Comparaison de systèmes de traduction automatique, probabiliste et neuronal, par analyse d'erreurs.

Emmanuelle Esperança-Rodier[1], Nicolas Becker [2]

(1) Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
(2) Univ. Grenoble Alpes, 38000 Grenoble, France
Emmanuelle.Esperanca-Rodier@univ-grenoble-alpes.fr,
Nicolas.Becker@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article présente les travaux d'analyse d'erreurs de 2 systèmes de TA maison, l'un probabiliste et l'autre neuronal. Après une description du corpus et des systèmes, nous analysons les deux systèmes en fonction d'une typologie d'erreurs en nous arrêtant sur quelques exemples de phrases pour lesquelles les deux systèmes ont effectué le même type d'erreurs.

ABSTRACT

Comparison of an SMT system and an NMT system thanks to error translation annotations.
This article presents the error analysis work of 2 in-house Machine Translation systems, one probabilistic and the other one neural. After a description of the corpus and the systems, we analyse the two systems according to a typology of errors by focusing on some examples from sentences for which the two systems made the same types of errors.

MOTS-CLES : analyse d'erreurs, qualité de traduction, évaluation de systèmes de TA.
KEYWORDS: error analysis, translation quality, MT system evaluation.

1. Introduction

La comparaison de systèmes de traduction automatique (TA) probabiliste et neuronal est désormais monnaie courante. On trouve, d'une part, les travaux d'évaluation à petite échelle et d'autre part ceux à grande échelle.

Parmi les travaux étudiant de petits corpus, nous remarquons les travaux de Bentivogli et al., (2016). Ces derniers évaluent la qualité avec la métrique HTER (Snover et al., 2016) d'un système de TA probabiliste et de quatre systèmes de TA neuronaux. Les travaux de Wu et al. (2016) effectuent une évaluation grâce à la métrique BLEU (Papineni, et al., 2002). Esperança-Rodier et al. (2016) évaluent de manière quantitative, en utilisant les métriques BLEU, Translation Edit Rate (TER) (Snover et al., 2006), Meteor-E (Servan et

* Institute of Engineering Univ. Grenoble Alpes

al., 2016), mais aussi qualitative (évaluation humaine de la qualité et de la perception) un système de TA probabiliste et un système de TA neuronal.

Du côté des travaux utilisant de grands corpus, plusieurs systèmes de TA des deux types, et plusieurs couples de langues ont été évaluées lors de la campagne WMT16 (Bojar et al., 2016) ou au cours des travaux de Toral et al., (2017). Castilho et al. (2017) combinent quant à eux, l'évaluation par des méthodes automatiques et l'évaluation par des humains (adéquation, fluidité). Tout aussi récemment, Isabelle et al. (2017) évaluent des systèmes de TA neuronaux et probabilistes sur des phrases créées en fonction de difficultés linguistiques.

La grande majorité de ces travaux démontre l'amélioration de la qualité de la traduction par les systèmes de TA neuronaux en fonction de certains critères que nous expliciterons dans la section 2 de cet article. Notre travail se positionne dans la comparaison d'annotation de type d'erreurs de données de traduction du domaine du tourisme, à petite échelle, soient 310 couples en langue source FR-langue cible GB pour chaque système de TA maison, l'un probabiliste et l'autre neuronal, en fonction de la typologie d'erreurs de traduction de Vilar et al. (2006). Après avoir présenté les résultats des principaux autres travaux d'évaluation de la qualité dans l'état de l'art en section 2, nous présenterons le corpus utilisé en section 3. Puis nous décrirons les deux systèmes de TA maison dans la section 4. Nous continuerons par l'explication de la méthodologie suivie en détaillant la typologie d'erreurs de traduction de Vilar ainsi que l'outil d'évaluation utilisé, dans la section 5. Les résultats généraux ainsi que quelques exemples choisis seront donnés dans la section 6 avant de conclure en section 7.

2. État de l'art

Nous partageons ici une partie de l'état de l'art présenté par Castilho et al. (2017) tout en y ajoutant d'autres recherches. Les travaux faisant état de la comparaison des systèmes de TA probabilistes et neuronaux sont de plus en plus nombreux. Ainsi, les travaux de Bentivogli et al., (2016), portant sur un peu plus d'une centaine de phrases, ont permis de comparer un système probabiliste avec quatre systèmes neuronaux en évaluant le taux d'effort de post-édition qui s'est vu réduit de 26% grâce aux systèmes neuronaux, montrant ainsi moins d'erreurs d'ordre des mots notamment sur le positionnement des verbes, ainsi que moins d'erreurs du point de vue lexical et morphologique.

