Adaptation thématique pour la traduction automatique de dépêches de presse

Souhir Gahbiche-Braham[1, 2] Hélène Bonneau-Maynard[1, 2] François Yvon[1]
(1) LIMSI-CNRS, B.P. 133, F-91403 Orsay Cedex, France
(2) Université Paris Sud
souhir@limsi.fr, hbm@limsi.fr, yvon@limsi.fr

Résumé. L'utilisation de méthodes statistiques en traduction automatique (TA) implique l'exploitation de gros corpus parallèles représentatifs de la tâche de traduction visée. La relative rareté de ces ressources fait que la question de l'adaptation au domaine est une problématique centrale en TA. Dans cet article, une étude portant sur l'adaptation thématique des données journalistiques issues d'une même source est proposée. Dans notre approche, chaque phrase d'un document est traduite avec le système de traduction approprié (c.-à-d. spécifique au thème dominant dans la phrase). Deux scénarios de traduction sont étudiés : (a) une classification *manuelle*, reposant sur la codification IPTC ; (b) une classification *automatique*. Nos expériences montrent que le scénario (b) conduit à des meilleures performances (à l'aune des métriques automatiques), que le scénario (a). L'approche la meilleure pour la métrique BLEU semble toutefois consister à ne pas réaliser d'adaptation ; on observe toutefois qu'adapter permet de lever certaines ambiguïtés sémantiques.

Abstract. Statistical approaches used in machine translation (MT) require the availability of large parallel corpora for the task at hand. The relative scarcity of thes resources makes domain adaptation a central issue in MT. In this paper, a study of thematic adaptation for News texts is presented. All data are produced by the same source : News articles. In our approach, each sentence is translated with the appropriate translation system (specific to the dominant theme for the sentence). Two machine translation scenarios are considered : (a) a *manual* classification, based on IPTC codification ; (b) an *automatic* classification. Our experiments show that scenario (b) leads to better performance (in terms of automatic metrics) than scenario (a) . The best approach for the BLEU metric however seems to dispense with adaptation alltogether. Nonetheless, we observe that domain adaptation sometimes resolves some semantic ambiguities .

Mots-clés : adaptation thématique, classification automatique, traduction automatique.

Keywords: domain adaptation, automatic classification, machine translation.

1 Introduction

En traitement automatique des langues, l'adaptation au domaine[1] est une question souvent abordée. Pour l'application de traduction automatique, adapter ou spécialiser des modèles de traduction à un genre ou à un thème peut permettre de résoudre certaines ambiguïtés qui ne pourraient être levées par des modèles généraux. Ainsi, si le domaine considéré est celui de la réservation de billets d'avions, on peut s'attendre à ce que la spécialisation de modèles de traduction à ce domaine permette de désambiguïser du mot anglais *book*, qui peut être traduit en français par le substantif *livre* ou par le verbe *réserver*. Conjointement, l'adaptation du modèle de langue permet de distinguer certaines traductions en utilisant le contexte de la phrase en langue cible. Sennrich (2012b) explore les différences conceptuelles entre l'adaptation des modèles de traduction et celle du modèle de langue ainsi que leurs effets sur la performance de la traduction.

Dans l'étude présentée ici, la question de l'adaptation est abordée à un niveau plus fin que celui qui est classiquement considéré : alors qu'on se pose le plus souvent la question de combiner des données *du domaine* avec des données *hors domaine*, on envisage ici une adaptation **thématique** au sein d'un même registre et genre, celui des dépêches journalistiques. Pour ces documents, une classification manuelle en thèmes (ou catégories) est attribuée par les journalistes lors de la rédaction des dépêches, que nous cherchons donc à utiliser. Notre corpus est en effet constitué d'un ensemble de

[1]Le terme « domaine » doit être compris dans une acception assez vague : le besoin d'adapter se révèle dès lors que les données d'apprentissage diffèrent des données de l'application, que ces différences soient des différences de genre, de registre, de modalité, ou encore de thème, qui est la situation considérée ici.

dépêches produites par l'AFP[2]. Ces documents sont classifiés selon les 17 catégories principales du standard IPTC[3] (voir section 3). L'utilisation de cette classification est une des particularités de notre étude. La deuxième particularité, qui est inhérente au fait que les données proviennent d'une même source, est qu'elles sont beaucoup plus homogènes que dans les cas standard d'adaptation au domaine. Par ailleurs, plusieurs catégories peuvent être affectées à une même dépêche. De plus, les observations sur le corpus montrent que les frontières entre les catégories sont relativement floues.

Dans ce contexte, différentes méthodes d'adaptation thématique sont explorées, qui impliquent aussi bien l'adaptation des modèles de traduction que l'adaptation des modèles de langue. Les premières expériences sont menées en utilisant la classification *manuelle* en catégories qui accompagne les dépêches ; dans ce, cas toutes les phrases d'une même dépêche sont traduites avec les mêmes modèles. Nous proposons ensuite de considérer une classification *automatique* des phrases, permettant de traduire chaque phrase par le modèle qui lui est le plus approprié.

Cet article est organisé comme suit : la section 2 présente un état de l'art des approches d'adaptation au domaine en traduction automatique. Les données utilisées pour cette étude sont décrites dans la section 3. La section 4 décrit notre approche et la section 5 expose les résultats obtenus. Les principales conclusions sont enfin présentées à la section 6.

2 État de l'art

Les premiers travaux sur l'adaptation des modèles de langue ont été publiés durant les années 90, particulièrement dans le domaine de la reconnaissance de la parole (De Mori & Federico, 1999). La relative rareté des ressources parallèles qui sont nécessaires à l'apprentissage de systèmes de traduction a progressivement fait émerger cette problématique en TA, et a donné lieu dans les années récentes à une littérature abondante que nous survolons rapidement ici.

