISC : 10 %)[10]. Il utilise le schéma d'annotation BIO et distingue donc neuf étiquettes. La répartition en EN est présentée dans le tableau 2.

	LOC	ORG	PERS	MISC
Entités nommées	4 431	2 026	3 602	1 117
Entités nommées distinctes	1 004	657	1 446	437

Tableau 2. *Répartition des entités nommées dans le corpus ANER*

La figure 2 montre un extrait d'une phrase en format BIO. L'étiquette B-X (*Begin*) indique le premier mot d'une EN de type X. L'étiquette I-X (*Inside*) indique qu'un mot fait partie d'une EN mais qu'il n'est pas le premier. L'étiquette O (*Outside*) est utilisée pour les mots qui ne sont pas des EN.

9. `http://users.dsic.upv.es/~ybenajiba/downloads.html`
10. Seuls les trois premiers types sont utilisés dans nos évaluations.

وقال	ستيفان	دوجاريك	المتحدث	باسم	الامم	المتحدة	.
O	B-PERS	I-PERS	O	O	B-ORG	I-ORG	O
Et a dit	Stéphane	Dujarric	le porte-parole	des	Nations	unies	.

Figure 2. *Exemple d'un extrait de phrase en format BIO*

Corpus monolingue AFP : nous disposons dans ce cadre de ressources supplémentaires pour adapter la détection des EN :

– des données du domaine (AFP), non annotées (130 000 phrases, 3 500 K mots) ;

– un corpus de test, Gold AFP, annoté manuellement en EN (LOC, PERS et ORG) et constitué de 900 phrases issues des données de l'AFP.

Les dépêches AFP traitées dans notre application diffèrent substantiellement des données du corpus ANER, qui contient à la fois des articles de presse et des données collectées en ligne, en particulier des extraits de Wikipédia. Il existe également un décalage temporel entre la constitution du corpus ANER (2007) et les données que nous devons traiter, qui sont postérieures à 2009.

2.3.2. *Corpus bilingues*

Pour la tâche de traduction, le corpus de développement utilisé pour l'optimisation du système est constitué de 1 000 phrases extraites de dépêches AFP (décembre 2010). Pour évaluer notre méthode nous disposons de deux corpus de test : le corpus AFP (Gahbiche-Braham *et al.*, 2011) et le corpus Arcade II (Chiao *et al.*, 2006).

Corpus parallèle AFP : le corpus de test AFP est constitué de 1 000 phrases extraites de dépêches datées de décembre 2010 également : ces données sont représentatives de la tâche de traduction que nous essayons de traiter.

Corpus parallèle Arcade II : le corpus Arcade II est un corpus parallèle d'articles du *Monde diplomatique*, pour lequel les EN sont annotées à la fois en arabe et en français ; ces textes sont un peu plus éloignés des données de la tâche que le corpus précédent. Nous avons vérifié ce corpus manuellement pour réduire les EN à trois types. Des alignements ont été également corrigés ainsi que certaines phrases du corpus afin d'avoir le même nombre d'EN en source et en cible. Finalement, le corpus de test utilisé pour les expériences contient 1 312 lignes, soit 1 079 noms de lieux, 525 noms de personnes et 354 noms d'organisations.

2.3.3. *Dictionnaires d'EN*

Deux types de dictionnaires d'EN sont utilisés : les dictionnaires monolingues pour la détection des EN (voir la section 5.2) et les dictionnaires bilingues pour la traduction des EN (voir la section 6.1). Ces dictionnaires contiennent des entités nommées monolexicales et polylexicales pour chaque catégorie d'EN considérée. Ces dictionnaires sont constitués à partir de ressources accumulées durant le projet SAMAR ; le détail de ces ressources est présenté dans le tableau 3.

	Sources	LOC	PERS	ORG
Dictionnaires monolingues	ANERGazet	x	x	x
	Wikipédia	x	x	x
	Geonames[11]	x	-	x
	Total	4 920	18 098	1 043
Dictionnaires bilingues	ANERGazet	x	-	-
	Wikipédia	x	x	-
	JRC[12]	x	x	-
	Geonames	x	x	x
	Total	3 810	16 470	728

Tableau 3. *Constitution des dictionnaires monolingues et bilingues*

3. État de l'art

Les travaux sur le traitement automatique de l'arabe se sont multipliés ces dernières années (Habash, 2011). Dans cette section, nous passons brièvement en revue ceux qui sont les plus pertinents pour notre étude, en présentant dans un premier temps le problème du repérage des entités nommées, puis celui de la traduction automatique. Comme mentionné ci-dessus, les problèmes que posent les EN en traduction automatique depuis l'arabe se retrouvent peu ou prou lorsque l'on traduit depuis d'autres langues, ce qui justifie notre discussion de quelques études qui ne traitent pas spécifiquement de l'arabe.

3.1. *Étiquetage en EN*

3.1.1. *Repérage dans un cadre statique*

Les premiers travaux sur la reconnaissance des EN pour l'arabe remontent à 1998 et reposent sur des méthodes à base de règles (Maloney et Niv, 1998), voir également le travail plus récent de Shaalan et Raza (2009) ou encore de Zaghouani *et al.* (2010). Les expériences de Zitouni *et al.* (2005) reposent, quant à elles, sur l'utilisation des techniques d'apprentissage automatique (*Maximum Entropy Markov Models*) : en considérant des jeux de descripteurs idoines, les auteurs obtiennent de très bons résultats sur les données de la campagne ACE 2004.

Ces travaux sont prolongés en particulier par Benajiba et ses coauteurs et ont notamment donné lieu à la construction du corpus ANER. Dans une première approche (Benajiba et Rosso, 2007), un étiqueteur statistique estimé par maximisation d'entropie est exploré. Cette approche est ensuite étendue en décomposant la prédiction en

11. `http://download.geonames.org/export/dump/`
12. *Joint Research Center* de la Communauté européenne : `http://langtech.jrc.it/JRC-Names.html`

deux temps : d'abord les frontières de l'EN en introduisant des catégories morphosyntaxiques (POS), puis la détermination de son type. Une seconde approche (Benajiba et Rosso, 2008), fondée sur l'utilisation de champs aléatoires conditionnels (CRF, (Lafferty *et al.*, 2001)) a permis d'explorer l'intégration de l'ensemble des traits dans un modèle unique, amenant à de meilleures performances, essentiellement en termes de rappel. Benajiba *et al.* (2008) montrent également l'efficacité d'un pré-traitement des textes pour séparer les différents constituants du mot (préfixes, lemme, et suffixes). Abdul Hamid et Darwish (2010) intègrent des traits intramots (notamment n-grammes de caractères) dans un modèle CRF. Cette approche permet de capturer implicitement les caractéristiques morphosyntaxiques, qui sont introduites explicitement par l'analyse préalable réalisée dans les expériences de Benajiba et Rosso (2008).

Dans des travaux plus récents, Al-Jumaily *et al.* (2012) présentent un système de détection des EN fondé sur une architecture à base de motifs, qui permet de détecter en temps réel l'apparition de certaines EN et des événements dans des dépêches. Ce système peut être utilisé pour les applications Web.

Mentionnons pour finir les travaux de Samy *et al.* (2005), qui utilisent un corpus parallèle arabe et espagnol pour extraire des EN en arabe. Ces auteurs repèrent les EN en espagnol à l'aide d'un étiqueteur à base de règles et d'un lexique monolingue ; les EN sont ensuite translittérées vers l'arabe. L'avantage principal de cette approche est que le corpus parallèle joue un double rôle : c'est à la fois une ressource et une cible. Cette technique de projection d'EN translingue a depuis été utilisée à plusieurs reprises : ainsi Zhang (2012) comme Che *et al.* (2013) utilisent des données parallèles chinois et anglais pour améliorer les performances d'un détecteur d'EN en chinois (voir également la section 3.2, qui présente diverses études portant sur l'extraction de corpus des EN bilingues). Elle repose sur l'hypothèse qu'une EN en langue source se traduit presque toujours par une EN en langue cible, hypothèse que nous exploiterons en utilisant le corpus Arcade II.

3.1.2. *Adaptation de systèmes de détection des EN*

En apprentissage automatique, l'adaptation consiste à développer un système de traitement pour un domaine cible à partir de données appartenant à un domaine source. D'un point de vue statistique, cela implique que les distributions des exemples observés sont différentes au moment de l'apprentissage et au moment du test.

Cette problématique a fait l'objet de multiples propositions en modélisation statistique des langues (par exemple l'étude de Bellagarda (2001) pour les modèles statistiques de langue), ou encore les états de l'art produits par Daumé III et Marcu (2006) et par Blitzer (2008) : combinaison linéaire de systèmes entraînés sur les domaines source et cible, utilisation de pondérations différentielles pour les exemples des domaines source et cible (Jiang et Zhai, 2007), utilisation de descripteurs spécifiques pour les exemples source et cible (Daumé III, 2007), etc. On se reportera à (Daumé III *et al.*, 2010) pour un échantillon de travaux récents. Dans un cadre non supervisé, la stratégie la plus commune est l'autoapprentissage (*self-training*), qui consiste à gé-

nérer automatiquement des données d'apprentissage pour le domaine cible à partir du système source (Mihalcea, 2004).