Wu et al. (2016), quant à eux, ont évalué cinq cents phrases sur 3 couples de langues dans les deux sens, par la métrique BLEU (Papineni, et al., 2002), et ont montré que les systèmes de TA neuronaux étaient plus performants pour les langages morphologiquement riches.

Nos précédents travaux (Esperança-Rodier et al., 2016) évaluant de manière quantitative, en utilisant les métriques BLEU, Translation Edit Rate (TER), Meteor-E, ainsi que qualitative (évaluation humaine de la qualité et de la perception) un système de TA probabiliste et un système de TA neuronal maison, sur un couple de langue et une cinquantaine de phrases sélectionnées dans le domaine du tourisme, ont montrés que les deux systèmes étaient perçus par les traducteurs de manière équivalente bien qu'une légère préférence ait été donnée au système de TA neuronal.

La campagne WMT16 (Bojar et al., 2016) portant sur 12 couples de langues a démontré que pour 6 couples sur 12 les systèmes neuronaux donnaient de meilleurs résultats que les autres systèmes pour la tâche de traduction, et qu'ils amélioraient grandement la réalisation pour la tâche de post-édition.

En ce qui concerne les travaux plus récents, trois se dénotent du contexte en émettant des réserves sur l'ampleur de l'amélioration apportée par les systèmes neuronaux. Le premier étant l'étude de Toral et al. (2017) sur 9 couples de langues, comparant des systèmes neuronaux et probabilistes en mettant en valeur le fait que les systèmes neuronaux ont obtenu de meilleurs scores BLEU que les systèmes probabilistes. Toral et al. (2017) mitigent ces résultats en précisant que les systèmes probabilistes obtiennent de meilleurs scores en comparaison des systèmes neuronaux, sur des phrases de plus de 40 mots. Cette étude a corroboré les travaux de Bentivogli et al. (2016) en montrant que les traductions issues de systèmes de TA neuronaux avaient obtenu un meilleur score de fluidité que les traductions issues de systèmes probabilistes.

Constat appuyé par les résultats de Castilho et al. (2017) qui ont également indiqué que les traductions issues des systèmes neuronaux nécessitaient moins de post-édition, bien que relativisant la qualité de ces traductions en soulignant les nombreuses erreurs d'omissions et de mauvaises traductions.

Enfin, les travaux d'Isabelle et al. (2017) se singularisent par la création d'un ensemble de phrases comportant des difficultés linguistiques sélectionnées pour leur mauvais traitement en TA. Cette étude démontre également que les systèmes neuronaux obtiennent de moins bons scores lorsqu'il s'agit de traduire des expressions idiomatiques.

Le travail décrit dans cet article s'inspire de ces évaluations bien qu'analysant les erreurs de traductions et non pas l'effort de post-édition. Nous utilisons un petit échantillonnage de phrases, environ 300 dans le domaine du tourisme, décrit dans la prochaine section, pour deux systèmes de TA l'un neuronal et l'autre probabiliste, présentés en section 4, sur un seul couple de langues français-anglais.

3. Corpus

Nous avons sélectionné le corpus BTEC (Basic Travel Expression Corpus) qui, comme décrit dans la tâche BTEC de la campagne d'évaluation IWSLT 2010[1], est un corpus linguistique multilingue contenant des phrases liées au domaine du tourisme et similaires à celles qui se trouvent généralement dans les guides de conversation pour les touristes partant à l'étranger. Le BTEC est composé des phrases courtes (10 mots / phrase en moyenne).

BTEC	French Words	English Words
train	201k	189k
dev	12,2k	11,5k
test1	3,9k	3,6k
test2	3,8k	3,6k

TABLE 1 : Corpus BTEC en nombre de mots

[1] http://iwslt2010.fbk.eu/

Nous avons travaillé sur les traductions de BTEC test1, qui représentent, comme indiqué dans la table 1, 3,9k mots pour le français et 3,6k pour l'anglais, soit 469 segments alignés.