Langlais (2002) présente les premiers travaux sur l'adaptation du domaine en traduction automatique. Il implémente une stratégie qui consiste à compléter le modèle de traduction avec lexiques adaptés. De nombreuses stratégies d'adaptation ont ensuite été proposées, pour l'essentiel fondées sur l'interpolation de modèles du domaine et de modèles hors-domaine, comme par exemple dans (Koehn & Schroeder, 2007), où les meilleures performances sont obtenues par une interpolation linéaire de modèles de langue et une interpolation log-linéaire de modèles de traduction.

Des approches d'adaptation fondées sur les modèles de mélange ont été également été proposées par Foster & Kuhn (2007) et par Sennrich (2012a). Ces auteurs utilisent des modèles de langue et des modèles de traduction adaptés combinés par interpolation linéaire et log-linéaire. Sennrich (2012b) adapte les modèles de traduction en minimisant les scores de perplexité et pour optimiser les coefficients d'interpolation. Si l'adaptation thématique améliore souvent les performances, il apparaît également que les données hors-domaine peuvent dégrader les performances de traduction en introduisant des ambiguïtés lexicales. En particulier, Haddow & Koehn (2012) montrent que l'ajout de données hors domaine à un corpus d'apprentissage peut améliorer la traduction des mots rares (les moins fréquents) mais en revanche dégrade la qualité de la traduction des mots les plus fréquents.

Une stratégie alternative d'adaptation consiste à pondérer différentiellement les données du domaine et hors-domaine ; cette piste est explorée notamment par Foster *et al.* (2010), Shah *et al.* (2010) ou encore Niehues & Waibel (2010) : des poids sont assignés aux phrases et aux segments de phrases avant la création des modèles de traduction. Il est par exemple possible de nuancer l'importance des phrases hors-domaine sur la base d'un degré de similarité avec des phrases du domaine.

Selon les applications, il est possible de considérer des distinctions plus fines que *in domain* versus *out-of-domain*, Dans ce contexte, Yamamoto & Sumita (2008) détectent les domaines de chaque phrase à traduire, qui est ensuite traduite en utilisant les modèles spécifiques au domaine détecté. Nakov (2008) utilise une interpolation log-linéaire et combine plusieurs modèles de langue de différents domaines. De nouveaux traits sont ajoutés, un pour chaque modèle de traduction interpolé. Ce trait indique pour chaque paire de phrases, si elle provient du modèle de traduction en question.

Zhao *et al.* (2004) explorent des techniques pour l'adaptation non-supervisée des modèles de langue. Ces modèles sont construits à partir de données monolingues extraites en se basant sur des scores de similarité et des traits (*features*) sémantiques. Ces modèles sont interpolés avec un modèle de langue général contenant l'ensemble des données et permettent d'obtenir une amélioration de la qualité de traduction. Dans des travaux plus récents, Sennrich *et al.* (2013) proposent une approche d'adaptation non supervisée dans laquelle plusieurs modèles de langue sont interpolés log-linéairement lors de

[2]Agence France Presse, `http://www.afp.fr`.

[3]L'IPTC – *International Press Telecommunications Council* – est un consortium qui rassemble les grandes agences de presse mondiales. L'IPTC fournit des schémas de classification, de normalisation et de codage des métadonnées, voir `http://www.iptc.org`.

la traduction. Des scores sont calculés lors du décodage pour choisir la meilleure hypothèse de traduction pour chaque phrase source.

Dans la plupart des travaux existants, les données adaptées proviennent de sources bien séparées. Un exemple extrême est donné par Sennrich (2012b), qui combine des données extraites à partir du journal *Alpine Club* (dédié à l'alpinisme) et des données issues du corpus *Europarl* (débats politiques). Il est clair que ces données dérivent de sources bien séparées. La situation est un peu moins marquée dans le travail de Banerjee (2012) sur la traduction de documents techniques, qui utilise deux catégories de données : ceux qui traitent de logiciels du domaine *availability* (récupération de données, sauvegarde) et ceux qui traitent du domaine *security* (vulnérabilité de logiciels malveillants, protection contre les attaques).

Eidelman *et al.* (2012) proposent une approche de classification thématique non-supervisée fondée sur les modèles d'allocation de Dirichlet latente (LDA). Le thème de chaque phrase est induit de manière non supervisée : des distributions thématiques sont utilisées pour calculer les probabilités de pondération lexicales du thème dépendant et les intégrer dans le modèle de traduction comme étant des traits (*features*). Les données utilisées proviennent du corpus FBIS[4] (audio), et du corpus NIST pour la traduction du chinois vers l'anglais.

Dans cet article, nous proposons plusieurs approches pour adapter les modèles de traduction et les modèles de langue. Nous nous intéressons à deux approches de classification (*manuelle* et *automatique*), qui servent à faire en sorte que chaque phrase soit traduite avec le système approprié construit à partir de modèles adaptés. Notre approche est semblable à celle de Sennrich (2012a) et de celle de de Yamamoto & Sumita (2008). La différence est que dans notre cas, les données proviennent de la même source et que l'on dispose d'une catégorisation manuelle des documents.

3 Description des données

Les données utilisées dans cette étude sont constituées d'un ensemble de dépêches journalistiques en arabe et en français produites par l'AFP entre décembre 2009 et juillet 2012. Chaque dépêche est catégorisée manuellement par les journalistes en utilisant les 17 catégories principales de la nomenclature IPTC (voir tableau 1). Une dépêche peut être affectée à plusieurs catégories. Il peut arriver également qu'une dépêche ne soit pas catégorisée par la suite d'un oubli.

