Évaluation morphologique pour la traduction automatique: adaptation au français

Franck Burlot François Yvon
LIMSI, CNRS, Univ. Paris-Sud, Université Paris Saclay, 91 403 Orsay, France
prénom.nom@limsi.fr

RÉSUMÉ
Le nouvel état de l'art en traduction automatique (TA) s'appuie sur des méthodes neuronales, qui différent profondément des méthodes utilisées antérieurement. Les métriques automatiques classiques sont mal adaptées pour rendre compte de la nature du saut qualitatif observé. Cet article propose un protocole d'évaluation pour la traduction de l'anglais vers le français spécifiquement focalisé sur la compétence morphologique des systèmes de TA, en étudiant leurs performances sur différents phénomènes grammaticaux.

ABSTRACT
Morphological Evaluation for Machine Translation : Adaptation to French

While the state of the art in machine translation has recently changed, it is regularly acknowledged that automatic metrics do not provide enough insights to fully measure the observed qualitative leap. This paper proposes an evaluation protocol for translation from English into French specifically focused on the morphological competence of a system with respect to various grammatical phenomena.

MOTS-CLÉS : Traduction automatique, évaluation de la TA, morphologie.

KEYWORDS: Machine translation, MT evaluation, morphology.

1 Introduction

Le domaine de la traduction automatique (TA) statistique a été récemment transformé par l'arrivée à maturité de nouveaux systèmes de TA reposant massivement sur des architectures neuronales (Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014), qui constituent aujourd'hui le nouvel état de l'art du domaine. Ces nouvelles architectures semblent en particulier capables de détecter (dans la langue source) et de modéliser (dans la langue cible) des dépendances entre mots distants et ainsi de mieux restituer des associations lexicales (collocations, expressions figées) ainsi que des accords grammaticaux (Bentivogli *et al.*, 2016; Isabelle *et al.*, 2017; Sennrich, 2017). Cette amélioration des performances s'est faite au détriment de la prédictibilité et de la transparence des architectures de calcul, dont le fonctionnement s'avère particulièrement opaque et complexe à diagnostiquer.

L'avènement des systèmes neuronaux doit donc s'accompagner du développement de nouvelles méthodes d'évaluation automatique : d'une part parce que le score BLEU (Papineni *et al.*, 2002) ne suffit plus à distinguer des systèmes qui produisent tous des sorties extrêmement fluides ; d'autre part afin de mieux comprendre la capacité des méthodes neuronale à résoudre plus ou moins bien telle ou telle difficulté de traduction, et ainsi d'orienter les évolutions de ces systèmes. Cette ambition a

donné lieu dans les années récentes à une floraison de travaux sur l'évaluation de la TA neuronale, que nous survolons ci-dessous (§ 2.1).

La principale contribution de ce travail est d'étendre l'approche récemment proposée par Burlot & Yvon (2017) pour évaluer les performances morphologiques des architectures neuronales au cas de la traduction vers le français. En plus de traiter une langue supplémentaire, nous introduisons également de nouveaux tests pour le français, qui pourront également être utilisés pour d'autres langues. Après avoir rappelé les principes de la méthode (§ 2), nous décrivons les principaux tests utilisés pour le français (§ 3), puis présentons les résultats d'une comparaison des performances morphologiques de plusieurs systèmes de TA (§ 4), qui permettent d'éclairer l'apport des méthodes neuronales par rapport aux systèmes à base de segments (Koehn, 2010; Allauzen & Yvon, 2011) pour la traduction vers le français. Les scripts et données utilisées dans cette étude sont librement disponibles [1].

2 Principes de l'évaluation morphologique

2.1 Motivations et fondements

La littérature récente sur le diagnostic de TA neuronale peut être organisée en quatre grandes familles : la première s'appuie sur des typologies d'erreurs (Vilar *et al.*, 2006; Lommel *et al.*, 2014) pour catégoriser des erreurs dans les sorties des systèmes, et peuvent impliquer soit une analyse manuelle, souvent difficile ; soit une analyse automatisée (Popović & Ney, 2011; Toral Ruiz & Sánchez-Cartagena, 2017; Klubička *et al.*, 2017), qui se fonde alors sur une comparaison de surface entre la sortie et une traduction de référence [2]. La seconde reprend la tradition ancienne (King & Falkedal, 1990) des jeux de tests (*test suites*) spécifiquement conçus pour mettre en défaut la résolution, par les systèmes de TA, d'un problème linguistique particulier. Isabelle *et al.* (2017) propose un tel jeu de test pour la paire (anglais, français) qui inclut à la fois des difficultés d'ordre morpho-syntaxique (phénomènes d'accord, concordance des modes et temps, etc.), lexical (mots polysémiques, idiomes et expressions figées, etc.), et syntaxique (divergences dans la construction de groupes verbaux, dans la construction de propositions relatives, etc). Cette approche peut être critiquée au regard de l'expertise humaine nécessaire à la création des phrases test comme à l'évaluation des erreurs du système ; par ailleurs, la représentativité des tests et la gravité des erreurs n'est également pas prise en compte.