4. Systèmes de TA maison

Nous avons choisi de comparer deux systèmes de TA maison. Nous décrirons le système de TA neuronal développé maison Seq2Seq, dans la section 4.2 après avoir succinctement décrit, en section 4.1, le système de TA probabiliste maison LIG-Moses.

Corpus BTEC	LIG-Moses (BLEU/TER/Méteor 1.5)	Seq2Seq (BLEU/TER/Méteor 1.5)
dev	54,35/28,66/43,70	51,56/30,75/41,09
test1	49,44/32,20/42,26	47,07/33,16/40,17
test2	45,75/35,08/40,82	47,88/33,09/40,10

TABLE 2 : Scores monoréférence (BLEU/TER/Meteor 1.5)

Comme le montre la table 2 ci-dessus, malgré un corpus d'entrainement relativement petit, le système Seq2Seq donne, en ce qui concerne le score BLEU, des résultats comparables à un système probabiliste état de l'art.

4.1 LIG-Moses

La base de référence de notre système de traduction probabiliste, nommé LIG-Moses, que nous utilisons pour cette évaluation, est un modèle utilisant Moses Toolkit (Koehn et al., 2007), entrainé sur le corpus d'entrainement BTEC train. Ce corpus d'entrainement représente 201k mots pour le français et 189k mots pour l'anglais, mais il ne prend pas en compte toutes les données monolingues ajoutées. Le système LIG-Moses est optimisé sur le corpus de développement BTEC dev représente 12,2k mots pour le français, et 11,5k pour l'anglais.

4.2 Seq2Seq

Notre modèle de TA neuronal est un réseau neuronal encodeur-décodeur utilisant un mécanisme d'attention (Sutskever et al., 2014, Bahdanau et al., 2015). La mise en œuvre du LIG, décrite dans Bérard et al. (2016) est fondée sur le modèle Seq2Seq implémenté avec TensorFlow (Abadi et al., 2015). Il réutilise certaines de ses composantes, tout en ajoutant un certain nombre de fonctionnalités, comme un encodeur bidirectionnel (Bahdanau et al., 2015), un décodeur de recherche de faisceaux, un mécanisme d'attention et un encodeur hiérarchique (Chorowski et al., 2015). Le modèle de TA neuronal utilise un encodeur-décodeur fondé sur un modèle bi-LSTM avec 2 couches cachées de 256 neurones, avec des représentations vectorielles de mots (embeddings) de taille 256. Nous utilisons un modèle d'attention standard. Pour l'entrainement, nous utilisons l'algorithme Adam avec un taux d'apprentissage initial de 0.001 (Kingma et Ba, 2015), et une taille de mini-lot de 64. Nous appliquons un "dropout" (avec un taux de 0.5) pendant l'entraînement sur les connexions entre les couches LSTM dans l'encodeur

et le décodeur (Zaremba et al., 2014). L'entrainement est réalisé avec le même corpus d'entrainement BTEC train que celui utilisé pour LIG-Moses.

5. Méthodologie

Dans le cadre d'un projet, Pôle Grenoble Cognition, un étudiant de Master 2 de Traductologie à l'université de Grenoble Alpes, a effectué l'annotation des erreurs de traduction.

Afin de mener à bien cette tâche, nous avons fourni en entrée les 469 segments du BTEC test1, à chaque système de TA maison. Nous avons ainsi obtenu 469 hypothèses de traduction pour le système de TA probabiliste (notées ci-après SMT) et 469 hypothèses de traduction pour le système de TA neuronal (notées ci-après NMT).

L'étudiant a utilisé la plateforme collaborative d'annotation d'erreurs de traduction ACCOLÉ (Brunet-Manquat et Esperança-Rodier, 2018). Cette plateforme en ligne permet en plus de l'annotation selon la typologie d'erreurs de traduction de Vilar et al. (2006) décrite ci-après, d'effectuer des recherches d'erreurs sur les corpus annotés. L'étudiant a annoté les erreurs de traduction des hypothèses de traduction SMT, ainsi qu'en parallèle, les erreurs de traduction des hypothèses de traduction NMT. Durant sa période de stage, il a annoté 310 segments alignés pour le SMT et 310 segments alignés pour le NMT. Ce sont les annotations sur ces 310 segments pour les deux systèmes de TA, que nous avons analysées pour cet article.