Le corpus parallèle utilisé pour construire un système de traduction automatique statique est constitué de 265 000 paires de phrases extraites d'un corpus comparable avec la méthode décrite dans (Gahbiche-Braham *et al.*, 2011). Le tableau 1 donne la répartition des phrases selon ces 17 catégories. Pour chaque catégorie et pour chaque langue, la valeur indiquée est le pourcentage des phrases issues de dépêches étiquetées avec cette catégorie. Une phrase pouvant appartenir à une ou plusieurs catégories, on notera que la somme de ces pourcentages est supérieure à 100 %.

On peut observer que les valeurs diffèrent pour l'arabe et le français. En effet, bien que les dépêches en arabe soient souvent des (quasi)-traductions des dépêches en français, la catégorie thématique des dépêches est réattribuée manuellement par les journalistes après la traduction. Notons que la distribution des données est très inégale ; la catégorie dominante est la catégorie POL avec environ 60 % des phrases appartenant à cette catégorie. Nous limiterons notre étude aux trois catégories les plus fréquentes : *politique* (POL), *guerre* (WAR), et *finance* (FIN), les autres catégories étant présentes en trop petit nombre pour pouvoir mener des études. Le tableau présente également la répartition des phrases parallèles en catégories pour les corpus d'entraînement, de développement et de test initial. Le tableau 2 met en évidence le nombre de phrases communes pour chaque paire des catégories POL, WAR et FIN. On observe notamment que les catégories POL et WAR ont une intersection importante, représentant plus de 16 % de l'ensemble des phrases des deux catégories, et que la proportion des phrases de la catégorie WAR qui sont également étiquetées par la catégorie POL atteint presque la moitié (42,7%). Bien que non négligeables, les intersections sont moins importantes pour les deux autres paires de catégories, avec cependant encore 40,3 % des phrases de la catégorie FIN également affectées à la catégorie POL.

Les données collectées pendant le mois de novembre 2011 ont été isolées afin de construire un ensemble de test initial et un ensemble de développement. Ces ensembles sont constitués de respectivement 1 000 et 1 178 phrases. Trois corpus de test et de développement spécifiques à chaque catégorie ont été également constitués. Ces derniers sont également constitués de 1 000 (test) et 1 178 (développement) phrases chacun.

[4]FBIS : Foreign Broadcast Information Service.

				# phrases parallèles		
Catégorie	% AR	% FR	Nom de la catégorie	Entraînement	dev	test
ACE	1,1	2,7	*arts, culture and entertainment*	2 825	3	1
CLJ	9,2	12,1	*crime, law and justice*	24 330	25	32
DIS	5,3	6,8	*disaster and accident*	13 951	19	10
FIN	**14,8**	**15,3**	*economy, business and finance*	39 227	132	162
EDU	0,2	0,2	*education*	420	0	0
EVN	0,8	1,3	*environmental issue*	2 137	1	0
HTH	0,9	1,0	*health*	2 302	2	0
HUM	0,4	0,7	*human interest*	1 067	2	1
LAB	0,8	2,0	*labour*	1 986	2	3
LIF	0,03	0,03	*lifestyle and leisure*	93	0	0
POL	**58,9**	**62,3**	*politics*	156 352	663	561
REL	3,4	3,9	*religion and belief*	8 966	6	8
SCI	1,0	1,3	*science and technology*	2 746	2	2
SOI	2,0	2,6	*social issue*	5 367	2	3
SPO	1,3	1,3	*sport*	3 327	8	8
WAR	**38,3**	**42,9**	*unrest, conflicts and war*	101 655	310	207
WEA	0,9	0,1	*weather*	2 393	1	2
-	0,1	2,3	-	-	-	-

TAB. 1 – Distribution des catégories IPTC en arabe et en français et nombre de phrases pour les corpus d'entraînement, de développement et de test initial.

			Pourcentage des phrases communes dans chaque cat.	
Catégorie 1	Catégorie 2	Phrases en commun (%)		
POL	WAR	16,35	27,7 % POL	42,7 % WAR
POL	FIN	5,97	12,5 % POL	40,3 % FIN
WAR	FIN	1,36	3,6 % WAR	9,2 % FIN

TAB. 2 – Pourcentage des phrases communes entre les paires de catégories POL-WAR, POL-FIN et WAR-FIN pour le corpus d'entraînement, et pourcentage des phrases communes existant dans chaque catégorie pour ces trois paires.

4 Adaptation thématique

L'idée principale consiste à utiliser une classification thématique du corpus d'entraînement pour produire des systèmes de traduction spécifiques aux différentes thématiques. Il s'agit d'étudier l'impact de l'utilisation de modèles de traduction et de modèles de langue spécifiques sur la qualité de la traduction, en particulier dans le but de lever certaines ambiguïtés de mots polysémiques. Par exemple le verbe سجل (enregistre) en arabe peut avoir deux sens *marquer* ou *enregistrer* qui sont traduits différemment en français. La phrase سجل فرحان شكور هدف العراق (*Farhan Chakour* ***a marqué*** *le but de l'Iraq*) est extraite d'une dépêche à laquelle la catégorie SPO (sport) est attribuée, alors que la phrase سجل سنودن ومساعدته اسميهما على رحلة ايرفلوت (*Snowden et son assistante* ***ont enregistré*** *leurs noms dans le vol Airfloat*) est extraite d'une dépêche à laquelle la catégorie POL (politics) lui est affectée. La connaissance de la catégorie de la phrase peut permettre de choisir entre les deux sens du mot سجل.