Une troisième manière de procéder est plus indirecte et consiste à n'utiliser que les scores des systèmes (et non leur sortie) : la qualité d'un système se mesure alors à sa capacité à donner un meilleur score (une plus forte probabilité) à une phrase correcte par rapport à une phrase délibérément altérée pour simuler une faute particulière. Cette méthodologie est utilisée, par exemple, pour évaluer les modèles de langue neuronaux par Linzen *et al.* (2016), qui s'intéressent spécifiquement aux erreurs d'accord (entre sujet et verbe) : le système sera alors considéré comme défaillant s'il assigne à la phrase altérée un meilleur score qu'à la phrase correcte. Sennrich (2017) applique cette stratégie à grande échelle à la TA (de l'anglais vers l'allemand), en engendrant automatiquement des erreurs reflétant des fautes d'accord, des mots inconnus, etc. Si cette méthode permet de s'affranchir de l'intervention d'un expert humain, elle ne permet qu'une évaluation approximative des performances, puisqu'il n'est pas assuré que les deux phrases comparées correspondent à des sorties réelles du système.

1. Voir `https://morpheval.limsi.fr/`
2. Une erreur de morphologie correspond alors à l'observation dans la sortie de la TA d'un lexème présent dans la référence, mais avec une marque flexionnelle différente.

	source	cible
base	he is very happy	il est très heureux
variante	he **was** very happy	il **était** très heureux

FIGURE 1 – Un exemple de test contrastif - après manipulation du temps verbal, on vérifie que la traduction de la variante présente bien un passage au passé par rapport à la traduction de la base.

La méthode proposée par Belinkov *et al.* (2017) est encore plus détournée : elle consiste à comparer les plongements lexicaux (*embeddings*) appris par l'encodeur (ou le décodeur) du système de traduction du point de vue de leur capacité à prédire correctement des tâches de nature morphologique, en l'espèce un étiquetage morpho-syntaxique. Elle n'apporte donc que peu d'information pouvant aider au diagnostic. Si cette approche permet de comparer de manière automatique plusieurs manières de décomposer les mots sources ou cibles, ou de comparer plusieurs couches du réseau, elle dit en revanche peu de chose sur les erreurs morphologiques dans un contexte de traduction (voir également dans la même lignée le travail de Vania & Lopez (2017)).

Comme détaillé ci-dessous, l'approche proposée dans (Burlot & Yvon, 2017) se distingue des méthodes existantes sous divers aspects : (a) elle vise à obtenir un diagnostic *entièrement automatique* portant sur (b) des difficultés morphologiques spécifiques et (c) avérées dans les sorties de traduction automatique ; l'intervention humaine est limitée à la conception des tests, et permet (d) de produire des tests en grande quantité, permettant d'éviter que les systèmes s'adaptent à un jeu de test particulier. La contrepartie est le caractère approximatif de la détection d'erreurs, qui peut toutefois être contrôlé en augmentant le nombre de cas tests.

2.2 Les tests contrastifs

La méthode que nous avions initialement proposée (Burlot & Yvon, 2017) repose sur la notion de *test contrastif*. Dans sa version la plus élémentaire, elle consiste à construire des paires formées de phrases sources comportant une différence minimale : l'une (*la base*) par exemple contiendra un pronom objet masculin, et l'autre (*la variante*) le même objet au féminin. On observe alors les différences entre les traductions de ces deux phrases - lorsqu'elles ne se distinguent que par l'expression (en cible) du trait morphologique qui est manipulé dans la source, on considère que le système a bien reproduit le contraste ; dans le cas inverse on le jugera défaillant.

Deux ensembles de tels tests sont considérés (les ensembles A et B de Burlot & Yvon (2017)). Il existe une seconde famille de tests contrastifs (l'ensemble C) qui se focalisent sur la cohérence des traductions : elle vise à vérifier que les choix de traduction restent cohérents lorsque l'on produit plusieurs variantes de la même base. Ainsi, on pourra remplacer un nom par des synonymes, et vérifier que les propriétés morphologiques de la traduction (en langue cible) de toutes les variantes sont les mêmes et ne dépendent pas de spécificités lexicales : on évalue donc ici plutôt le caractère systématique du fonctionnement du système.

Dans cette approche, les traitements automatiques interviennent à deux étapes : (a) lors de la génération de la ou des variantes, qui demande une analyse grammaticale de la phrase source ; (b) lors du calcul des différences minimales, qui n'exploite que des dictionnaires. Ces deux étapes étant susceptibles d'introduire des erreurs, nous avions proposé de multiplier les paires minimales afin d'obtenir une mesure approximativement correcte du comportement du système. Une évaluation de précision des

Base/Variante(s)	Sortie	Évaluation
	Test-A	
I am hungry	j'ai faim	verbe au passé
I was hungry	j'**avais** faim	trouvé
	Test-B	
I see them	je les vois	le nom et l'adjectif
I see crazy researchers	je vois des **chercheurs fous**	sont au pluriel
	Test-C	
a big responsibility	une **grande** responsabilité	tous les adjectifs
a small responsibility	une **petite** responsabilité	sont au féminin
an important responsibility	une **importante** responsabilité	
a ridiculous responsibility	une responsabilité **ridicule**	
a terrible responsibility	une **terrible** responsabilité	Entropie = 0.0

FIGURE 2 – Exemples de phrases réussissant les tests.

résultats est présentée à la section 4.4.