La typologie d'erreurs de traduction de Vilar, décrite dans la figure 1, identifie cinq catégories principales d'erreurs de traduction à savoir :

- Mots manquants, pour les mots qui n'ont pas été traduits,
- Ordre des mots, pour un mauvais ordre des mots dans la séquence traduite,
- Mots incorrects, pour une erreur de traduction,
- Mots inconnus, pour les mots restés dans la langue source,
- Ponctuation, lorsque les règles de ponctuation de la langue cible n'étaient pas respectées.

En ce qui concerne les deux premières catégories d'erreurs de traduction, les mots manquants et l'ordre des mots, des sous-catégories ont été créées pour affiner la classe d'erreur.

Pour la catégorie d'erreur Mots manquants, la distinction entre les mots signifiants et les mots outils permet de voir si le mot manquant était significatif ou non. Cette sous-catégorie illustre le fait que le sens complet de la phrase a été conservé ou non, ce qui est évidemment l'un des buts d'une évaluation de la qualité de traduction. En ce qui concerne la catégorie d'erreur Ordre des Mots, la sous-catégorie Mot ou Syntagme indique si l'erreur de traduction entraîne une réorganisation des mots eux-mêmes ou un réordonnancement des segments dans la phrase. Il permet de localiser à quel niveau, lexical ou sémantique, le système a échoué.

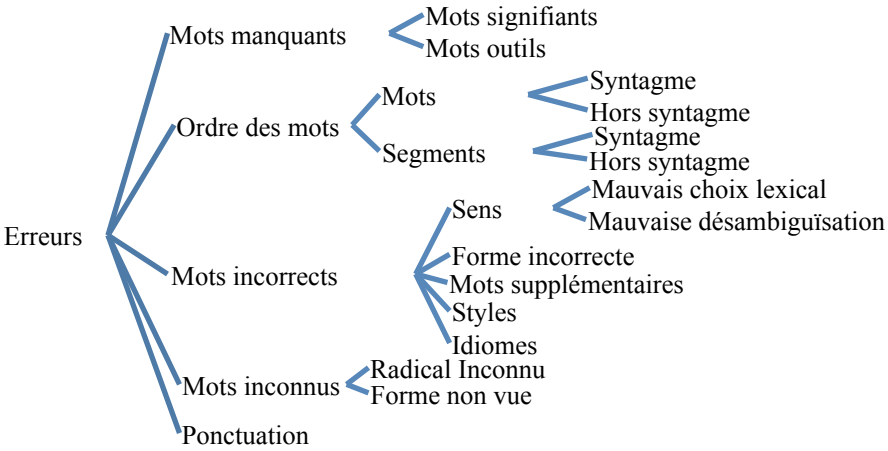

FIGURE 1 : Typologie d'erreurs de Vilar et al. (2006)

En regardant la troisième catégorie d'erreur de traduction, Mots incorrects, nous pouvons voir qu'il existe plusieurs sous-catégories visant à distinguer la raison de l'erreur de traduction, ce qui peut être dû au fait que le système n'a pas été capable de désambiguïser correctement le sens d'un mot source ou de produire la bonne forme du mot, bien que le lemme du mot ait été bien traduit.

Pour la quatrième catégorie d'erreur de traduction, Mots inconnus, on peut distinguer si le lemme des mots était connu par le système ou non.

Et enfin, la cinquième catégorie d'erreur de traduction, qui est Ponctuation, n'a pas reçu toute notre attention.

6. Résultats

6.1 Analyse générale

Une première analyse indique que certaines hypothèses de traduction issues des deux systèmes de TA ont été annotées avec le même type d'erreurs, c'est-à-dire que pour la même phrase source chaque hypothèse de traduction comportait un segment qui avait été mal traduit pour la même raison. Nous avons donc fait figurer dans la figure 2 ci-après, le nombre de phrases communes aux 2 systèmes pour lesquelles le même type d'erreurs a été annoté. La figure 2 montre pour chaque type d'erreurs de traduction de la typologie de Vilar et al. (2006) :

- le nombre de phrases annotées pour au moins une erreur de ce type uniquement pour le système SMT, noté SMT-C.
- Le nombre de phrases annotées pour au moins une erreur de ce type, pour le système SMT et pour le système NMT, noté Communs (C).
- Le nombre de phrases annotées pour une erreur de ce type uniquement pour le système NMT, noté NMT-C.