Les difficultés de cette étude viennent du fait que nous nous intéressons au cas où les données spécialisées dérivent d'une même source et sont telles que les frontières entre thématiques sont relativement floues. Les catégories que nous traitons ne sont pas exclusives et l'intersection entre deux catégories peut être importante.

Deux scénarios d'adaptation sont alors proposés : une adaptation utilisant *une classification manuelle* des phrases et une adaptation utilisant *une classification automatique* des phrases. La section 4.1 présente l'approche de classification automatique utilisée dans cet article. Les méthodes d'adaptation sont décrites dans la section 4.2.

4.1 Classification manuelle versus classification automatique

Dans une première approche, chaque phrase du corpus d'entraînement est classifiée *manuellement* selon la catégorie de la dépêche originale dont elle est extraite (catégorisée par l'AFP). Seulement 46 % des phrases d'entraînement appartiennent à une seule catégorie. Les phrases appartenant à plusieurs catégories sont affectées à chacune de ces catégories.

L'approche par *classification automatique* remet en question l'hypothèse précédente. Plutôt que de projeter systématiquement la (ou les) catégorie(s) d'une dépêche sur les phrases la constituant, chaque phrase est catégorisée indépendamment des autres phrases du document. La figure 1 justifie cette seconde approche. Les deux phrases sont extraites d'une dépêche étiquetée par la catégorie FIN. Si la catégorie FIN est appropriée pour la première phrase de cet extrait, cela est moins clair pour la seconde, qui n'a en fait rien de spécifique à la catégorie finance.

> La jeune entreprise TimoCom est rapidement devenue une entreprise de taille moyenne, [...]
> Le texte du communiqué issu d'une traduction ne doit d'aucune manière être considéré comme officiel.

FIG. 1 – Un extrait d'une dépêche AFP affectée à la catégorie FIN.

Aux trois catégories traitées dans cette étude, une catégorie générique *Autre* est ajoutée, qui permet comme expliqué ci-dessous de catégoriser toutes les phrases. Un classifieur est donc entraîné pour affecter automatiquement une catégorie parmis (POL, WAR, FIN ou Autre) à chaque phrase.

Ce classifieur de phrases (en arabe) implémente une version simplifiée de l'algorithme Espérance-Maximisation (EM) pour le modèle de mélange de lois multinomiales (Rigouste *et al.*, 2007), initialisé avec les catégories IPTC. La principale singularité de l'approche consiste à estimer également un modèle « généraliste » en plus des modèles spécialisés. Lorsqu'une phrase est trop courte, ou bien qu'aucune des trois catégories n'obtient une vraisemblance meilleure que le modèle généraliste, alors la phrase n'est affectée à aucune des trois catégories et sera traduite (au test) par un modèle généraliste agrégeant toutes les données disponibles. Cette stratégie a pour effet de spécialiser les catégories automatiques, mais également de réduire les données utilisées pour apprendre les modèles spécialisées.

La probabilité *a posteriori* de chaque phrase est calculée par rapport aux quatre modèles disponibles. La catégorie assignée à chaque phrase est celle du modèle pour lequel elle présente la probabilité *a posteriori* la plus grande. De nouveaux modèles spécifiques sont alors appris sur la base de cette nouvelle annotation, puis utilisés pour reclassifier les données jusqu'à la convergence comme schématisé dans la figure 2.

La figure 3 donne la répartition du nombre de phrases pour chaque catégorie. Le nombre de phrases pour les catégories spécifiques est beaucoup plus réduit avec une classification automatique qu'avec la classification manuelle car dans le premier cas on impose la contrainte qu'une phrase ne peut être affectée qu'à une seule catégorie.

Après apprentissage, toutes les catégories des phrases d'entraînement sont recalculées. De nouveaux modèles spécifiques constitués sur la base de cette classification automatique sont alors construits : un effet induit important est que les phrases « génériques »sont retirées du corpus d'apprentissage des modèles thématiques initiaux (construits par classification manuelle), ce qui rend les modèles thématiques plus spécialisés. On note également qu'avec la nouvelle classification, les phrases extraites d'une même dépêche peuvent être affectées à des catégories différentes.

4.2 Méthodes d'adaptation

Pour chacune des méthodes de classification considérées, quatre approches d'adaptation sont explorées :

(a) Fusion des modèles de traduction : des modèles de traduction spécifiques sont entraînés séparément pour chaque