3 Génération de tests pour le français

3.1 Sélection des phrases tests

La sélection des phrases de base se fonde principalement sur un critère de simplicité, qui accélère la traduction et facilite les traitements automatiques : on se limite à des phrases d'au plus 15 tokens. Ces phrases sont extraites des corpus anglais monolingues News Crawl 2007 et 2008. [3] La production des variantes est plus complexe, en particulier pour les variations lexicales. Le principe général est de remplacer un mot de la base par un mot ou groupe de mots pour produire la variante. Nous utilisons à cet effet l'étiqueteur morpho-syntaxique CoreNLP (Manning *et al.*, 2014) afin de localiser la partie du discours du mot à remplacer, puis le générateur morphologique Pymorphy [4] pour produire la flexion désirée. La génération de synonymes et d'antonymes est effectuée avec WordNet (Miller, 1995). Une dernière étape consiste à employer un modèle de langue (Heafield, 2011) entraîné sur toutes les données monolingues anglaises mises à disposition pour la campagne WMT 2015 [5] pour sélectionner les variantes les plus fluides. Cette sélection aboutit à 500 groupes de phrases pour chaque test.

3.2 Test-A : transformations morpho-syntaxiques

Le premier groupe de tests consiste à modifier la flexion d'un mot dans la base, puis à évaluer la présence du même contraste côté cible. Les tests présentés ici adaptent, pour partie, les propositions de Burlot & Yvon (2017). C'est le cas des tests portant (a) sur le temps des verbes, où un verbe au présent est remplacé respectivement par les formes au `passé` et au `futur` ; (b) sur le `nombre` d'un

3. `http://statmt.org/wmt17/translation-task.html`
4. `http://pymorphy.readthedocs.io/`
5. `http://statmt.org/wmt15/translation-task.html`

pronom objet initialement au singulier ; (c) sur la `négation` d'une base à l'affirmatif ; (d) sur le `comparatif` où un adjectif au comparatif dans la base est remplacé par une forme neutre.

Le français présentant une morphologie verbale riche, deux tests supplémentaires sont proposés. Le premier évalue la génération du `conditionnel` : pour ce faire, nous remplaçons l'auxiliaire *will* dans la base par le modal *would* et testons le contraste indicatif/conditionnel dans les phrases cibles. Le second évalue le passage d'un verbe de l'indicatif au `subjonctif` ; les paires contrastives sont produites en recherchant des phrases introduites par des propositions principales du type *I believe*, changées dans les variantes en *I don't believe*. Il est attendu que, pour la variante, la traduction de la proposition subordonnée comprenne un verbe au subjonctif. Enfin, nous ajoutons également un test concernant le `superlatif`, produit de manière identique à celui du comparatif.

De manière générale, l'évaluation se déroule de la manière suivante : chacun des mots de la variante traduite qui n'est pas présent dans la base traduite est récolté et analysé au moyen du dictionnaire Lefff[6] (Sagot, 2010). Si le dictionnaire ne propose aucune analyse pour le mot trouvé, la paire de phrase est rejetée du test, puisqu'il est dans ce cas impossible de déterminer si elle reflète ou non le contraste attendu. Ne sont retenus que les mots qui contiennent dans leur analyse la partie du discours évaluée (par exemple les verbes dans le test du subjonctif). Enfin, l'étiquette morphologique reflétant le phénomène grammatical du test est recherchée : si elle est présente, un succès est rapporté.

3.3 Test-B : transformations lexicales

Ces tests ont pour objectif d'évaluer la capacité d'un système à modéliser différentes formes d'accord grammatical. De même que dans la partie précédente, certains tests sont repris de Burlot & Yvon (2017). Ainsi, `verbes coord` consiste à changer le verbe de la base en un groupe verbal coordonné (*he eats* → *he eats and drinks*), puis à vérifier dans la traduction que les deux verbes coordonnés contiennent bien les mêmes marques de nombre, de personne et temps/mode. Le test de syntagmes nominaux (`synt nom`) est similaire : un pronom dans la base est modifié en syntagme nominal `ADJ+NOM` dans la variante. L'évaluation vérifie (séparément) l'accord en genre et l'accord en nombre entre l'adjectif et le nom dans la cible française.

Un nouveau test (`coréf`) concerne les liens de coréférence qui existent entre un pronom personnel et son antécédent nominal. Les bases sont sélectionnées lorsqu'elles contiennent un lien de coréférence détecté par l'étiqueteur de CoreNLP[7] (Manning *et al.*, 2014) ; l'antécédent nominal dans la base est substitué par un synonyme et l'on vérifie alors que les pronoms sont correctement accordés en nombre ou en genre.

Notons que pour `verbes coord` et `coréf`, l'utilisation de contrastes permet de projeter des annotations depuis l'anglais vers le français, ce qui est un usage quelque peu différent de celui utilisé pour le premier jeu de tests. Pour `coréf`, en observant ce qui co-varie dans les traductions de la base et de la variante, il devient possible de localiser l'antécédent du pronom dans les deux phrases cibles. Dans ce cas, chaque paire nous permet d'évaluer deux traductions, et les scores sont récoltés sur la base et sur la variante.