Concernant le repérage des EN, le problème de l'adaptation se pose avec une acuité particulière car les EN sont souvent associées avec un thème particulier et ont également des distributions d'occurrences très variables dans le temps. Ce problème est étudié en particulier par Béchet *et al.* (2011) qui (i) combinent deux approches d'étiquetage en EN (l'une symbolique, l'autre probabiliste) pour le français, puis (ii) adaptent le système probabiliste fondé sur un processus discriminant à base de CRF, au domaine des données de test.

Plusieurs travaux se sont également focalisés sur l'adaptation temporelle du vocabulaire d'une application, comme ceux de Allauzen (2003) qui portent sur l'adaptation du vocabulaire et du modèle de langue d'un système de transcription automatique par fouille de flux d'information textuels sur Internet.

3.2. *Traduction des EN*

L'amélioration de la traduction des mots inconnus, et en particulier des EN, a également fait l'objet de nombreux travaux. Une stratégie minimaliste, principalement pertinente pour les noms de lieux et de personnes, consiste à les translittérer en alphabet latin. C'est ainsi que Al-Onaizan et Knight (2002b) utilisent des ressources bilingues et monolingues et combinent à la fois la translittération (Al-Onaizan et Knight, 2002a) avec la traduction par accès dictionnairique, afin de trouver la meilleure traduction en anglais d'une entité nommée initialement en arabe. Une approche similaire (traduction par accès dictionnairique et par translittération) est utilisée par Hassan et Sorensen (2005), qui s'intéressent plus spécifiquement à l'intégration de ces techniques au sein d'un système statistique. La même préoccupation anime Kashani *et al.* (2007), qui évaluent comment au mieux combiner prétraduction ou prétranslittération dans un système statistique à base de segments, et montrent qu'il est possible d'améliorer la qualité de la traduction en proposant des translittérations de tous les mots inconnus.

Notons que l'utilisation de translittérations peut s'avérer utile dès l'apprentissage : par exemple Santanu *et al.* (2010) translittèrent les EN de la langue source (anglais) vers la langue cible (bengali), afin de les aligner ; obtenir de meilleurs alignements pour les EN permet d'améliorer les performances de la traduction automatique jusqu'à 4,6 points BLEU[13].

13. BLEU est une mesure de précision, dont le principe est de calculer le degré de similitude entre une traduction automatique et une ou plusieurs références en se basant sur la précision n-gramme : si une traduction automatique est identique à une des références, alors le score BLEU est égal à 100. En revanche, si aucun des n-grammes de la traduction n'est présent dans aucune référence, alors le score BLEU est égal à 0.

Translittérer inconditionnellement tous les mots inconnus est pourtant une stratégie risquée, comme le montre l'étude déjà citée de Habash (2008). L'auteur y développe une approche dans laquelle il améliore la traduction des mots inconnus en arabe en étendant un modèle de traduction préexistant d'une façon qui prend en compte les différents types de mots inconnus. Pour les verbes et noms communs, des traductions nouvelles sont ajoutées en mettant en correspondance chaque mot inconnu avec une variation morphologique et/ou orthographique de ce mot. Des traductions extraites de dictionnaires créés manuellement – à partir du glossaire de BAMA (Buckwalter, 2002) – ainsi que des translittérations d'une liste de noms propres collectée automatiquement sont également utilisées pour améliorer le modèle de traduction.

En fait, il s'avère même que traiter de manière uniforme toutes les EN peut dégrader globalement la qualité de la traduction, comme le montrent Hermjakob *et al.* (2008) : ceci est causé par les erreurs du système de détection des EN, combinées aux erreurs du système de translittération, mais également par l'existence de formes qui ne doivent être que partiellement translittérées, etc. ; comme le proposent les auteurs, il est donc nécessaire de choisir avec soin les EN qui seront translittérées.

Pour limiter ces erreurs, la principale alternative consiste à accumuler des dictionnaires bilingues d'EN en utilisant des méthodes de fouille de textes translingues, appliquant aux EN des techniques qui se sont avérées utiles pour d'autres unités (cf. (Rapp, 1995) pour la découverte de traductions de mots inconnus, ou encore (Fung et McKeown, 1997) pour l'acquisition de traductions de termes). Ainsi, Moore (2003) effectue ce travail en exploitant un corpus parallèle : les entités nommées (dans ce cas dans un domaine technique) sont extraites du corpus en langue source (anglais) ; des mesures d'association statistiques sont ensuite utilisées pour décider quelles sont les séquences de mots espagnols ou français qui correspondent le mieux aux segments extraits. Plus proche de nos préoccupations, Hassan *et al.* (2007) enrichissent les systèmes de traduction avec des EN et leurs traductions extraites à la fois de corpus parallèles et comparables ; une approche similaire est également adoptée par Abdul Rauf (2012), qui propose d'améliorer la traduction des EN en prolongeant le corpus parallèle avec des dictionnaires constitués à partir de mots inconnus et de leurs traductions extraites de corpus comparables, en utilisant des techniques de recherche d'information. Des principes voisins sont enfin utilisés pour fouiller des corpus toujours moins parallèles, notamment par Jiang *et al.* (2007), qui combinent la translittération avec des données extraites du Web et choisissent ainsi la meilleure hypothèse de traduction, et par Ling *et al.* (2011), qui utilisent des liens Web pour récupérer les traductions des EN et enrichir les systèmes de traduction.

Mentionnons finalement le travail récent de Fehri *et al.* (2011), qui repose sur le développement et l'utilisation de ressources linguistiques riches : ces auteurs proposent des traductions isolées d'EN à l'aide de transducteurs, de dictionnaires bilingues, de translittérations et d'un ensemble de règles de réordonnancement ; l'utilisation de ces diverses ressources linguistiques permet d'atteindre une très bonne précision.

L'architecture que nous proposons dans la suite intègre une partie des méthodes de la littérature, en se focalisant plus spécifiquement sur (i) l'identification des EN en

prétraitement, dans un contexte d'adaptation au domaine, et (ii) l'intégration de dictionnaires bilingues dans un système statistique. Ni l'intégration de translittérations, ni l'acquisition de nouveaux dictionnaires bilingues ne sont, *a contrario*, considérées ici, même si notre système pourrait certainement bénéficier de l'utilisation de ces deux techniques. Notre démarche emprunte donc beaucoup à celle de Hálek *et al.* (2011) qui cherche à améliorer la traduction automatique des EN depuis le tchèque vers l'anglais. Les EN sont détectées en utilisant des CRF, puis des propositions de traductions extraites de Wikipédia sont proposées. Les résultats montrent une baisse de la qualité de traduction (en BLEU) même si des annotateurs humains jugent cette évolution positivement. L'utilisation d'un deuxième modèle de traduction permet d'améliorer les performances quantitatives (en BLEU) par rapport au modèle de base.

4. Prétraitements de l'arabe

La langue arabe est une langue sémitique, qui a la particularité d'avoir un vocabulaire à base de racines de mots trilitères consonantiques. À cette forme de base, peuvent s'ajouter des préfixes, des suffixes, ainsi que des clitiques. Les clitiques et les affixes sont agglutinés au mot de base pour former d'autres mots de plus en plus complexes, voire des phrases entières. Par exemple la forme وسيأكلها, (*Et il va la manger*), est constituée de deux proclitiques و (et) et سـ (marque du futur, *va* dans cet exemple), un préfixe يـ (marque de l'inaccompli et de la troisième personne du singulier *il*) et un enclitique ها (pronom *la*).

L'absence des voyelles dans les textes arabes engendre une certaine ambiguïté en ce qui concerne le sens du mot d'une part, et augmente la difficulté à identifier sa fonction dans la phrase, d'autre part.

Dans (Gahbiche-Braham *et al.*, 2012), nous avons décrit un outil de prétraitement de l'arabe – SAPA (*Segmentor and Part-of-speech tagger for Arabic*) – à base de CRF. Ce dernier utilise des modèles de prédiction de préfixes et d'étiquettes morphosyntaxiques entraînés par Wapiti (Lavergne *et al.*, 2010) sur l'*Arabic Tree Bank* (Maamouri *et al.*, 2005a ; Maamouri *et al.*, 2005b). SAPA a été comparé aux deux outils les plus connus de prétraitement de l'arabe (MADA (Habash *et al.*, 2009) et MorphTagger (Mansour, 2010)). Nous avons montré que SAPA est aussi performant que ces deux outils de prétraitement pour la tâche de traduction automatique. Les

particularités de SAPA sont (i) la vitesse dela prétraitement (trente fois plus rapide que MADA et 30 % plus rapide que MorphTagger) et (ii) l'indépendance vis-à-vis de toute autre ressource externe d'analyse morphologique ou de désambiguïsation : en particulier il ne nécessite pas d'utiliser l'analyseur de Buckwalter (2002).