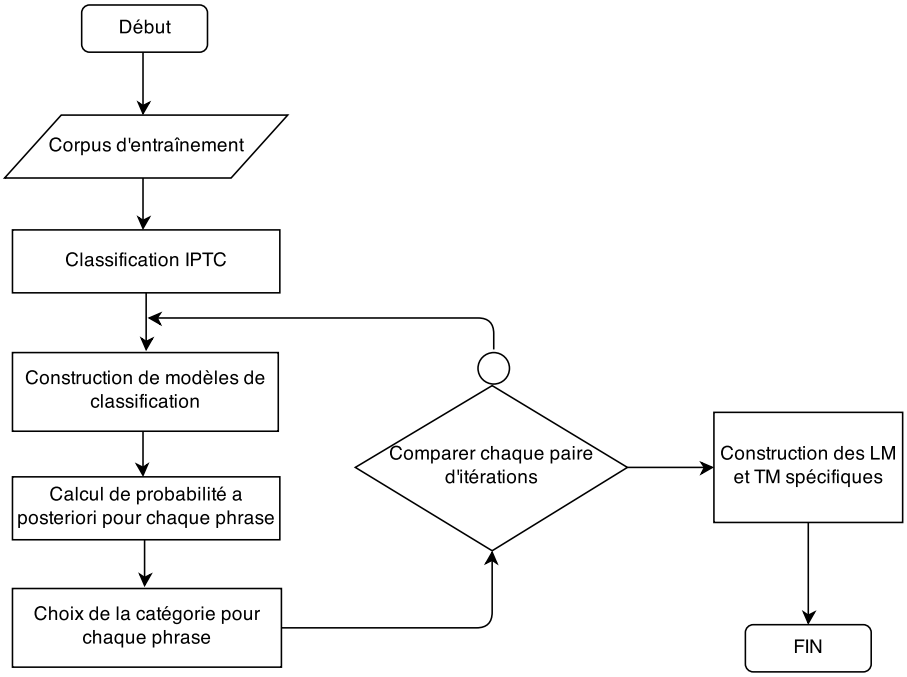

FIG. 2 – Processus itératif pour la construction du classifieur automatique.

catégorie (POL, WAR et FIN). Ils sont ensuite fusionnés en un seul modèle de traduction global. Chaque segment (chaque *phrase*) de la table de traduction globale (*Merge*), est ainsi nanti de 17 paramètres : les 4 paramètres classiques pour chaque modèle ($P(s|t)$, $P(t|s)$, $lex(s|t)$ et $lex(t|s)$) – qui sont soit recopiés des modèles spécifiques si le segment y est présent, soit affectés d'un score de probabilité faible – et un score qui représente la probabilité de distorsion (constante 2,718) ;

(b) Interpolation log-linéaire de modèles de traduction : consistant à employer deux modèles de traduction (spécifique et générique) en privilégiant le premier modèle de traduction spécifique par rapport au modèle de traduction générique (mode *either* dans Moses avec l'option *decoding-graph-backoff*). Leurs poids sont optimisés simultanément avec MERT (Och, 2003) afin d'optimiser les performances de traduction ;

(c) Interpolation log-linéaire de modèles de langue, consistant à utiliser deux modèles de langues (spécifique et générique) ;

(d) Interpolation linéaire des modèles de langue : chacun des modèles de langue est donc d'abord entrainé sur un corpus spécifique, puis les modèles sont interpolés linéairement en utilisant des coefficients d'interpolation estimés en minimisant la perplexité sur un corpus de développement.

Les modèles génériques *Gen* englobant toutes les données, sont construits dans le but d'avoir des modèles de traduction et de langue indépendant du domaine.

5 Expérimentations et résultats

Pour la traduction automatique, le décodeur à base de segments Moses[5] (Koehn *et al.*, 2007) est utilisé ; pour la phase d'entraînement, nous avons recours à l'aligneur sous-phrastique MGIZA++[6] (Gao & Vogel, 2008). 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 (Gahbiche-Braham *et al.*, 2012) a été utilisé pour le prétraitement de l'arabe, consistant en particulier à normaliser les proclitiques et à présegmenter les tokens arabes en des unités plus courtes et plus facilement appariables avec des mots français. Le protocole expérimental est présenté dans

[5] http://www.statmt.org/moses/.
[6] http://www.kyloo.net/software/doku.php/mgiza:overview.

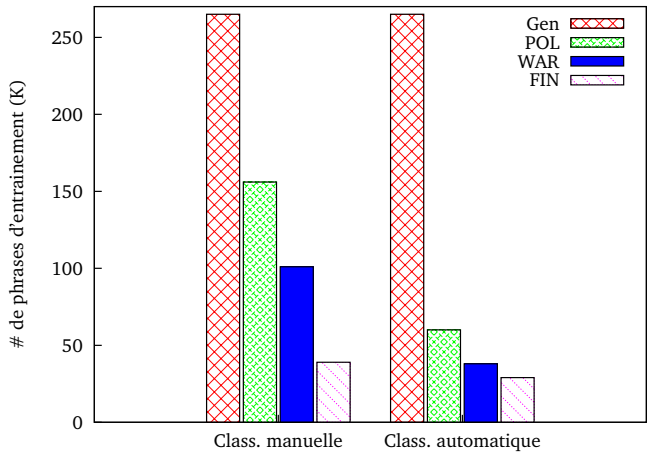

FIG. 3 – Nombre de phrases d'apprentissage par catégorie pour les deux scénarios : classification manuelle et automatique.

la section 5.1. Les modèles spécifiques sont évalués dans la section 5.2 et les différentes stratégies d'adaptation dans la section 5.3.

5.1 Protocole expérimental

Le corpus d'entraînement initial décrit dans la section 3 a été subdivisé pour construire des sous-corpus spécifiques à chacune des catégories IPTC choisies. Une *vérification manuelle du corpus de test initial* a été réalisée pour attribuer une seule catégorie à chaque phrase de l'ensemble de test initial (contenant initialement 40 % de phrases multicatégoriques). Pour les phrases multicatégoriques, la catégorie attribuée manuellement est une des catégories attribuées à la dépêche dont elle est extraite. Les phrases ne correspondant à aucune des catégories (POL, WAR et FIN) considérées dans l'étude sont affectées à la catégorie *Autre*. Le tableau 3 permet de comparer la répartition des phrases du corpus de test initial selon chaque catégorie, dans le cas d'un étiquetage manuel en comparaison à un étiquetage automatique. Avec la classification

	# phrases (%)	
Catégorie	Class. manuelle	Class. automatique
POL	56,1	33,7
WAR	20,7	15,8
FIN	16,2	14,9
Autre	7	35,6

TAB. 3 – Pourcentage des phrases pour chaque catégorie du corpus de test initial par rapport à l'ensemble des phrases, vérifié manuellement et classifié automatiquement.

automatique, 35,6 % des phrases sont attribuées à d'autres catégories que les trois catégories spécifiques.