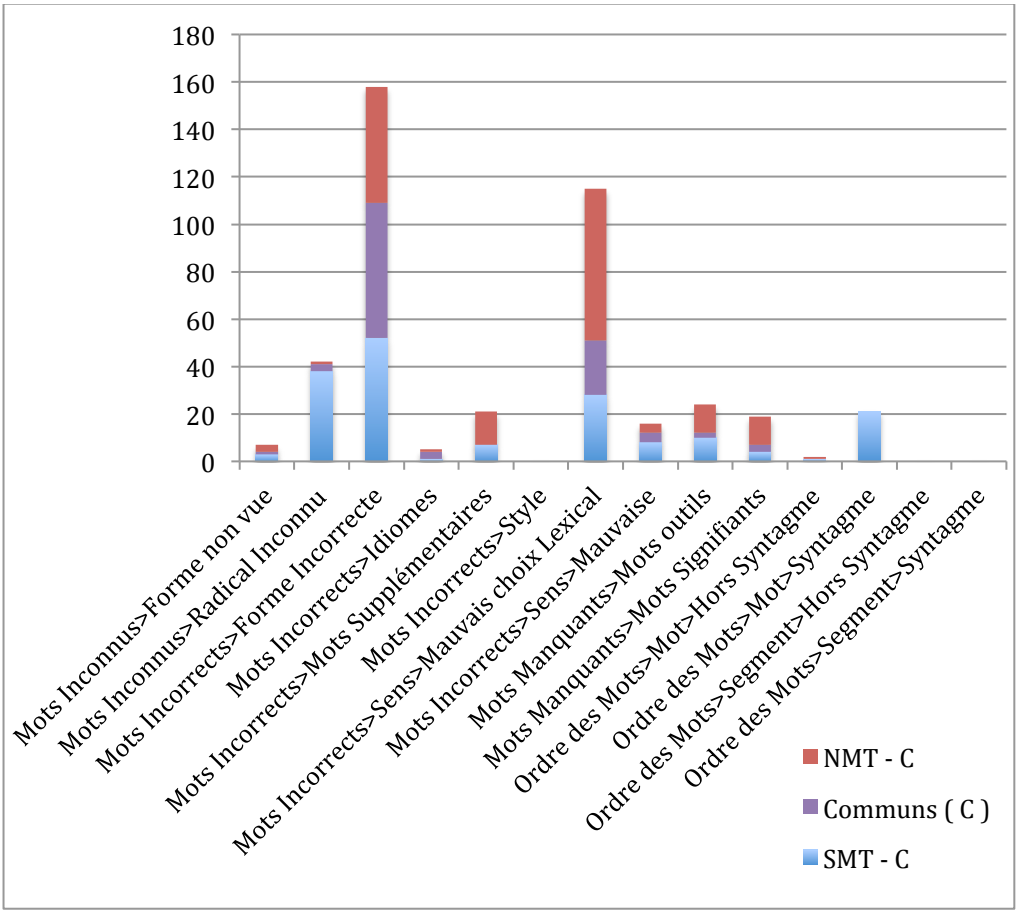

FIGURE 2 : Nombre de phrases annotées pour au moins un type d'erreurs pour chaque système de TA avec indication des phrases communes aux 2 systèmes

Nous remarquons ensuite qu'aucune erreur de Style, ni d'Ordre des Mots au niveau du Segment, à l'extérieur ou à l'intérieur d'un Syntagme, n'a été annotée pour aucun des systèmes. En contraste, les annotations du type d'erreurs Mots Incorrects, Forme Incorrecte ainsi que du type Mots Incorrects, Sens, Mauvais Choix Lexical sont les plus nombreuses pour les deux systèmes. Nous rappelons que ces types d'erreurs permettent d'indiquer que le sens de la phrase source n'a pas été reproduit dans l'hypothèse de traduction (notion d'adéquation).

Pour le type d'erreurs Mots Incorrects, Forme Incorrecte, 52 phrases ont été annotées pour le SMT, 57 phrases annotées sont communes aux deux systèmes et 49 phrases ont été annotées pour le NMT. Il en va de même pour le type Mots Incorrects, Sens, Mauvais Choix Lexical, 28 phrases annotées pour le SMT, 23 phrases communes, et 64 phrases pour le NMT. Notons pour ce dernier type une prépondérance d'erreurs pour le NMT (64 phrases contre 28 pour le SMT). Ainsi le NMT montre une plus mauvaise adéquation de

ses hypothèses de traductions que le SMT, ce qui rejoint les travaux de Castilho et al. (2017).