Le prétraitement consiste en quatre étapes principales : le texte en arabe est tout d'abord transcodé en utilisant le schéma de Buckwalter ; il est ensuite analysé morphosyntaxiquement par un premier CRF, puis traité par un second CRF qui utilise la connaissance des parties du discours pour reconnaître les frontières d'un certain nombre de proclitiques et d'affixes. Sur la base de ce double étiquetage, les décisions de segmentation sont alors réalisées par application d'un petit ensemble de règles déterministes ; diverses normalisations sont également appliquées à ce stade. Notre prétraitement permet en particulier de (i) séparer les proclitiques و (et), ف (alors), ب (avec), ك (comme), ل (pour) et س (marque du futur) de la forme de base, et (ii) de produire un texte étiqueté en parties du discours (sur la base de vingt-quatre étiquettes morphosyntaxiques) (Gahbiche-Braham *et al.*, 2012). En revanche, les enclitiques (corrrespondant à des pronoms personnels agglutinés aux formes de base) ne sont pas segmentés : lors du développement de SAPA, nous avons choisi de ne pas les séparer car ce traitement supplémentaire ne semblait pas améliorer la traduction.

5. Détection des entités nommées

À la suite de nombreux travaux, nous abordons cette tâche avec des outils d'apprentissage automatique et utilisons le modèle des CRF, tels qu'implémentés dans le logiciel Wapiti[14]. Cet outil permet d'utiliser de très gros modèles incorporant des centaines d'étiquettes et des centaines de millions de descripteurs, ainsi qu'une stratégie d'optimisation permettant de sélectionner les descripteurs les plus utiles par le biais d'une pénalité L_1 (Sokolovska *et al.*, 2009). De manière standard, le repérage d'EN est effectué en utilisant une représentation BIO des frontières d'EN ; trois types d'EN étant distingués, le modèle prédit à chaque position une des dix balises différentes.

L'utilisation de modèles statistiques pose la question de la pertinence des corpus d'apprentissage au regard des données de test. Nous traitons cette question en explorant une adaptation non supervisée. Le protocole expérimental est exposé dans la

14. `http://wapiti.limsi.fr`

section 5.1. Les caractéristiques sélectionnées pour la détection des EN sont décrites dans la section 5.2. Le système de détection des EN de base est détaillé dans la section 5.3. La section 5.4 montre les expériences d'adaptation du système de détection des EN. Une évaluation de la détection des EN sur le corpus Arcade II est réalisée dans la section 5.5.

5.1. *Protocole expérimental*

Les expériences sont réalisées à partir de données translittérées[15] et segmentées avec l'outil de segmentation SAPA décrit ci-dessus (section 4). Les scores sont calculés en utilisant l'outil d'évaluation développé pour la tâche de repérage des EN proposée dans le cadre de CoNLL 2002[16], qui calcule le rappel, la précision et la F-mesure comme suit :

$$\text{Précision} = \frac{100 * C}{T} \qquad \text{Rappel} = \frac{100 * C}{TC}$$

$$\text{F}_{\beta = 1} = \frac{2 * Precision * Rappel}{Precision + Rappel}$$

où *C* représente le nombre des étiquettes correctes pour une EN, *T* le nombre des étiquettes trouvées, et *TC* le nombre des EN présentes dans la référence.

Les modèles sont évalués sur le corpus ANER par validation croisée à 10 partitions, sur des tests d'environ 25 000 mots chacun.

5.2. *Sélection de caractéristiques*

Différentes versions du modèle de détection des EN ont été développées, qui incluent des jeux de descripteurs de richesse croissante. Nous décrivons ci-dessous les principales familles de descripteurs ; chaque réalisation x d'un trait d'une de ces familles donne lieu à un ensemble de fonctions booléennes testant x avec chaque étiquette et avec chaque bigramme d'étiquettes possible.

Ponctuation et nombres : ce trait teste la présence de caractères de ponctuation et de chiffres dans le mot courant ainsi que dans les deux mots voisins.

N-grammes de mots : ces caractéristiques testent tous les unigrammes, trigrammes et tétragrammes dans une fenêtre de taille 5 centrée autour de la position courante, ainsi que tous les bigrammes dans une fenêtre de taille 3.

Préfixes et suffixes : chaque séquence d'une, deux, ou trois lettres observée à l'initiale ou à la finale d'un mot du corpus d'apprentissage donne lieu à un ensemble

15. `http://www.qamus.org/transliteration.htm`
16. `http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt`

de traits. L'apparition de ces préfixes et suffixes est testée dans une fenêtre de taille 5 centrée sur le mot courant.

POS-tags : ce trait concerne les étiquettes morphosyntaxiques prédites en utilisant un modèle entraîné par Wapiti sur l'*Arabic Tree Bank*[17] ; des détails sur cet étiqueteur sont donnés dans (Gahbiche-Braham *et al.*, 2012). Les tests évaluent les unigrammes de POS-tags dans une fenêtre de taille 5, et les bigrammes de POS-tags dans une fenêtre de taille 3.

Dictionnaires monolingues : plusieurs traits s'appuient sur l'exploitation des dictionnaires monolingues : pour chaque mot w, on teste si w est une entrée simple du dictionnaire (**Dict**), ou bien s'il y figure précédé d'un ou deux proclitiques contenant un ou deux caractères (**Pref.Dict** sur la figure 3). En langue arabe, il est en effet possible d'avoir des EN précédées par la conjonction *et* (و) – constituée d'un seul caractère et collée au mot suivant –, ou à la fois précédées simultanément par une conjonction et une préposition, comme par exemple ولفرنسا (*et pour la France*). On teste enfin si w apparaît, comme composant d'une unité polylexicale, et si w, sans ses éventuels proclitiques, est le premier mot d'une unité polylexicale.

Les résultats de ces expériences sont reportés sur la figure 3, qui représente la variation globale de la précision, du rappel et de la F-mesure, ainsi que le nombre de traits actifs. On constate qu'au fur et à mesure que de nouveaux traits sont ajoutés au modèle précédent, le rappel et la F-mesure augmentent, parfois au prix d'une légère dégradation de la précision. Il est à noter que les traits relatifs à la *ponctuation et aux nombres* sont utilisés dans tous les modèles.

5.3. *Système de détection des EN de base et comparaison avec l'état de l'art*

Dans cette section, nous présentons les performances du système de détection des EN de base (NERAr pour *Named Entity Recognition for Arabic*). Ce dernier est entraîné sur des données segmentées avec SAPA et englobe toutes les caractéristiques présentées dans la section 5.2. Le tableau 4 résume les performances de NERAr sur le corpus ANER en comparaison avec les performances du système de base de Benajiba et Rosso (2008). Ce dernier est fondé sur les CRF et utilise des données segmentées avec l'outil de segmentation de Diab (2009) pour la détection des EN.

Les résultats de ces premières expériences NERAr donnent des performances comparables à l'état de l'art. La segmentation du texte, avant l'entraînement du modèle de détection des EN, améliore les résultats, puisqu'une évaluation du système de base sans segmentation préalable des mots du texte a donné une $F_{\beta=1}$ totale égale à $81,96$.

17. `http://www.ircs.upenn.edu/arabic/`

18. Le total reporté dans (Benajiba et Rosso, 2008) inclut l'EN MISC. Le total ici a été fait en calculant la moyenne pour les trois catégories.