Le tableau 4 montre le pourcentage des phrases ayant gardé les mêmes catégories pour la classification manuelle et automatique ainsi que le pourcentage des phrases ayant été classifiées dans les autres catégories. On observe que la classification automatique attribue la même catégorie que celle de l'étiquetage manuel dans environ 44 % des cas (la somme des termes sur la diagonale). Pour la catégorie POL par exemple, 25,3 % des phrases parmi 56,1 % ont gardé la même catégorie (c'est-à-dire 45 % des phrases étiquetées initialement POL). Avec la classification automatique, 33,7 % des phrases du corpus de test initial sont affectées à la catégorie POL (voir tableau 3), parmi ces phrases 5 % ont été initialement attribuées à la catégorie WAR, 2,2 % à la catégorie FIN et 1,2 % à d'autres catégories (première colonne du tableau 4).

On observe que pour certaines phrases, la nouvelle catégorie assignée automatiquement est plus appropriée que la catégorie assignée manuellement. C'est par exemple le cas de la phrase « *Afghanistan : six civils tués par une bombe artisanale*

Man \ Auto	POL	WAR	FIN	Autre
POL	**25,3**	6,1	5,5	19,2
WAR	5	**6,9**	0,9	7,9
FIN	2,2	1,2	**7,8**	5
Autre	1,2	1,6	0,7	**3,5**

TAB. 4 – Pourcentage des phrases ayant gardé les mêmes catégories pour la classification manuelle et automatique (en diagonale), ainsi que le pourcentage des phrases ayant été classifiées dans les autres catégories.

dans l'est. » initialement attribuée à la catégorie POL et reclassifiée dans la catégorie WAR ; ou encore de la phrase « *La zone euro sous pression pour maîtriser l'incendie de la dette.* » initialement attribuée à la catégorie WAR et reclassifiée automatiquement dans la catégorie FIN.

La traduction automatique de ce corpus de test initial, en utilisant l'ensemble des données d'entraînement, sans effectuer aucune adaptation aux catégories, donne un score BLEU de $33,47$.

5.2 Évaluation des modèles spécifiques

En se basant sur la classification de l'AFP, des modèles de langue (LM) spécifiques et des modèles de traduction spécifiques (TM) ont été construits. Un système de traduction général et trois systèmes de traduction spécifiques (POL, WAR et FIN) ont été construits. Le tableau 5 montre les résultats de la traduction automatique des trois tests spécifiques ainsi que le test initial (décrit dans le tableau 3) sur ces trois modèles spécifiques.

	BLEU			
Test	POL	WAR	FIN	Initial
POL	**30,94**	32,82	**29,89**	**31,73**
WAR	28,10	**32,96**	25,61	29,72
FIN	25,25	26,67	28,00	25,34

TAB. 5 – Évaluation des systèmes spécifiques – optimisés sur les corpus de développement spécifiques – sur les corpus de tests spécifiques POL, WAR et FIN et sur le test initial (présenté dans le tableau 3).

On observe que les tests spécifiques POL et WAR sont mieux traduits par les modèles spécifiques correspondants, ce qui confirme l'intérêt d'une traduction dépendant de la catégorie. En revanche, ceci n'est pas vrai pour le test spécifique FIN, pour lequel le meilleur résultat est obtenu avec le modèle spécifique POL. Ceci peut s'expliquer d'une part par le fait que le modèle spécifique FIN est entraîné sur quatre fois moins de données que le modèle POL, et d'autre part par le fait que 40 % des données utilisées pour entraîner le modèle POL sont aussi utilisées pour entraîner le modèle FIN (voir tableau 2). Le système spécifique POL donne également le meilleur résultat de traduction sur le corpus de test initial, puisque la plupart des phrases de ce test sont affectées à la catégorie POL dominante (voir tableau 3) et les modèles POL sont entraînés sur plus de données que pour les modèles WAR et FIN (voir figure 3).

5.3 Évaluation des stratégies d'adaptation

Les performances en BLEU des différentes méthodes d'adaptation décrites en section 4.2 sont reportées dans le tableau 6. Le tableau présente les résultats sur les tests spécifiques (classifiés manuellement) et sur le test initial (classifié manuellement et automatiquement).

Les résultats du système sans adaptation (*Baseline*) constitué d'un modèle de langue général (LM Gen) et d'un modèle de traduction général (TM Gen) sont comparés aux différentes méthodes d'adaptation qui combinent les modèles généraux avec les modèles spécifiques (Spe).

Dans le cas de l'adaptation, chaque phrase est traduite avec le système adapté à sa catégorie. Deux types de systèmes

sont considérés : les premiers entraînés sur des données classifiées automatiquement (colonnes 1, 2, 3 et 5), les seconds entraînés sur des données classifiées manuellement (AFP) (colonne 4). Les phrases classifiées dans la catégorie *Autre* sont traduites avec le système *Baseline* dans lequel les modèles de traduction et de langue sont généraux.