Aucune annotation n'a été effectuée sur des phrases communes aux deux systèmes pour les types d'erreurs suivants :

- Mots Incorrects, Mots Supplémentaires,
- Ordre des mots, Mot, Hors Syntagme,
- Ordre des mots, Mot, Syntagme.

Pour le type Mots Incorrects, Mots Supplémentaires, 7 phrases ont été annotées pour le SMT, et 14 l'ont été pour le NMT, soit le double. Les erreurs annotées, soulignées dans la table 3 ci-dessous, pour le SMT portent pour 3 phrases sur l'ajout d'un même segment non présent dans la source, comme le montre l'exemple 1 de la table 3, et pour 1 sur l'ajout d'un adjectif (exemple 2). Enfin, pour 3 autres phrases, un élément grammatical a été ajouté (exemple 3). Cela peut s'expliquer par le traitement d'apprentissage statistique du SMT, pour lequel la traduction I really liked it a été plus fréquemment rencontrée pour la traduction de J'ai bien aimé.

Pour le NMT, il s'agit pour moitié, soit pour 7 phrases, d'ajout de termes (exemple 4) et pour l'autre moitié, de bégaiement (exemple 5). Ceci est inhérent à la jeunesse du système, mais surtout au fonctionnement interne du système NMT qui utilise un décodeur de recherche de faisceaux afin d'agrandir l'espace de traduction. Lorsqu'il ne peut trouver la meilleure solution, le système continue son traitement et produit une mauvaise traduction.

#	Source	SMT	NMT
1	au secours !	help something like	help .
2	où se trouve l' ambassade japonaise ?	where is the japanese embassy japanese ?	where is the japanese embassy ?
3	j' ai bien aimé le film . merci de m' avoir invité .	i really liked it the movie . thank you for inviting me .	i really liked the movie . thank you for inviting me .
4	qui est en train de parler , s' il vous plaît ?	who 's talking , please ?	who 's speaking with you , please ?
5	je serai là dans environ cinq minutes .	i 'll be there in about fifteen minutes .	i 'll be there in about about five minutes .

TABLE 3 : Exemples de phrases du SMT ou du NMT pour lesquelles une annotation d'erreurs du type Mots Incorrects, Mots Supplémentaires a été associée

Nous notons également que le SMT effectue plus d'erreurs de type Mots Inconnus, Radical Non Vu, 38 phrases plus 3 phrases communes aux deux systèmes contre une seule phrase pour le NMT et les 3 phrases communes. Pour les 38 phrases où le SMT a fourni une mauvaise traduction, il a copié le mot source dans la cible comme le montre les exemples sélectionnés 6, 7, 8 et 9 dans la table 4 ci-dessous. Il est à noter que pour cet exemple, le NMT a produit une mauvaise traduction, mais annotée avec un autre type d'erreur. Pour les exemples 7, 8 et 9, les deux systèmes ont commis le même type d'erreurs de traduction sur le même segment. Ceci s'explique par le fait que les systèmes probabilistes sont plus enclins à

reproduire pour traduction un mot de la phrase source lorsqu'il ne le reconnaît pas. Le système NMT produit soit une traduction hallucinée, qui n'a rien à voir avec la phrase source (exemples 6 et 9) ou bien un bégaiement (exemples 7, 8 et 10). Ce qui s'explique, comme dit plus haut, par l'utilisation d'un décodeur de recherche de faisceaux.

#	Source	SMT	NMT
6	je n' arrive pas à trouver la bonne traduction .	i can 't find the good traduction .	i can 't find the wrong change .
7	je voudrais manger de la vraie nourriture indienne .	i 'd like to have true food indienne .	i 'd like to eat some food food .
8	je voudrais une voiture non fumeuse .	i 'd like a car no fumeuse .	i 'd like a car car .
9	bonjour , je m' appelle junko . je travaille pour le département des relations publiques . je suis nouveau ici .	hello , my name is junko . i work for the département any relationships publiques . i 'm new here .	hello , my name is oda . i 'm just looking for the lost-and-found . i 'm new here .
10	avez-vous des petites voitures ?	do you have any small cars ?	do you have any cars cars ?