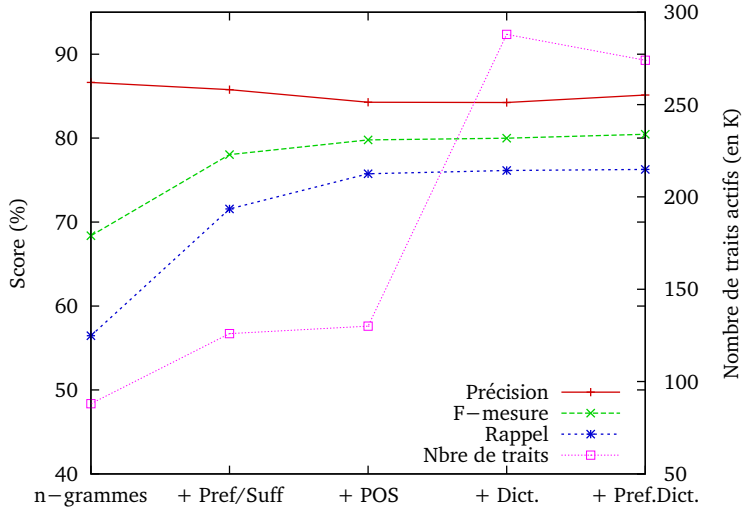

Figure 3. *Précision (en %), rappel (en %), F-mesure et nombre de traits actifs pour des modèles de complexité croissante (à chaque nouveau modèle, de nouveaux traits sont ajoutés)*

(Benajiba et Rosso, 2008)				Système de base (NERAr)			
	Precision	Recall	$F_{\beta=1}$		Précision	Rappel	$F_{\beta=1}$
LOC	93,03 %	86,67 %	89,74	LOC	89,75 %	89,39 %	89,57
ORG	84,23 %	53,94 %	65,76	ORG	83,37 %	66,01 %	73,68
PERS	80,41 %	67,42 %	73,35	PERS	82,49 %	78,53 %	80,46
Total	85,89 %	69,34 %	76,28[18]	Total	86,02 %	80,78 %	83,32

Tableau 4. *Précision, rappel et F-mesure du système de détection des EN de base (NERAr) sur le corpus ANER (en validation croisée) en comparaison avec le système de détection des EN de Benajiba et Rosso (2008)*

5.4. *Adaptation non supervisée du système de détection des EN*

Pour adapter le détecteur d'EN, nous avons utilisé le corpus d'apprentissage constitué de données du domaine (AFP) ; les évaluations sont alors réalisées en utilisant un corpus de test monolingue *Gold AFP* constitué de dépêches annotées manuellement en EN.

Le système de base, construit à partir du corpus ANER, est utilisé pour annoter automatiquement le corpus AFP. Deux systèmes adaptés sont alors obtenus en utilisant comme corpus d'entraînement (AFP) soit le corpus étiqueté automatiquement

seul, soit l'union des deux corpus (ANER + AFP). Le tableau 5 donne les résultats

	Précision	Rappel	$F_{\beta=1}$	Modèle
LOC	91,30 %	80,31 %	85,46	ANER
ORG	52,20 %	33,81 %	41,04	
PERS	70,21 %	71,74 %	70,97	
Total	80,07 %	69,77 %	74,57	
LOC	90,57 %	82,11 %	86,14	AFP
ORG	55,44 %	38,08 %	45,15	
PERS	71,51 %	69,57 %	70,52	
Total	80,55 %	71,07 %	75,51	
LOC	87,20 %	85,04 %	86,10	ANER + AFP
ORG	49,79 %	41,99 %	45,56	
PERS	71,16 %	73,10 %	72,12	
Total	77,13 %	74,32 %	75,70	

Tableau 5. *Comparaison et adaptation du système de reconnaissance d'entités nommées sur le corpus de test* Gold AFP

des trois systèmes sur les données de test *Gold AFP*. Le changement de domaine entraîne une baisse sensible des performances du système de base (la F-mesure passe de 83,32 sur les données de test ANER à 74,57 sur les données de test *Gold AFP*). Après adaptation (AFP et ANER + AFP), on constate une amélioration du rappel et de la $F_{\beta=1}$ pour tous les types d'EN. On note que, bien que le modèle AFP ait été annoté automatiquement, il conduit à de meilleurs scores que le modèle ANER.

Les tailles des deux corpus sont assez différentes : environ 5 000 phrases pour le corpus ANER et 130 000 phrases pour le corpus AFP. Les données AFP sont annotées automatiquement, ce qui peut engendrer du bruit. Un test supplémentaire a été effectué en sélectionnant seulement les 5 000 phrases qui ont le meilleur score de confiance donné par Wapiti. La F-mesure totale obtenue pour le modèle limité ANER + AFP est de 74,90. On constate que plus on a de données d'adaptation, même imparfaitement annotées, plus la détection des EN s'améliore.

5.5. *Détection des EN pour le corpus Arcade II*

Le corpus Arcade II annoté en EN est différent à la fois du corpus ANER et du corpus AFP. Afin d'avoir une idée des performances du détecteur d'EN sur le corpus Arcade, – qui sera, par la suite, utilisé pour évaluer l'impact de la traduction des EN – nous avons évalué notre modèle de détection des EN de base sur ce corpus.

Le tableau 6 montre les performances du système de détection des EN de base ainsi que celles du système adapté (ANER + AFP) sur le corpus de test Arcade II. On constate que le modèle adapté ANER + AFP donne de meilleures performances que le modèle ANER seul sur le corpus Arcade II en termes de rappel et de F-mesure. Nous avons donc choisi le modèle adapté pour détecter les EN sur les corpus de test AFP et Arcade.

Modèle ANER				Modèle ANER + AFP			
	Precision	Recall	$F_{\beta=1}$		Précision	Rappel	$F_{\beta=1}$
LOC	81,69 %	74,50 %	77,93	LOC	78,75 %	78,90 %	78,83
ORG	71,67 %	22,22 %	33,93	ORG	60,54 %	28,94 %	39,16
PERS	70,47 %	77,92 %	74,01	PERS	69,61 %	79,04 %	74,02
Total	77,08 %	65,38 %	70,75	Total	74,01 %	69,35 %	71,60

Tableau 6. *Précision, rappel et F-mesure sur le corpus Arcade II en utilisant le modèle de détection des entités nommées ANER puis ANER + AFP*

5.6. *Bilan de la détection des EN*

Dans cette section, nous avons présenté le système de détection d'EN qui sera utilisé pour préparer les textes à traduire. Les expériences réalisées dans cette section montrent que le système atteint des performances comparables à celles des systèmes statistiques de l'état de l'art. On observe également une dégradation très sensible des résultats lorsqu'on le teste sur des données qui proviennent de sources différentes de celles des données d'entraînement, dégradation que l'autoadaptation permet de limiter. On notera finalement que pour l'arabe, le problème du repérage automatique des EN reste mal résolu, puisque, bien que l'on ne considère qu'un étiquetage très grossier (distinguant trois catégories), les meilleures performances plafonnent autour de 80 de F-mesure.

6. Traduction automatique

Pour la traduction automatique, nous utilisons le décodeur à base de segments Moses[19] (Koehn *et al.*, 2007), ainsi que l'aligneur sous-phrastique MGIZA++[20] (Gao et Vogel, 2008) pour la phase d'entraînement. La table de traduction est constituée en rendant symétriques les alignements selon l'heuristique *grow-diag-final-and* de Moses, et contient des segments dont la longueur va jusqu'à sept mots. L'outil SAPA a été utilisé pour le prétraitement de l'arabe (section 4). La section 6.1 présente la méthode d'intégration des traductions des EN à partir de dictionnaires. Les expériences et résultats sont présentés dans la section 6.2.

6.1. *Intégration de dictionnaires pour la traduction des EN*

L'idée générale de l'architecture que nous étudions est d'utiliser la détection d'EN lors d'un prétraitement, et de prétraduire par accès dictionnairique les EN qui ont été

19. `http://www.statmt.org/moses/`
20. `http://www.kyloo.net/software/doku.php/mgiza:overview`

détectées ; les autres tokens sont passés tels quels au système de traduction statistique. La figure 4 détaille ces différentes étapes de prétraitement.

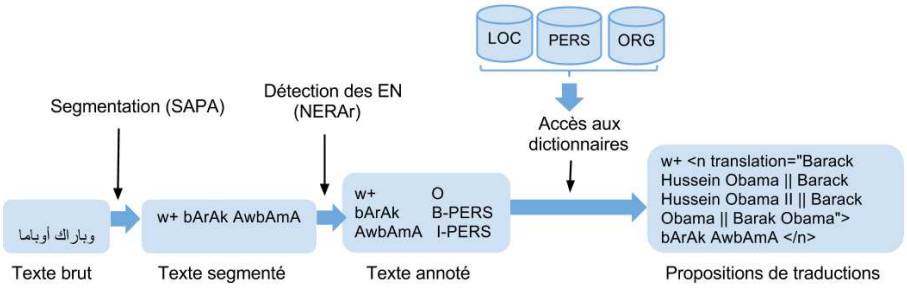

Figure 4. *Méthode de prétraitement du texte en arabe en combinant segmentation du texte, détection des EN et propositions de traductions*

Lors de la première phrase de traitement, le texte en arabe est segmenté et les EN sont détectées et étiquettées. Selon le type d'EN détectée, le dictionnaire bilingue approprié est consulté (lieu, personne et organisation) afin d'éviter les ambiguïtés : un nom de rue (LOC) peut être identique au nom d'une personne (PERS), comme par exemple *Charles-de-Gaulle* qui peut être un nom de rue, un nom d'aéroport ou un nom de personne.

Les propositions de traductions – extraites des dictionnaires – de cette EN sont ajoutées dans le texte à traduire sous forme de balises. Par exemple, pour le nom propre باراك اوباما (*Barack Obama*) – donné en entrée sous forme d'un texte transcodé avec le schéma d'encodage de Buckwalter (bArAk AwbAmA) –, quatre prétraductions possibles, qui rendent compte de la variabilité des formes écrites de ce nom propre, sont proposées au décodeur ; elles sont représentées sous le format suivant[21] :
<n translation="Barack Hussein Obama||Barack Hussein Obama II||Barack Obama||Barak Obama"> bArAk AwbAmA </n>

21. Chaque prétraduction peut également recevoir une probabilité ; lorsque ces probabilités ne sont pas explicitées, une distribution uniforme est appliquée.