			BLEU				
			Tests spécifiques Class. manuelle			Test initial	
						Class. manuelle	Class. automatique
Approche	TM	LM	POL	WAR	FIN		
Baseline	Gen	Gen	32,08	34,87	31,71	33,47	33,47
(a)	Merge	Gen	30,60	34,68	31,15	32,61	32,41
(b)	Spe+Gen	Gen	31,06	34,26	31,08	32,74	32,41
(c)	Gen	Spe+Gen (log-lin)	31,50	34,62	31,62	32,57	32,85
(b)+(c)	Spe+Gen	Spe+Gen (log-lin)	31,52	33,96	30,43	32,91	32,26
(d)	Gen	Spe+Gen (lin)	**31,94**	**34,84**	**31,83**	**32,98**	**33,14**
(b)+(d)	Spe+Gen	Spe+Gen (lin)	31,37	34,31	31,54	32,85	32,66

TAB. 6 – Traduction automatique en utilisant différentes méthodes d'adaptation, et évaluation sur les tests spécifiques POL, WAR et FIN ainsi que sur le corpus de test initial : chaque phrase est traduite avec le système de traduction spécifique optimisé sur le corpus de développement spécifique et correspondant à la catégorie qui lui est attribuée. Comparaison entre les systèmes construits à partir de modèles spécifiques construits par classification manuelle et les modèles spécifiques construits par classification automatique.

On observe que l'approche (d) - système construit par un modèle de langue interpolé linéairement (lin) et un modèle de traduction général - donne les meilleures performances pour les trois tests spécifiques. On note que le meilleur score pour la classification manuelle du test initial est obtenu également par la même approche.

Une petite amélioration de 0,18 points BLEU pour l'approche (c) et 0,16 points pour l'approche (d) sont observées pour la classification automatique. Bien que les modèles soient plus petits, la performance de traduction n'est pas dégradée. De même avec la classification automatique, les meilleures performances sont données par l'approche (d).

Finalement aucun des modèles adaptés ne permet d'obtenir des performances se traduisant par un meilleur BLEU que le système sans aucune adaptation (Baseline). Ces résultats doivent être relativisés par une étude des sorties de traduction. Les figures 4 et 5 montrent deux exemples de sorties extraites de traductions du corpus de test initial : la phrase en arabe, la référence, la traduction sans adaptation et la traduction avec adaptation en utilisant une catégorisation automatique.

Arabe	[...] ينبغي حل مجلس الامه واحالة النواب الفاسدين علي القضاء
Référence :	le **parlement doit être dissous** et les députés corrompus traduits en justice [...]
Trad. sans adaptation :	il faut une **solution du Conseil de la Nation** et pour traduire en justice les corrompus [...]
Trad. avec adaptation :	il faut la **dissolution du Parlement** et de traduire les députés corrompus [...]

FIG. 4 – Phrase appartenant à la catégorie POL avec sa traduction de référence et les traductions automatiques produites par un modèle sans adaptation et un modèle avec adaptation à la catégorie.

La figure 4 correspond à une phrase appartenant à la catégorie POL. En arabe, le mot حل est ambigu et a pour signification *solution* ou *dissolution* selon le contexte. L'approche avec adaptation permet de résoudre correctement cette ambiguïté ce qui n'est pas le cas pour la traduction sans adaptation. Dans le même temps, le modèle avec adaptation propose une traduction de مجلس الامة plus proche de la référence (*Parlement*) que le modèle sans classification (*Conseil de la Nation*). Ce dernier groupe de mots existe dans le modèle de langue général mais n'existe pas dans le modèle de langue spécifique

à la catégorie POL. Il est à noter également que les mots مجلس et الامة peuvent être traduits indépendamment par *conseil* et *nation*. L'utilisation d'un modèle de langue adapté aide dans ce cas à améliorer la traduction de ce groupe de mots.

L'exemple de la figure 5 permet d'observer également une amélioration d'une sortie de traduction par adaptation à la catégorie. La traduction proposée par le système avec adaptation à la catégorie FIN est plus proche de la référence que celle donnée par le système sans adaptation.

Arabe	وقد انسحبت منها بيونغ يانغ في نيسان / ابريل 2008 [...]
Référence	Pyongyang s' en était officiellement retiré en avril 2008 [...]
Trad. sans classification :	et officiellement , dont Pyongyang a claqué la porte en avril 2008 [...]
Trad. avec class. automatique :	Pyongyang a s' en était retiré officiellement en avril 2008 [...]

FIG. 5 – Amélioration de la traduction (phrase de la catégorie FIN)

On observe donc que, même si cela ne se traduit pas par une amélioration en BLEU, l'adaptation permet dans certains cas d'améliorer les sorties de traduction.

6 Conclusion

Dans cet article, nous avons analysé les résultats d'un ensemble d'expériences d'adaptation thématique pour la traduction automatique. La particularité de notre approche est que nous utilisons un corpus pré-classifié (par les journalistes de l'AFP selon une classification du standard IPTC). Cette classification est utilisée pour construire des modèles spécifiques et comparer plusieurs approches d'adaptation. Deux scénarios de traduction automatique sont proposés : une adaptation reposant sur la *classification manuelle* des phrases et l'autre sur une *classification automatique* en catégorie. Dans les deux cas la traduction est effectuée après la détection des catégories en choisissant le modèle adapté à la catégorie.

Bien que certains exemples montrent que l'adaptation des modèles de traduction à la catégorie permet dans certains cas de désambiguïser la bonne traduction, ceci ne se traduit pas par une amélioration en terme de BLEU.

Deux raisons principales expliquent ces performances décevantes : (i) la trop petite taille des corpus spécialisés, qui conduisent à des modèles probablement plus précis, mais également plus lacunaires ; (ii) le fait que les catégories IPTC (celles utilisées pour cette étude) sont parfois très proches. En particulier, l'effet du manque de données pour estimer les modèles de traduction a un impact important sur les performances finales.