6. http ://alpage.inria.fr/ sagot/lefff.html

7. Afin de privilégier la précision aux dépens du rappel, nous avons conservé les bases étiquetées positivement à la fois par le modèle de base et par le modèle neuronal de la boîte à outils.

	verbes							
Système	passé		futur		cond.		subj.	
Moses	69,5%	321/462	88,7%	422/476	62,3%	299/480	86,0%	430/500
Nematus	74,4%	349/469	76,2%	356/467	62,9%	303/482	85,5%	425/497
+rétro-trad.	84,3%	402/477	85,2%	410/481	83,1%	398/479	91,4%	457/500

	pronoms		adjectifs				autres			
Système	nb		compar.		super.		nég.		noms plur.	
Moses	82,8%	414/500	80,4%	402/500	78,0%	390/500	97,0%	485/500	82,6%	395/478
Nematus	76,4%	382/500	82,2%	411/500	87,4%	437/500	96,6%	483/500	86,8%	400/461
+rétro-tr.	87,8%	439/500	87,2%	436/500	90,2%	451/500	98,8%	494/500	89,2%	415/465

TABLE 1 – Évaluation des paires de phrases (test-A).

	verbes coord.					
Système	nb.		pers.		TM	
Moses	97,5%	394/404	97,0%	392/404	95,3%	385/404
Nematus	94,8%	423/446	94,6%	422/446	94,4%	421/446
+rétro-trad.	97,8%	435/445	98,4%	438/445	98,0%	436/445

	synt. nom.				coréférence	
Système	genre		nb.		genre	
Moses	94,4%	356/377	92,0%	347/377	83,2%	691/831
Nematus	95,8%	365/381	95,8%	365/381	89,4%	787/880
+rétro-trad.	97,9%	375/383	98,4%	377/383	88,4%	827/936

TABLE 2 – Évaluation des paires de phrases (test-B).

3.4 Test-C : tests de cohérence

La troisième famille de tests est quelque peu différente : pour chaque base, on produit 4 variantes, et l'on mesure la cohérence des choix du système de TA en mesurant l'entropie du trait morphologique contrôlé : un système idéal doit produire toujours les mêmes traits (ce qui correspond à une entropie nulle), alors qu'un système incohérent produira une entropie maximale. Cinq tests sont considérés, trois qui portent sur les verbes (respectivement sur le nombre, le genre, et le couple (temps, mode) TM , et deux sur les adjectifs (pour le genre et le nombre). Pour chacun, on rapporte l'entropie moyennée sur tous les groupes de phrases du test.

Système	BLEU
Moses	32,25
Nematus	33,06
+rétro-trad.	34,11

TABLE 3 – Scores BLEU (Newstest-2014).

	verbes			adjectifs	
Système	nb.	pers.	TM	genre	nb.
Moses	0,075	0,039	0,099	0,131	0,131
Nematus	0,040	0,033	0,080	0,076	0,052
+rétro-trad.	0,024	0,015	0,066	0,065	0,049

TABLE 4 – Évaluation des groupes de phrases (test-C).

4 Évaluation

4.1 Systèmes et données

Les systèmes choisis pour illustrer notre méthode d'évaluation de la morphologie sont représentatifs de l'évolution récente de la TA et nous en présentons ici deux types : statistique et neuronal.

Le système statistique est basé sur Moses (Koehn *et al.*, 2007). Il est entraîné sur 4 millions de phrases parallèles provenant des données fournies à WMT 2017 (plus précisément des corpus EUbookshop, MultiUN, News-Commentary-11 et Wikipedia). Le modèle de langue employé par le système a été entraîné avec KenLM (Heafield, 2011) sur le côté cible des données parallèles, auxquelles ont été ajoutées environ 10 millions de phrases issues du corpus news-2014 fourni à WMT 2015.

Le système neuronal a été entraîné avec la boîte à outils Nematus (Sennrich *et al.*, 2017) sur les mêmes phrases parallèles que le système statistique. Les vocabulaires source et cible ont été traités avec un modèle bilingue de *Byte Pair Encoding* (Sennrich *et al.*, 2016b), paramétré à 50 000 opérations de fusion. Ce traitement a conduit à un vocabulaire anglais de plus de 32 000 unités et à un vocabulaire français de moins de 36 000 unités.

Un second traducteur neuronal reprend les paramètres du système précédent, qui sont employés pour initialiser l'entraînement d'un nouveau système optimisé sur 2 millions de phrases sélectionnées aléatoirement parmi les données initiales, complété par 2 millions de phrases françaises extraites du corpus Europarl (Koehn, 2005) *rétro-traduites* (Sennrich *et al.*, 2016a) vers l'anglais au moyen d'un système neuronal français-anglais similaire au système décrit supra pour la direction anglais-français.

4.2 Résultats

Les scores BLEU (Papineni *et al.*, 2002) pour ces trois systèmes ont été calculés sur Newstest-2014 et sont dans le tableau 3. Ils distinguent sensiblement le système statistique des systèmes neuronaux qui obtiennent 1 à 2 points de plus. Les précisions obtenues sur les tâches du test-A apparaissent au tableau 1 et contredisent quelque peu les scores BLEU. C'est ce que l'on observe pour le futur, qui obtient la plus haute précision, mais aussi pour le conditionnel et le subjonctif, dont les précisions sont similaires au système Nematus. Cela révèle l'efficacité relative des systèmes neuronaux dans la transmission d'une caractéristique morphologique de la source vers la cible. En effet, si ces systèmes sont réputés pour fournir une sortie plus fluide que les modèles statistiques, cela se produit parfois aux dépens de l'adéquation de la cible avec la source.