Ces informations sont proposées à Moses en utilisant l'option XML selon deux modes :

– le mode *exclusive* impose au décodeur de choisir, pour le segment concerné, une des suggestions proposées en prétraduction ; notons que cela impose également au décodeur de découper l'entrée de façon à ce que le segment balisé soit considéré comme une unité de traduction. Les segments de la table de traduction qui seraient en concurrence avec ce segment sont ignorés ;

– le mode *inclusive* permet de mettre en concurrence les prétraductions avec toutes les entrées de la table de traduction ; pour chaque segment de la phrase à traduire, les scores des différentes hypothèses se trouvant dans la table de traduction, ainsi que ceux des hypothèses fournies en prétraduction, sont calculés. Toutes ces hypothèses sont ensuite comparées, afin de choisir celle qui conduit au meilleur score global.

6.2. *Expériences et résultats*

Les expériences sont effectuées sur deux corpus différents : le corpus de test AFP (section 2.3.2), et le corpus de test Arcade II annoté en EN à la fois en arabe et en français. Le corpus Arcade permet de vérifier plus précisément l'impact du traitement spécifique des EN sur leur traduction automatique.

Ainsi, nous avons évalué comparativement trois situations : sans traitement *a priori* des EN (configuration *default*), ou en proposant des traductions soit en mode *exclusive* soit en mode *inclusive*. Dans les trois configurations, un deuxième paramètre est utilisé pour le modèle de langue, permettant de pénaliser la génération de mots inconnus[22].

Des évaluations automatiques utilisant les métriques BLEU (Papineni *et al.*, 2002) et METEOR (Banerjee et Lavie, 2005 ; Lavie et Agarwal, 2007) ont été réalisées : rappelons que ces métriques se fondent sur des comparaisons de surface entre les traductions proposées par le système et des références humaines. Des évaluations manuelles ainsi que des analyses détaillées des résultats ont également été effectuées.

6.2.1. *Tests sur le corpus AFP*

Le tableau 7 présente les résultats de traduction automatique du corpus AFP pour les trois configurations *default* (système de base), *exclusive* et *inclusive* avec deux mesures de traduction[23] BLEU et METEOR (ce dernier intègre des équivalents sémantiques des EN). Le nombre de mots inconnus (#mots unk) est également présenté.

22. En traduction automatique, les modèles de langue cible doivent pouvoir évaluer des phrases contenant des mots inconnus ; ce paramètre permet donc de contrôler la probabilité assignée à une forme inconnue, qui, par définition ne peut être directement mesurée par décompte sur corpus.

23. Une augmentation de la qualité de la traduction se traduit par une augmentation des mesures BLEU et METEOR.

	#mots unk	BLEU	METEOR
Default	288	34,62	52,64
Exclusive	285	33,58	51,65
Inclusive	285	34,21	52,39

Tableau 7. *Scores BLEU et METEOR en traduction arabe-français sans (*default*) et avec (*exclusive *et* inclusive*) propositions de prétraductions sur le corpus de test AFP, constitué de* 1 000 *phrases extraites de dépêches datées de décembre 2010*

Les mesures BLEU et METEOR ne permettant pas d'apprécier les éventuelles améliorations de la traduction des EN, nous avons vérifié manuellement les 500 premières lignes du corpus de test.

Évaluation manuelle : la vérification manuelle a consisté à comparer la sortie des trois systèmes (*default*, *inclusive*, et *exclusive*) ; pour chaque différence nous avons vérifié la source et la référence pour savoir laquelle des traductions proposées était la meilleure. Sur les 500 lignes analysées, 871 EN ont été détectées, parmi lesquelles 60 % ont été trouvées dans les dictionnaires et ont pu être prétraduites. Si l'on compare les systèmes *default* et *exclusive*, on observe des différences pour seulement 48 entités nommées (5 % des cas) : on rencontre 13 cas où des erreurs de traduction d'EN du système *default* sont corrigées, conformément à la référence par le système *exclusive*, et 35 autres cas pour lesquels la traduction proposée par le système *exclusive*, bien qu'acceptable, diffère de celle de la référence. Ces configurations sont donc considérées comme erronées (à tort) par les mesures automatiques. Ainsi بيلاروسيا est traduit par *Bélarus* par le système *default*, ce qui est la traduction existante dans la référence, alors que l'option *exclusive* propose *Biélorusse*, qui est plus correct en français. L'analyse manuelle permet donc de conclure que pour l'ensemble de différences observées (soit 5 % des EN), le système *exclusive* améliore effectivement la traduction des EN.

La comparaison de la sortie de traduction *default* avec la sortie *inclusive* va dans le même sens, avec un nombre de différences observées plus réduit, correspondant à seulement 35 cas : 12 où la nouvelle traduction améliore le BLEU, et 23 cas où la traduction proposée est correcte, mais diffère de celle de la référence. Nous observons notamment que les EN agglutinées à des conjonctions (comme والمانيا, *et l'Allemagne*) et qui n'ont pas été traduites dans le cas *default*, ont été traduites correctement par les systèmes *inclusive* et *exclusive* : 7 erreurs de ce type ont été ainsi supprimées.

Au final, si peu de différences sont observées entre les systèmes, les propositions des dictionnaires sont toujours correctes. Elles sont cependant trop rares pour influer positivement sur les scores automatiques. En effet, forcer la segmentation de la phrase source, voire imposer l'insertion de certains mots inconnus du modèle de langue dans la cible, peut avoir un effet négatif sur la traduction du voisinage de ce mot. Ainsi, le système *default* propose la traduction *Séoul a annoncé que ces manœuvres [...]* alors que la meilleure hypothèse du système *exclusive* est *Séoul a indiqué que ces manœuvres [...]*, qui est induite par une autre segmentation de la source. Le système *default* obtient ici un meilleur score car le mot *indiqué* ne fait pas partie de la référence.

Dictionnaires : la qualité de la traduction des EN est dépendante de leur détection. L'évolution du nombre d'EN détectées varie en fonction de la taille des dictionnaires. Ceci est reflété par les résultats représentés sur la figure 5, qui montre l'évolution du nombre d'EN du corpus de test AFP donnant lieu à une prétraduction en fonction de la taille des dictionnaires. Nous avons initialement utilisé des dictionnaires monolingues

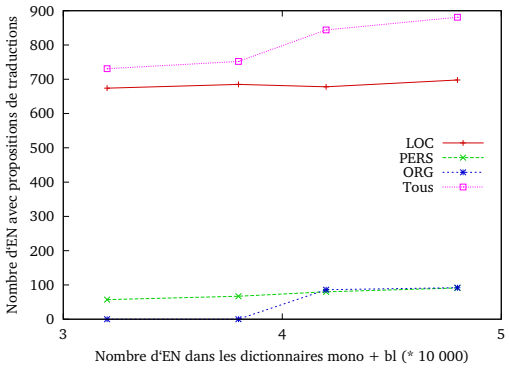

Figure 5. *Nombre d'entités nommées du corpus de test AFP trouvées dans les dictionnaires bilingues en fonction de l'évolution de la taille des dictionnaires*

et bilingues de taille réduite, qui, en particulier, ne contiennent aucun nom d'organisation, et correspondent aux points situés à gauche sur la figure 5. Ces dictionnaires sont ensuite enrichis : les dictionnaires monolingues par 1 143 lieux, 4 466 personnes et 726 organisations supplémentaires ; les dictionnaires bilingues par 1 586 paires de lieux, 3 129 paires de personnes et 726 paires d'organisations, correspondant aux deux points intermédiaires. Les points les plus à droite correspondent à l'ajout simultané des deux dictionnaires.

Les dictionnaires comme corpus : nous avons enfin effectué une dernière expérience afin de savoir si l'intégration des dictionnaires en tant que corpus pour l'entraînement des modèles de traduction peut être intéressante. Les évaluations ont été menées soit en concaténant les données des dictionnaires avec les données du système AFP et en construisant un seul modèle de traduction (1 table), soit en utilisant un deuxième modèle appris à partir des seules données contenues dans les diction-

naires (2 tables). Dans cette dernière situation, le mode *either* dans Moses et l'option *decoding-graph-backoff* ont été utilisés afin de privilégier le modèle de traduction AFP (pour lequel les EN sont présentées en contexte) par rapport au modèle de traduction construit à partir de données extraites des dictionnaires. Les scores de traduction obtenus sont donnés dans le tableau 8.