Parmi les 17 catégories IPTC, très peu sont représentées en quantité suffisante dans les données que nous avons traitées. Le choix des catégories (POL, WAR, FIN) a été effectué en se basant sur la quantité de données. Mais ces catégories contiennent beaucoup de phrases communes, ce qui rend leurs frontières floues.

Lors de l'étiquetage manuel du test initial nous avons constaté qu'il est parfois difficile de contraindre l'annotation d'une phrase à une seule catégorie. La phrase ci-dessous (en arabe, avec sa traduction en français) par exemple est extraite d'une dépêche très récente affectée par les journalistes aux catégories ACE (arts, culture et divertissement), POL (politique) et SPO (sport).

عبرت وسائل الاعلام الدولية عن اعجابها بحفل افتتاح دورة الالعاب الاولمبية الشتوية في مدينة سوتشي الروسية، لكن الكثير منها اشار الي الرسالة السياسية التي تنطوي عليها واهميتها في نظر الرئيس فلاديمير بوتين.

Les médias internationaux ont exprimé leur admiration pour la cérémonie d'ouverture des Jeux olympiques d'hiver à la

ville russe de Sotchi, mais beaucoup d'entre eux ont souligné le message politique en cause et son importance aux yeux du président Vladimir Poutine.

S'agissant des jeux olympiques, la phrase ci-dessus doit effectivement être affectée à la catégorie sport. Mais elle traite également du divertissement (cérémonie d'ouverture) et surtout bien sûr de politique. Plusieurs catégories doivent donc être affectées à cette phrase.

Parmi les perspectives pour la suite de nos travaux, il sera en particulier intéressant d'explorer des approches autorisant la multi-catégorisation. Des modèles combinés peuvent être crées et utilisés pour traduire des phrases multi-catégoriques.

Références

BANERJEE P. (2012). *Domain Adaptation for Statistical Machine Translation of Corporate and User-Generated Content*. PhD thesis, Dublin City University.

DE MORI R. & FEDERICO M. (1999). *Language Model Adaptation*, In K. PONTING, Ed., *Computational models of speech pattern processing volume 169, NATO ASI*, p. 280–303. Springer Verlag : Prague, Czech Republic.

EIDELMAN V., BOYD-GRABER J. & RESNIK P. (2012). Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers - Volume 2*, ACL '12, p. 115–119, Stroudsburg, PA, USA : Association for Computational Linguistics.

FOSTER G., GOUTTE C. & KUHN R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 451–459, Stroudsburg, PA, USA : Association for Computational Linguistics.

FOSTER G. & KUHN R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, p. 128–135, Stroudsburg, PA, USA : Association for Computational Linguistics.

GAHBICHE-BRAHAM S., BONNEAU-MAYNARD H., LAVERGNE T. & YVON F. (2012). Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier. In *Proc. of LREC'12*, p. 2107–2113, Istanbul, Turkey.

GAHBICHE-BRAHAM S., BONNEAU-MAYNARD H. & YVON F. (2011). Two ways to use a noisy parallel news corpus for improving statistical machine translation. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, p. 44–51, Portland, Oregon : Association for Computational Linguistics.

GAO Q. & VOGEL S. (2008). Parallel implementations of a word alignment tool. In *In* Software Engineering, Testing, and Quality Assurance for Natural Language Processing, p. 49–57.

HADDOW B. & KOEHN P. (2012). Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada : Association for Computational Linguistics.

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. (2007). Moses : open source toolkit for statistical machine translation. In *In* Proceedings of the 45th Annual Meeting of the ACL, p. 177–180.

KOEHN P. & SCHROEDER J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, p. 224–227, Stroudsburg, PA, USA : Association for Computational Linguistics.

LANGLAIS P. (2002). Improving a general-purpose statistical translation engine by terminological lexicons. In *COLING-02 on COMPUTERM 2002 : second international workshop on computational terminology - Volume 14*, COMPUTERM '02, p. 1–7, Stroudsburg, PA, USA : Association for Computational Linguistics.

NAKOV P. (2008). Improving English-Spanish statistical machine translation : experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, p. 147–150, Stroudsburg, PA, USA : Association for Computational Linguistics.

NIEHUES J. & WAIBEL A. (2010). Domain adaptation in statistical machine translation using factored translation models. *Proceedings of EAMT*.

OCH F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, p. 160–167, Stroudsburg, PA, USA : Association for Computational Linguistics.

RIGOUSTE L., CAPPÉ O. & YVON F. (2007). Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing and Management*, **43**(5), 1260–1280.

SENNRICH R. (2012a). Mixture-modeling with unsupervised clusters for domain adaptation in statistical machine translation. In *Proceedings of the 16th EAMT Conference*, p. 185–192, Trento, Italy.

SENNRICH R. (2012b). Perplexity minimization for translation model domain adaptation in statistical machine translation. In W. DAELEMANS, M. LAPATA & L. MÀRQUEZ, Eds., *EACL 2012, 13th Conference of the European Chapter of the ACL, Avignon, France, April 23-27, 2012*, p. 539–549 : Association for Computational Linguistics.

SENNRICH R., SCHWENK H. & ARANSA W. (2013). A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 832–840, Sofia, Bulgaria : Association for Computational Linguistics.

SHAH K., BARRAULT L. & SCHWENK H. (2010). Translation model adaptation by resampling. In *WMT, Association of Computational Linugistics (ACL)*, Uppsala (Sweden).

YAMAMOTO H. & SUMITA E. (2008). Bilingual cluster based models for statistical machine translation. *IEICE Transactions*, **91-D**(3), 588–597.

ZHAO B., ECK M. & VOGEL S. (2004). Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA : Association for Computational Linguistics.