Enfin, les comparatifs et superlatifs semblent mieux pris en charge par les systèmes neuronaux. Nous posons l'hypothèse que cela est dû au caractère ouvert de leurs vocabulaires, qui permettent ainsi de générer potentiellement n'importe quel mot forme. À l'opposé, les systèmes statistiques ont un vocabulaire fixe et lorsqu'un adjectif anglais au comparatif ou au superlatif n'a pas été observé à l'entraînement, le système est incapable de générer une traduction correcte. Cette remarque tend à expliquer la meilleure performance du système statistique par rapport au système neuronal sans rétro-traduction sur la tâche des pronoms au pluriel. En effet, les pronoms correspondant à une classe de mots fermée, Moses n'a aucune difficulté à les traiter correctement. Lorsqu'en revanche ce système est confronté à une classe ouverte, comme celle des noms au pluriel, le caractère fixe de son vocabulaire limite ses performances.

tâche	Fréquences (f) des mots produits dans la variante							
	$f = 0$		$0 < f < 50$		$50 < f < 1000$		$1000 < f$	
passé	0,0%	(0/0)	81,2%	(13/16)	71,7%	(66/92)	68,5%	(241/352)
comparatif	0,0%	(0/2)	50,0%	(9/18)	76,9%	(60/78)	79,0%	(226/286)
noms plur.	0,0%	(0/1)	79,5%	(31/39)	79,0%	(94/119)	84,9%	(270/318)
v. tps/mode	87,5%	(7/8)	97,7%	(42/43)	98,4%	(121/123)	92,3%	(120/130)
coréf. genre	40,0%	(2/5)	54,8%	(34/62)	68,3%	(153/224)	71,1%	(494/695)

TABLE 5 – Performance du système Moses selon la fréquence des mots à traduire.

tâche	Fréquences (f) des mots produits dans la variante							
	$f = 0$		$0 < f < 50$		$50 < f < 1000$		$1000 < f$	
passé	50,0%	(1/2)	61,1%	(11/18)	76,7%	(66/86)	87,3%	(322/369)
comparatif	0,0%	(0/2)	11,1%	(2/18)	82,1%	(64/78)	89,5%	(256/286)
noms plur.	66,7%	(2/3)	68,6%	(24/35)	96,5%	(110/114)	89,4%	(279/312)
v. tps/mode	100,0%	(10/10)	95,1%	(39/41)	98,5%	(133/135)	99,3%	(139/140)
coréf. genre	60,0%	(3/5)	69,4%	(43/62)	81,2%	(182/224)	84,5%	(587/695)

TABLE 6 – Performance du système Nematus + rétro-traduction selon la fréquence des mots à traduire.

Les tests-B (tableau 2) ne révèlent pas de différences aussi importantes entre les systèmes. Nous constatons toutefois que le système Nematus sans rétro-traduction est le plus mauvais pour le test de coordination verbale. Notons que le système statistique emploie un modèle de langue entraîné sur une grande quantité de données monolingues, ce qui n'est pas le cas des systèmes neuronaux qui n'ont observé que quelques millions de phrases cibles. Il est donc indéniable que les modèles neuronaux tirent un bien meilleur parti d'une moindre quantité de données monolingues. La tâche de coréférence place enfin le système statistique en-dessous des systèmes neuronaux qui semblent mieux prendre en charge de tels phénomènes d'accord distants.

Les tests-C (tableau 4) témoignent d'une progression claire mettant en valeur la supériorité du système neuronal avec rétro-traduction, qui a toujours une entropie inférieure aux deux autres. La variété lexicale inhérente à ce test révèle la faiblesse du système statistique qui peine à rester cohérent dans sa prédiction morphologique lorsque le syntagme nominal n'a pas été observé à l'entraînement.

4.3 Performance sur les mots rares et inconnus

Nous proposons ici d'affiner les résultats présentés en mesurant la performance d'un système selon la fréquence du mot source sur lequel porte une tâche. Ces fréquences sont calculées sur le côté source des mêmes données parallèles employées par chaque système. Des précisions sont ainsi rapportées pour un sous-ensemble de tâches aux tableaux 5 (statistique) et 6 (neuronal avec rétro-traduction). Ainsi, pour la tâche du `passé`, nous considérons le mot source introduit dans chaque variante qui porte la marque du passé. Les phrases sont classées selon quatre intervalles de fréquences : les mots inconnus, les mots dont la fréquence est inférieure à 50, inférieure à 1000 et supérieure à 1000.

Nous pouvons constater que la génération des variantes lors de la création du jeu de test a produit peu ou pas de mots inconnus. Le système neuronal est basé sur une segmentation des mots en BPE, ce qui lui permet en théorie d'interpréter n'importe quel mot inconnu par la combinaison de plusieurs unités en source. Toutefois, nous n'observons pas ici d'amélioration significative du système neuronal par