	#mots unk	BLEU	METEOR
1 table de traduction	272	34,97	52,99
2 tables de traduction	277	34,28	52,14

Tableau 8. *Scores BLEU et METEOR en traduction arabe-français sur le corpus de test AFP en utilisant les dictionnaires comme corpus d'entraînement des modèles de traduction*

L'approche par intégration des dictionnaires comme corpus s'avère donc ici paradoxalement meilleure qu'une approche par prétraduction (en comparaison des résultats du tableau 7) et améliore même de 0,3 point BLEU les performances par rapport au système de base (34,97 par rapport à 34,62 pour *default*). Intégrer les dictionnaires directement au sein des données parallèles permet au système d'optimiser leurs scores relatifs par rapport aux autres entrées du modèle de traduction, tout en assurant, comme précédemment, une meilleure couverture pour les EN. Des observations similaires sont rapportées dans (Kashani *et al.*, 2007).

6.2.2. *Tests sur le corpus Arcade II*

Afin d'évaluer au mieux l'impact de la détection des EN sur la traduction, nous avons utilisé le corpus Arcade II qui est annoté en EN en source et en cible. Environ 60 % des phrases du corpus Arcade II contiennent des EN, distribuées comme représenté sur la figure 6. Le tableau 9 présente une comparaison du score BLEU des différentes variantes de nos systèmes de traduction sur ce corpus. La détection des EN est effectuée avec le modèle adapté ANER + AFP. Une évaluation a été également réalisée sur le corpus de test en proposant les traductions correctes des EN présentes dans les références afin de mesurer le maximum d'amélioration atteignable (Oracle). Un tel test nécessite d'avoir un corpus annoté en EN.

Comme précédemment, le tableau 9 ne montre pas d'amélioration du score BLEU lorsque l'on utilise les options *inclusive* et *exclusive*, bien que le pourcentage des mots inconnus (unk) soit réduit de 1,13 %. L'option *exclusive* n'améliore pas les résultats de traduction *default*, même dans la situation idéale où toutes les EN proposées sont correctes (configuration Oracle).

Pour ce corpus, l'approche qui consiste à utiliser les dictionnaires comme données parallèles s'avère moins positive. Les scores BLEU et METEOR obtenus sont reportés dans le tableau 10. On note qu'en utilisant une seule table de traduction, le nombre de mots inconnus est réduit.

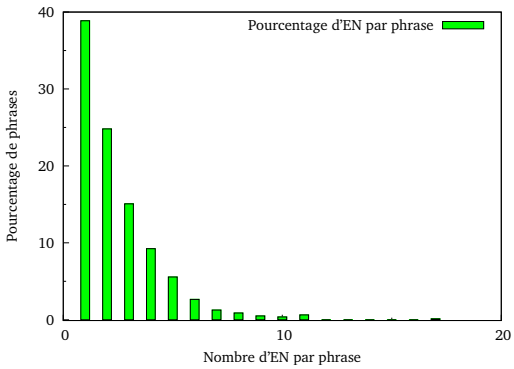

Figure 6. *Pourcentage de phrases du corpus Arcade selon le nombre d'entités nommées par phrase*

	#mots unk	BLEU	METEOR
Default	2 634	20,87	40,17
Exclusive	2 604	20,52	39,81
Inclusive	2 604	20,83	40,22
Oracle exclusive	2 493	20,60	39,24
Oracle inclusive	2 493	21,17	40,13

Tableau 9. *Scores BLEU et METEOR en traduction arabe-français sur le corpus de test Arcade II sans* default *et avec* exclusive *et* inclusive *proposition de prétraductions, en comparaison avec l'Oracle (*Oracle exclusive *et* Oracle inclusive*)*

6.2.3. *Analyse des résultats*

Cette section est consacrée à une analyse des résultats obtenus selon trois niveaux : (i) comparaison des deux types de corpus de test, (ii) couverture et pourcentage des EN existantes dans les dictionnaires et (iii) statistiques selon les types d'EN. Ces deux derniers niveaux sont réalisés grâce au corpus Arcade II permettant d'analyser plus en détail les EN.

Test AFP *vs* **Arcade II :** une observation troublante est la grande différence de performances de nos systèmes entre les corpus de test AFP et Arcade. On observe par exemple une chute de 14 points BLEU pour le système de base *default* sur le corpus Arcade (tableau 9, BLEU 20,87) en comparaison des résultats obtenus pour le même système sur le corpus AFP (tableau 7, BLEU 34,62). Nous avons donc procédé à une étude comparative de ces deux corpus qui montre que le style des données Arcade – issues du journal *Le Monde diplomatique* – est sensiblement différent de celui des dépêches de l'AFP. Ainsi, par exemple, 43,2 % des phrases du corpus de test AFP dé-

	#mots unk	BLEU	METEOR
1 table	2 536	20,21	40,18
2 tables	2 552	20,27	39,19

Tableau 10. *Scores BLEU et METEOR en traduction arabe-français sur le corpus de test Arcade II en utilisant les dictionnaires comme corpus d'entraînement des modèles de traduction*

butent par un syntagme verbal (position canonique du verbe en arabe), alors que c'est le cas pour seulement 9,6 % du corpus de test Arcade. Ce phénomène est illustré par la figure 7 qui présente deux phrases en arabe représentatives respectivement du corpus Arcade II et du corpus AFP avec leurs annotations morphosyntaxiques – extraites automatiquement avec SAPA – et leurs traductions en français. La phrase extraite du corpus Arcade débute par un syntagme nominal alors que celle extraite du corpus AFP commence par une conjonction suivie d'un verbe – ces deux unités étant agglutinées dans la forme brute (avant segmentation).

Arcade :

وبعد قليل من عودة السيد كيلي الي واشنطن اعلن مسوؤل اميركي امام الصحفيين ان اتفاق العام ١٩٩٤ الاساسي حول وقف العمل في مفاعل يونغ اصبح كانه لم يكن.

POS : conj prep noun prep noun noun noun_prop prep noun_prop verb noun adj prep noun conj_sub noun adj noun_num adj prep noun noun prep noun noun_prop noun_prop verb prep conj_sub part_neg verb punc

Réf. : peu de temps après le retour de M. Kelly à Washington , un responsable américain déclarait à des journalistes que l'accord-cadre de 1994 sur le gel du réacteur de Yongbyon était nul et non avenu .

AFP :

وقال برنار فاليرو المتحدث باسم الخارجية الفرنسية باشرنا عملية متواسلة لضمان الامن حفاضا علي سرية الوثائق.

POS : conj verb noun_prop noun_prop adj prep noun noun adj punc verb noun adj prep noun noun noun prep noun noun punc punc

Réf. : « afin de préserver la confidentialité des documents , nous sommes engagés dans un processus permanent de sécurisation » , déclare Bernard Valero , le porte-parole du ministère .

Figure 7. *Comparaison de la structure de deux phrases issues de deux corpus différents : Arcade II et AFP. La figure présente les phrases en arabe, les séquences d'étiquettes morphosyntaxiques correspondantes, et la traduction en français.*

Le décalage temporel entre les deux corpus peut également expliquer cette différence de performances. Nos systèmes de traduction ont été en effet entraînés sur des données issues de dépêches datant de 2009 à 2012. Si les données de test AFP ont été

collectées dans la même période (décembre 2010), celles du corpus Arcade l'ont été entre 2001 et 2003.

Couverture des EN par les dictionnaires bilingues : l'analyse des résultats sur le corpus Arcade, seul corpus annoté à la fois en EN en source et en cible, montre que la proportion des EN – présentes dans la partie source en arabe – pour lesquelles les traductions ont été retrouvées dans les dictionnaires bilingues est faible : en moyenne 16,4 % (24,2 % des lieux, 20 % des noms de personnes et seulement 5 % des organisations). Cette faible couverture s'explique par deux raisons : d'une part, le corpus Arcade contient des noms d'entités très spécifiques, comme par exemple *Barrio Alicia Pietri de Caldera* (un nom de quartier baptisé en hommage à la femme du précédent président du Venezuela), ou encore *Organisation Libérale* (en Tunisie) et *Tribunal pénal international pour l'ex-Yougoslavie de La Haye*. D'autre part, un certain nombre de noms de lieux qui ne sont pas reconnus existent dans nos dictionnaires mais sous une autre forme : c'est le cas de *territoires palestiniens* qui existe en tant que *Palestine* ou encore de *zones kurdes* qui existe en tant que *Kurdistan*, voire qui ont une orthographe différente de celle des dictionnaires : c'est le cas par exemple pour *l'Angleterre*, انكلترا au lieu de انجلترا ou pour *les Philippines*, الفيليبين au lieu de الفلبين.

Enfin, parmi les propositions de traductions, seulement 84,2 % des EN présentes dans la partie source du corpus Arcade existent dans la référence en langue cible (se répartissant comme suit : 75,6 % des lieux – pour lesquels il y a des propositions de traductions –, 81 % des noms de personnes et 96,1 % des organisations). Par conséquent, avec les dictionnaires existants, la traduction ne pourrait être potentiellement améliorée que pour au plus 13,8 % (16,4 % x 84,2 %) de ces ENs.