tâche	moyenne	1000	750	500	250	100
			Test-A			
passé	76.8	±2.6	±3.0	±3.7	±5.2	±8.2
futur	83.6	±2.3	±2.6	±3.2	±4.6	±7.2
conditionnel	81.9	±2.4	±2.8	±3.4	±4.8	±7.5
subjonctif	92.8	±1.6	±1.8	±2.3	±3.2	±5.0
pron. nb	85.0	±2.2	±2.6	±3.1	±4.4	±6.9
comparatif	80.7	±2.4	±2.8	±3.5	±4.9	±7.7
superlatif	91.4	±1.7	±2.0	±2.5	±3.5	±5.4
négation	97.2	±1.0	±1.2	±1.4	±2.0	±3.0
noms plur.	85.8	±2.2	±2.5	±3.1	±4.3	±6.8
			Test-B			
v. nombre	94.8	±1.4	±1.6	±1.9	±2.7	±4.2
v. personne	94.4	±1.4	±1.6	±2.0	±2.8	±4.4
v. tps/mode	93.9	±1.5	±1.7	±2.1	±2.9	±4.6
SN genre	96.7	±1.1	±1.3	±1.6	±2.2	±3.4
SN nombre	98.1	±0.8	±1.0	±1.2	±1.6	±2.4
coréf. genre	89.7	±1.9	±2.2	±2.7	±3.7	±5.9

TABLE 7 – Significativité des mesures avec le système Nematus + rétro-traduction

rapport au système statistique.

Comme l'on pouvait s'y attendre, plus la fréquence augmente et plus la performance est élevée. Le nombre de phrases comportant des fréquences inférieures à 50 est généralement réduit, toutefois, nous constatons que le système statistique à tendance à mieux gérer ces mots rares : passé, comparatif, noms plur. et v. tps/mode présentent dans ce cas une plus grande précision. En revanche, au-delà de 50 occurrences, strictement toutes les précisions du système neuronal sont supérieures. Ce résultat tend à montrer que le problème des mots rares reste important en traduction neuronale, et leur segmentation en unités plus fréquentes ne garantit pas une bonne traduction.

4.4 Une évaluation de la métrique

4.4.1 Significativité des scores

Chaque tâche introduite jusqu'ici suppose une précision (ou une entropie) calculée sur la base de 500 paires (ou groupes) de phrases. Pour mesurer l'impact de la quantité d'exemples sur les mesures de qualité, nous estimeons la significativité des scores obtenus selon différentes tailles du jeu de tests.

Nous avons constitué un nouveau jeu de test indépendant du premier et comportant 1000 exemples par tâche, qui a été traduit avec le système neuronal bénéficiant des données rétro-traduites (tableau 7). Sont considérés différents sous-ensembles comprenant 100 à 1000 exemples issus des tests A et B. Pour chacun d'entre eux, 10 000 différents tests de la même taille ont été échantillonnés aléatoirement parmi les 1000 disponibles. Des scores moyens ont ainsi été obtenus, ainsi qu'une mesure de significativité, selon un intervalle de confiance de 95%.

Ces mesures de significativité ont été réalisées avec pour objectif de rechercher un nombre d'exemples par tâche qui satisfasse deux considérations opposées. D'une part, ce nombre doit être assez élevé pour permettre un calcul de précisions assez fin dans le but de comparer deux systèmes. D'autre part, il doit être limité pour que le jeu de test ne soit pas trop volumineux, ce qui conduirait à des temps de

décodage trop longs et peu pratiques pour les systèmes neuronaux.

Nous constatons ainsi que la différence entre les système Moses et Nematus+rétro-traduction sur la tâche du `comparatif` (respectivement 80,4 et 87,2) n'aurait pas été significative si elle avait été obtenue sur 100 exemples ($\pm$ 7,7). Sur 500 exemples, et avec variation de $\pm$ 3,5, nous sommes en mesure de distinguer les deux systèmes avec un certain degré de confiance. Nous constatons par ailleurs que tous deux ne montrent pas de différence significative ($\pm$ 1,4) avec 500 exemples pour la tâche de `v. nombre` (respectivement 97,5 et 97,8). Notons qu'une telle différence n'aurait pas été plus pertinente sur 1000 exemples ($\pm$ 1,4). Ainsi, notre choix de sélectionner 500 exemples permet d'avoir des tests significatifs et relativement rapides à mettre en œuvre.

4.4.2 Évaluation qualitative

Nous présentons ici quelques exemples concrets de traduction et de leur évaluation, telle qu'elle s'opère automatiquement dans le cadre du protocole présenté.

La tâche `SN nombre` consiste à changer dans la source un pronom en syntagme nominal et de vérifier si un système modélise correctement l'accord entre l'adjectif et le nom.

source	I don't want to kill **you**. I don't want to kill the **impartial compositors**.	
Moses	Je ne veux pas vous tuer. Je ne veux pas tuer les **clavistes impartial**.	✗
Nematus	je ne veux pas tuer . Je ne veux pas tuer les **compositeurs impartiaux**.	✓
+ rétro-trad.	Je ne veux pas vous tuer. Je ne veux pas tuer les **compositeurs impartiaux**.	✓

Le mot *compositors* est peu fréquent dans les données d'entraînement, si bien que le modèle de langue du système Moses n'a observé ni *clavistes impartial*, ni *clavistes impartiaux*, bien que les deux formes de l'adjectif soient présentes dans les données. En revanche, le système neuronal bénéficie d'une plus grande généralisation en segmentant l'adjectif en *imparti- aux*, ce qui distingue une terminaison univoque du pluriel.