Statistiques selon les types d'EN : le taux d'EN traduites correctement a été calculé pour chaque type en comparant les EN traduites aux EN dans la référence (français) sur le corpus Arcade. Les résultats sont dans le tableau 11.

	LOC	PERS	ORG
Default	70,4 %	51,2 %	44,6 %
Inclusive	70,7 %	52,8 %	44,6 %
Oracle inclusive	70,9 %	53,0 %	45,2 %
Dictionnaires comme corpus			
1 table	72,0 %	57,3 %	44,4 %
2 tables	66,7 %	54,9 %	45,8 %

Tableau 11. *Pourcentage des EN – dans le corpus Arcade II – traduites correctement en utilisant notre approche de propositions de traductions en comparaison avec l'Oracle et avec l'utilisation des dictionnaires comme corpus pour l'entraînement des modèles de traduction*

D'après le tableau 11, on note qu'en utilisant l'option *inclusive*, la traduction de tous les types d'EN a été améliorée (jusqu'à 1,52 %). Ce résultat est à comparer au score qui serait atteint en traduisant correctement toutes les EN *détectées et qui*

existent dans les dictionnaires (ligne *Oracle inclusive*) : les prétraductions proposées par le dictionnaire sont donc presque toujours justes.

L'utilisation des dictionnaires comme corpus pour l'entraînement des modèles de traduction est encore meilleure et réduit jusqu'à 2,5 % le taux d'EN inconnues lors de la traduction. Ces résultats quantitatifs montrent que l'utilisation de dictionnaires améliore effectivement la traduction des EN, même si cela ne se traduit pas toujours par une augmentation des métriques de traduction automatique.

7. Conclusion

Dans ces travaux, nous nous sommes intéressés à la réduction du taux d'erreur des mots inconnus et à l'amélioration de la traduction des EN. Les EN représentent en effet une large proportion des mots inconnus, et leur traduction est particulièrement informative. Nous nous sommes donc focalisés sur le traitement automatique des EN afin de réduire les erreurs de traduction de ces entités. Nous avons développé un outil de détection des EN en arabe (NERAr) construit par des méthodes d'apprentissage supervisé. Ce système, qui embarque des centaines de milliers de descripteurs, obtient des performances comparables aux meilleurs systèmes de l'état de l'art. Nous avons ensuite adapté ce système par autoapprentissage conduisant à une légère amélioration des performances.

Notre deuxième contribution consiste à proposer une méthode de traduction des EN en arabe à l'aide de dictionnaires bilingues. Nous avons montré que la traduction préalable des EN améliore légèrement la qualité de la traduction automatique (telle que mesurée par BLEU et METEOR). Le nombre de mots inconnus a été réduit de 1,65 %. Une évaluation manuelle ainsi qu'une évaluation plus détaillée des EN obtenues en sortie de traduction montrent que la traduction automatique des EN a été améliorée pour les trois types d'EN : personne, lieu et organisation ; une amélioration que les métriques automatiques peinent à détecter. Il est intéressant de noter que dans notre architecture les erreurs de NERAr ne se répercutent pas sur la traduction : il est exceptionnel qu'une EN présente dans les dictionnaires bilingues ne soit pas détectée ; à l'inverse les mots incorrectement étiquetés comme EN ne se trouvent jamais dans les dictionnaires bilingues et ne sont donc pas prétraduits.

Une analyse des résultats à trois niveaux a été réalisée : tout d'abord une comparaison des styles des deux corpus de test, ensuite des statistiques sur la couverture des EN par les dictionnaires bilingues (afin de voir si les données ajoutées au décodeur sont intéressantes), et finalement une étude sur l'effet de la traduction des EN. Les résultats obtenus sur le corpus Arcade II ainsi que l'analyse détaillée des résultats montrent que notre approche améliore globalement la traduction des EN, même si cette amélioration ne se traduit pas toujours par une meilleure évaluation automatique.

Comme perspective, nous envisageons d'améliorer notre méthode de traduction des EN en rajoutant des translittérations pour les noms propres détectés par NERAr mais qui n'existent pas dans les dictionnaires, de façon à disposer d'une solution de

repli qui assure qu'aucune EN n'est copiée verbatim dans la sortie. Il semble également important de mettre en œuvre des techniques plus actives pour augmenter la couverture des dictionnaires d'EN, par exemple en exploitant des corpus comparables. Une troisième voie d'amélioration de notre système consistera à mettre en œuvre une stratégie plus effective pour garantir que non seulement les EN sont mieux traduites, mais que cette amélioration touche aussi positivement la traduction des mots voisins, idéalement la traduction de toute la phrase[24]. Ceci passera, comme nos dernières expériences le suggèrent, par un meilleur couplage entre la détection des EN et l'apprentissage du modèle de traduction. À moyen terme, nous envisageons d'exploiter des méthodes d'alignement au niveau des mots pour projeter des annotations depuis le français, langue dans laquelle les entités sont plus faciles à détecter, vers l'arabe, à l'instar du travail de Zhang (2012) pour le chinois : ceci nous permettrait d'avoir des corpus d'adaptation au moins partiellement annotés en EN. Une autre perspective consiste à réordonner une liste des meilleurs candidats (nbests) afin d'améliorer la traduction, en utilisant le fait que les bons candidats en langue cible doivent contenir le même nombre d'EN qu'il en existe en langue source. Cette approche demandera toutefois de mettre en œuvre des outils d'étiquetage dans les deux langues.

Remerciements

Les auteurs remercient Hacene Cherfi et Jérôme Mainka pour les données Geonames et JRC, Sophie Rosset pour de nombreuses discussions intéressantes, ainsi que Thomas Lavergne pour la mise en oeuvre de Wapiti. Merci également à Dominique Ferrandini, Leila Zighem et Sylvie Guillemin-Lanne pour le corpus de référence Gold AFP.

8. Bibliographie

Abdul Hamid A., Darwish K., « Simplified Feature Set for Arabic Named Entity Recognition », *Proc. of the 2010 Named Entities Workshop*, Uppsala, p. 110-115, July, 2010.

Abdul Rauf S., Efficient corpus selection for Statistical Machine Translation, PhD thesis, Université du Maine, Le Mans, 2012.

Al-Jumaily H. T., Martínez P., Martínez-Fernández J. L., der Goot E. V., « A real time Named Entity Recognition system for Arabic text mining », *Language Resources and Evaluation*, vol. 46, n° 4, p. 543-563, 2012.

Al-Onaizan Y., Knight K., « Machine transliteration of names in Arabic text », *Proc. of the ACL-02 workshop on Computational approaches to semitic languages*, p. 1-13, 2002a.

Al-Onaizan Y., Knight K., « Translating named entities using monolingual and bilingual resources », *Proc. of the 40th Annual Meeting on ACL*, ACL '02, Philadelphia, PA, USA, p. 400-408, 2002b.

24. Nous remercions un des relecteurs pour ses remarques sur l'interaction entre ces deux composants de notre système.

Allauzen A., Modélisation linguistique pour l'indexation automatique de documents audiovisuels, PhD thesis, Université Paris Sud, Orsay, 2003.

Banerjee S., Lavie A., « METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments », *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, Ann Arbor, Michigan, p. 65-72, 2005.

Béchet F., Sagot B., Stern R., « Coopération de méthodes statistiques et symboliques pour l'adaptation non supervisée d'un système d'étiquetage en entités nommées », *actes de la conférence TALN*, Montpellier, France, 2011.

Bellagarda J. R., « An overview of statistical language model adaptation », *Proc. of the ISCA Workshop on Adaptation Methods for Speech Recognition*, Sophia Antipolis, p. 165-174, 2001.

Benajiba Y., Diab M., Rosso P., « Arabic named entity recognition using optimized feature sets », *Proc. of EMNLP*, EMNLP, Stroudsburg, PA, USA, p. 284-293, 2008.

Benajiba Y., Rosso P., « Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information », *Proceedings of Workshop on Natural Language-Independant Engineering*, IJCAI, 2007.

Benajiba Y., Rosso P., « Arabic named entity recognition using Conditional Random Fields », *Proceedings of the Conference on Language Resources and Evaluation*, 2008.

Benajiba Y., Rosso P., Benedí J.-M., « ANERsys : An Arabic Named Entity Recognition System Based on Maximum Entropy », *CICLing*, p. 143-153, 2007.

Blitzer J., Domain adaptation of natural language processing systems, PhD thesis, University Philadelphia, PA, USA, 2008.

Buckwalter T., *Buckwalter Arabic Morphological Analyzer Version 1.0*, Catalog No. LDC2002L49, Linguistic Data Consortium, University of Pennsylvania, 2002.

Che W., Wang M., Manning C. D., Liu T., « Named Entity Recognition with Bilingual Constraints », *NAACL-HLT*, 2013.