Il arrive toutefois que certaines erreurs syntaxiques du système Moses ne puissent pas être repérées dans notre protocole. C'est ce que l'on observe dans cet exemple de la tâche `passé`, où la variante générée est une mauvaise traduction, mais où le test est néanmoins réussi, puisqu'un verbe au passé a été automatiquement détecté.

source	That **prompts** Tara to ask when she can eat. That **prompted** Tara to ask when she can eat.	
Moses	Tara qui se demander quand elle peut manger. Tara, qui **ont conduit** à se demander si elle peut manger.	✓
Nematus	Cela amène Tara à se demander quand elle peut manger. Cela **a incité** Tara à se demander quand elle peut manger.	✓
+ rétro-trad.	Cela incite Tara à se demander quand elle peut manger. Cela **a poussé** Tara à se demander quand elle peut manger.	✓

Dans le cas où la tâche `SN nombre` présente un mot inconnu dans les données,(ici *signallers*), notre protocole permet d'écarter facilement l'hypothèse fournie par Moses qui ne fait qu'une copie. Dans

ces cas, les systèmes neuronaux sont capables de générer des phrases parfaitement fluides, quoique peu fidèles à la source. Notre protocole consiste ici à repérer dans la variante deux mots qui sont absents de la base et à vérifier leur accord : ainsi, *signaux truqués* et *messages truqués* reflètent bien l'accord voulu, et nous ne sommes pas en mesure de juger la qualité de la traduction.

source	What more do **you** need to say ? What more do the **truthful signallers** need to say ?	
Moses	Plus ce que vous voulez dire ? Ce qu' il faut faire la **vérité signallers** à dire ?	✗
Nematus	Qu'avez-vous besoin de dire ? Qu'en est-il des **signaux truqués** ?	✓
+ rétro-trad.	Qu'est-ce que vous devez dire ? Quels sont les **messages truqués** qu'il faut dire ?	✓

Des erreurs peuvent par ailleurs provenir de la génération du test. Dans cet exemple de la tâche `futur`, l'analyseur en parties du discours a interprété *call* dans la base comme un verbe. La variante qui en résulte est donc agrammaticale et aucune traduction correcte ne peut être attendue.

source	Telephone **calls** to Khan and Kearney were not immediately returned. Telephone **will call** to Khan and Kearney were not immediately returned.	
Moses	Les appels téléphoniques à Khan et Kearney n'étaient pas immédiatement retourné. Téléphone fera appel à Khan et Kearney n'étaient pas immédiatement retourné.	✓
Nematus	Les appels téléphoniques à Khan et à Kearney n'ont pas été immédiatement retournés. Le téléphone fera appel à Khan et à Kearney.	✓
+ rétro-trad.	Les appels téléphoniques à Khan et à Kearney n'ont pas été immédiatement renvoyés. Téléphoner à Khan et à Kearney n'a pas été immédiatement retourné.	✗

Une fois encore, nous observons la grande fluidité des traductions neuronales, même lorsque les systèmes ne parviennent pas à interpréter la source correctement. En effet, le système statistique se montre incapable de traduire une séquence de deux verbes d'état, ce que la traduction neuronale résout dans la tâche `v. tps/mode`.

source	Our responsibility **lies** in communicating this information ! Our responsibility **rests and always lies** in communicating this information !	
Moses	Notre responsabilité est de communiquer cette information ! Notre responsabilité **est toujours et communiquer** cette information !	✗
Nematus	Notre responsabilité réside dans la communication de cette information ! Notre responsabilité **repose et réside toujours** dans la communication de cette information !	✓
+ rétro-trad.	Notre responsabilité réside dans la communication de cette information ! Notre responsabilité **repose et réside toujours** dans la communication de cette information !	✓

Une tâche semble toutefois rester difficile pour tous les systèmes : la `coréférence`. Ici, le système statistique omet tout simplement les pronoms. Les modèles neuronaux produisent bien un pronom, mais le système +rétro-trad semble choisir *administration* comme antécédent du pronom, ce qui le conduit à prédire le mauvais genre. Quant au système Nematus, il génère un pronom correct pour la base et un pronom ambigu du point de vue de la tâche : *l'* pouvant être à la fois masculin et féminin, il est toujours considéré comme correct.

source	The Bush administration should support the UN **process** and not undermine it. The Bush administration should support the UN **effort** and not undermine it.	
Moses	L'administration Bush doit soutenir le **processus** de l'ONU et ne pas saper. L'administration Bush doit appuyer les **efforts** de l'ONU et ne pas saper.	✗ ✗
Nematus	L'administration Bush devrait soutenir le **processus** de l'ONU et ne pas **le** saper. L'administration Bush devrait soutenir l'**effort** de l'ONU et ne pas **l'**affaiblir.	✓ ✓
+ rétro-trad.	L'administration Bush devrait soutenir le **processus** des Nations unies et ne pas **la** saper. L'administration Bush devrait soutenir l'**effort** des Nations unies et ne pas **la** saper.	✗ ✗

Cette évaluation permet également de distinguer les deux systèmes neuronaux. Dans la tâche `conditionnel`, Nematus ne traduit de la source ni le sens, ni la valeur du conditionnel. Le système +rétro-trad. est lui capable de produire une traduction fidèle à la source, tout en produisant le conditionnel attendu dans la variante. Ce système bénéficie de données synthétiques (rétro-traduites automatiquement) qui sont plus littérales que les traductions humaines (Crego & Senellart, 2016), et permettent au système de mieux transférer un sens grammatical depuis la source.

source	That is what **will** keep you alive. That is what **would** keep you alive.	
Moses	C'est ce que vous **permettront** de maintenir en vie. C'est ce que vous **permettre** de maintenir en vie.	✗
Nematus	C'est ce qui **va** rester en vie. C'est ce qui **est** en vie.	✗
+ rétro-trad.	C'est ce qui vous **tiendra** en vie. C'est ce qui vous **tiendrait** en vie.	✓

5 Conclusion

Dans cet article, nous avons présenté un protocole d'évaluation de la TA depuis l'anglais vers le français spécialisé dans l'analyse de la compétence morphologique des systèmes. Contrairement aux métriques automatiques qui mettent en évidence la supériorité des systèmes neuronaux sur les systèmes statistiques, les tests présentés suggèrent que certains phénomènes grammaticaux sont moins bien modélisés dans la traduction neuronale, notamment lorsqu'il s'agit de transmettre une caractéristique morphologique depuis la source, ou lorsqu'il s'agit de traduire des mots peu fréquents.

Le protocole présenté implique la génération automatique d'un jeu de tests, au cours de laquelle certaines erreurs peuvent apparaître. Pour diminuer l'impact de ces erreurs, il est trivial d'augmenter le nombre d'exemples, dans les limites imposées par le coût de décodage ; 500 exemples par tâche semblant constituer un bon compromis. L'avantage d'une telle approche réside dans son caractère automatique, qui réduit l'intervention humaine à la conception de tâches. L'analyse est donc basée sur de nombreux exemples et permet une focalisation sur des phénomènes linguistiques précis.

Remerciements

Ce travail a été partiellement financé par le programme H2020 de l'Union Européenne dans le cadre de l'accord de subvention No. 645452 (QT21).

Références

ALLAUZEN A. & YVON F. (2011). Méthodes statistiques pour la traduction automatique. In E. GAUSSIER & F. YVON, Eds., *Modèles Probabilistes pour l'accès à l'information*, chapter 7, p. 271–356. Hermès, Paris.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, **abs/1409.0473**.

BELINKOV Y., DURRANI N., DALVI F., SAJJAD H. & GLASS J. (2017). What do neural machine translation models learn about morphology ? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 861–872.

BENTIVOGLI L., BISAZZA A., CETTOLO M. & FEDERICO M. (2016). Neural versus phrase-based machine translation quality : a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 257–267, Austin, Texas.

BURLOT F. & YVON F. (2017). Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation, Volume 1 : Research Papers*, p. 43–55, Copenhagen, Denmark.

CREGO J. M. & SENELLART J. (2016). Neural machine translation from simplified translations. *CoRR*, **abs/1612.06139**.

HEAFIELD K. (2011). KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 187–197, Edinburgh, Scotland.

ISABELLE P., CHERRY C. & FOSTER G. (2017). A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2486–2496, Copenhagen, Denmark.

KING M. & FALKEDAL K. (1990). Using test suites in evaluation of machine translation systems. In *COLNG 1990 Volume 2 : Papers presented to the 13th International Conference on Computational Linguistics*, Helsinki, Finland.

KLUBIČKA F., TORAL RUIZ A. & SÁNCHEZ-CARTAGENA V. M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. In *Proceedings of the European Conference on Machine Translation*, EAMT'17, p. 121—-132, Prague, Czech Republic.

KOEHN P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings : the tenth Machine Translation Summit*, p. 79–86, Phuket, Thailand : AAMT AAMT.

KOEHN P. (2010). *Statistical Machine Translation*. Cambridge University Press.

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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, p. 177–180, Stroudsburg, PA, USA : Association for Computational Linguistics.

LINZEN T., DUPOUX E. & GOLDBERG Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, **4**, 521–535.

LOMMEL A., BURCHARDT A., POPOVIC M., HARRIS K., AVRAMIDIS E. & USZKOREIT H. (2014). Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the conference of the European Association for Machine Translation*, EAMT 2014, Dubrovnik, Croatia.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

MILLER G. A. (1995). WordNet : A Lexical Database for English. *Communications of the ACM*, **38**(11), 39–41.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 311–318, Stroudsburg, PA, USA : Association for Computational Linguistics.

POPOVIĆ M. & NEY H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, **37**(4), 657–688.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

SENNRICH R. (2017). How grammatical is character-level neural machine translation ? assessing mt quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 376–382 : Association for Computational Linguistics.

SENNRICH R., FIRAT O., CHO K., BIRCH A., HADDOW B., HITSCHLER J., JUNCZYS-DOWMUNT M., LÄUBLI S., BARONE A. V. M., MOKRY J. & NADEJDE M. (2017). Nematus : a toolkit for neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Software Demonstrations*, p. 65–68.

SENNRICH R., HADDOW B. & BIRCH A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 86–96, Berlin, Germany.

SENNRICH R., HADDOW B. & BIRCH A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1715–1725.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, p. 3104–3112, Cambridge, MA, USA : MIT Press.

TORAL RUIZ A. & SÁNCHEZ-CARTAGENA M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 1063–1073, Valencia, Spain : Association for Computational Linguistics (ACL).

VANIA C. & LOPEZ A. (2017). From characters to words to in between : Do we capture morphology ? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2016–2027 : Association for Computational Linguistics.

VILAR D., XU J., LUIS FERNANDO D. & NEY H. (2006). Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC'06, Genoa, Italy.