Chiao Y.-c., Kraif O., Laurent D., Minh T., Nguyen H., Semmar N., Stuck F., Véronis J., Zaghouani W., « Evaluation of multilingual text alignment systems : the ARCADE II project », *In Proc. of LREC*, 2006.

Chinchor N., Robinson P., « MUC-7 Named Entity Task Definition », *Proceedings of the seventh Message Understanding Conference (MUC 7)*, 1997.

Daumé III H., « Frustratingly Easy Domain Adaptation », *Proc. of the 45th Annual Meeting of the ACL*, Prague, Czech Republic, June, 2007.

Daumé III H., Deoskar T., McClosky D., Plank B., Tiedemann J. (eds), *Proc. of the 2010 Workshop on Domain Adaptation for NLP*, Uppsala, Sweden, July, 2010.

Daumé III H., Marcu D., « Domain Adaptation for Statistical Classifiers », *Journal of Artificial Intelligence Research*, vol. 26, p. 101-126, 2006.

Diab M., « Second Generation Tools (AMIRA 2.0) : Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking », *in* K. Choukri, B. Maegaard (eds), *Proc. of the Second International Conference on Arabic Language Resources and Tools*, The MEDAR Consortium, Cairo, Egypt, April, 2009.

Fehri H., Haddar K., Hamadou A. B., « Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model », *in* M. Constant, A. Maletti, A. Savary (eds), *FSMNLP, 9th International Workshop*, ACL, Bois, France, p. 134-142, 2011.

Fung P., McKeown K., « Finding Terminology Translations from Non-parallel Corpora », *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong-Kong, p. 192-202, 1997.

Gahbiche-Braham S., Bonneau-Maynard H., Yvon F., « Two Ways to Use a Noisy Parallel News Corpus for Improving Statistical Machine Translation », *Proc. of Workshop on Building and Using Comparable Corpora*, Portland, OR, p. 44-51, 2011.

Gahbiche-Braham S., Maynard H., Lavergne T., Yvon F., « Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier », *in* N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds), *Proc. of the 8th LREC*, ELRA, Istanbul, Turkey, 2012.

Gao Q., Vogel S., « Parallel implementations of word alignment tool », *In* Software Engineering, Testing, and Quality Assurance for Natural Language Processing, p. 49-57, 2008.

Habash N., « Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation », *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, p. 57-60, 2008.

Habash N., *Arabic Natural Language Processing*, Synthesis Lectures on Human Language Technologies, organ & Claypool Publishers, 2011.

Habash N., Rambow O., Roth R., « MADA+TOKAN : A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization », *in* K. Choukri, B. Maegaard (eds), *Proc. of the 2nd International Conference on Arabic Language Resources and Tools*, The MEDAR Consortium, Cairo, Egypt, April, 2009.

Hal D., Jagarlamudi J., « Domain adaptation for machine translation by mining unseen words », *Proceedings of the 49th Annual Meeting of the ACL :HLT : short papers - Volume 2*, HLT '11, ACL, Stroudsburg, PA, USA, p. 407-412, 2011.

Hálek O., Rosa R., Tamchyna A., Bojar O., « Named Entities from Wikipedia for Machine Translation », *in* M. Lopatková (ed.), *ITAT 2011 Information Technologies – Applications and Theory*, vol. 788, p. 23-30, September, 2011.

Hassan A., Fahmy H., Hassan H., « Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora », *Proc. of the Conference on Recent Advances in Natural Language Processing (RANLP '07)*, Borovets, Bulgaria, 2007.

Hassan H., Sorensen J., « An integrated approach for Arabic-English named entity translation », *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Semitic '05, ACL, Stroudsburg, PA, USA, p. 87-93, 2005.

Heintz I., « Arabic Language Modeling with Finite State Transducers », *Proc. of the ACL-08 : HLT Student Research Workshop*, ACL, Columbus, Ohio, p. 37-42, June, 2008.

Hermjakob U., Knight K., Daumé III H., « Name Translation in Statistical Machine Translation - Learning When to Transliterate », *Proc. of ACL-08 : HLT*, Ohio, p. 389-397, June, 2008.

Jiang J., Zhai C., « Instance Weighting for Domain Adaptation in NLP », *Proc. of the 45th Annual Meeting of the ACL*, Prague, Czech Republic, p. 264-271, June, 2007.

Jiang L., Zhou M., Chien L.-F., Niu C., « Named Entity Translation with Web Mining and Transliteration », *Proc. of the 20th International Joint Conference on Artificial Intelligence*, p. 1629-1634, 2007.

Kashani M. M., Joanis E., Kuhn R., Foster G., Popowich F., « Integration of an Arabic transliteration module into a statistical machine translation system », *Proc. of the 2nd Workshop on Statistical Machine Translation*, StatMT '07, ACL, Stroudsburg, PA, USA, p. 17-24, 2007.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E., « Moses : open source toolkit for statistical machine translation », *In* Proceedings of the 45th Annual Meeting of the ACL, p. 177-180, 2007.

Lafferty J., McCallum A., Pereira F., « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data », *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Morgan Kaufmann, San Francisco, CA, p. 282-289, 2001.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *Proceedings of the 48th Annual Meeting of the ACL*, p. 504-513, July, 2010.

Lavie A., Agarwal A., « METEOR : An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments », *Proc. of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic, p. 338-231, 2007.

Ling W., Calado P., Martins B., Trancoso I., Black A., Coheu L., « Named Entity Translation using Anchor Texts », *Proc. of the IWSLT*, San Francisco, USA, 2011.

Maamouri M., Bies A., Buckwalter T., Jin H., Mekki W., *Arabic Treebank : Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis)*, n° LDC2005T20, Linguistic Data Consortium, 2005a.

Maamouri M., Bies A., Buckwalter T., Jin H., Mekki W., *Arabic Treebank : Part 4 v 1.0 (MPG Annotation)*, n° LDC2005T30, Linguistic Data Consortium, 2005b.

Maloney J., Niv M., « TAGARAB : a fast, accurate Arabic name recognizer using high-precision morphological analysis », *Proc. of the Workshop on Computational Approaches to Semitic Languages*, Semitic '98, Stroudsburg, PA, USA, p. 8-15, 1998.

Mansour S., « MorphTagger : HMM-Based Arabic Segmentation for Statistical Machine Translation », *IWSLT*, Paris, France, p. 321-327, December, 2010.

Mihalcea R., « Co-training and Self-training for Word Sense Disambiguation », *in* H. T. Ng, E. Riloff (eds), *HLT-NAACL Workshop : CoNLL-2004*, Boston, p. 33-40, May, 2004.

Moore R. C., « Learning translations of named-entity phrases from parallel corpora », *Proceedings of the tenth conference on European chapter of the ACL - Volume 1*, EACL '03, ACL, Stroudsburg, PA, USA, p. 259-266, 2003.

Papineni K., Roukos S., Ward T., Zhu W.-J., « BLEU : a method for automatic evaluation of machine translation », *Proc. of the 40th Annual Meeting on ACL*, Stroudsburg, PA, USA, p. 311-318, 2002.

Rapp R., « Identifying word translations in non-parallel texts », *Proceedings of the 33rd annual meeting on ACL*, ACL '95, ACL, Stroudsburg, PA, USA, p. 320-322, 1995.

Samy D., Moreno A., Ma Guirao J., « A Proposal For An Arabic Named Entity Tagger Leveraging a Parallel Corpus », *Proceedings of RANLP'05*, 2005.

Santanu P., Kumar Naskar S., Pecina P., Bandyopadhyay S., Way A., « Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation », *Proc. of the COLING 2010 Workshop on Multiword Expressions : from Theory to Applications (MWE 2010)*, Beijing, China, p. 46-54, August, 2010.

Shaalan K., Raza H., « NERA : Named Entity Recognition for Arabic », *Journal of the American Society for Information Science and Technology*, vol. 60, n° 9, p. 1652-1663, 2009.

Sokolovska N., Cappé O., Yvon F., « Sélection de caractéristiques pour les champs aléatoires conditionnels par pénalisation L_1 », *TAL*, vol. 50, n° 3, p. 139-171, 2009.

Zaghouani W., Pouliquen B., Ebrahim M., Steinberger R., « Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic », Proc. of the Seventh conference on International Language Resources and Evaluation (LREC'10), p. 563-567, 2010.

Zhang M., Li H., Kumaran A., Liu M., « Report of NEWS2011 Machine Transliteration Shared Task », *Proc. of the 2011 Named Entities Workshop*, 2011.

Zhang Y., The Application of Source Language Information in Statistical Machine Translation, PhD thesis, RWTH Aachen University, 2012.

Zitouni I., Sorensen J., Luo X., Florian R., « The Impact of Morphological Stemming on Arabic Mention Detection and Coreference Resolution », *Proc. of Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, Michigan, p. 63-70, June, 2005.