TABLE 4 : Exemples de phrases du SMT ou du NMT pour lesquelles une annotation d'erreurs du type Mots Inconnus, Radical Non Vu a été associée

Il en est de même, mais dans une moindre mesure pour le type d'erreurs Ordre des Mots, Mot, Syntagme avec 21 phrases pour le SMT et aucune en commun et aucune pour le NMT. Ce qui rejoint les résultats de Bentivogli et al. (2016) en montrant une meilleure fluidité du système neuronal. Principalement, les erreurs de traduction du SMT sont dues au fait que la position de l'adjectif n'est pas respectée dans le syntagme nominal, exemple 11 de la table 5, ou bien que le mot source reproduit dans la traduction (type d'erreurs Mots Inconnus, Radical Non Vu) est resté à la position qu'il occupait dans la source, exemple 12. Il est à remarquer que les traductions du NMT comportent des erreurs d'autres types pour les mêmes segments sources.

#	Source	SMT	NMT
11	quel genre d' excursions de nuit avez-vous ?	what kind of tours night do you have ?	what kind of excursion cruises do you have ?
12	y a-t-il une visite guidée ?	is there a tour guidée ?	is there a tour ?

TABLE 5 : Exemples de phrases du SMT pour lesquelles une annotation d'erreurs du type Ordre des Mots, Mots, Syntagme a été associée

L'analyse des deux systèmes a permis de trouver une ou plusieurs erreurs dans 275 phrases pour le SMT et 258 phrases pour le NMT. Ce qui représente 178 phrases comportant au moins une erreur pour le SMT, 97 phrases communes aux deux systèmes, et 161 phrases comportant au moins une erreur pour le NMT. Ainsi ces résultats

corroborent nos premiers travaux (Esperança-Rodier et al., 2017) en montrant une équivalence des résultats pour le NMT et le SMT, tout en indiquant une plus mauvaise adéquation pour le NMT mais une meilleure fluidité.

Nous allons donc nous intéresser aux phrases communes annotées pour un même type d'erreurs de TA en présentant dans la section suivante quelques exemples choisis.

6.2 Analyse des phrases communes aux deux systèmes pour les deux types d'erreurs prépondérants

Pour le type d'erreurs, Mots Incorrects, Forme Incorrecte, nous avons pu repérer, parmi les 57 phrases communes au SMT et NMT, c'est-à-dire pour lesquelles les deux systèmes n'ont pas su traduire correctement un segment, sept grandes catégories illustrées par les exemples contenus dans la table 5 ci-dessous.

#	Source	SMT	NMT
13	combien de temps devrons-nous attendre ?	how long should we wait ?	how long should we wait ?
14	non , mais j' ai envie de vomir . je pense que j' ai trop mangé hier soir .	no , but i feel like vomiting . i think i 've eaten too last night .	no , but i feel like vomiting . i think i 've got drunk last night .
15	pouvez-vous garder mes sacs ici jusqu' à cinq heure ?	can you keep my bags here until five hour ?	can you hold my bags here until five time ?
16	la population totale du japon est d' environ cent trente millions d' habitants .	the total population of japan is about one hundred thirty million .	the total population of japan is about one hundred thirty million .
17	il fait moins trois degrés centigrade .	it 's least three degrees centigrade .	it 's about three days .
18	pourriez-vous m' appeler un taxi ?	could you call a taxi ?	could you call a taxi for me ?
19	ce pourcentage est beaucoup plus élevé que les chiffres des précédentes études pour seize autres produits .	this pourcentage is much more high the figures précédentes a college for four other products .	this product is much more than the most popular items for new people .

TABLE 5 : Exemples de phrases communes au SMT et au NMT pour lesquelles une annotation d'erreurs du type Mots Incorrects, Forme Incorrecte a été associée

Dans l'exemple 13, les deux systèmes fournissent la même traduction erronée (en souligné dans la table), portant sur la catégorie temps du verbe, alors que dans l'exemple 14, pour la même catégorie, en plus de commettre la même erreur, l'hypothèse de traduction du NMT change le sens de la phrase (utilisation de "drunk"/saôul au lieu de "eaten"/mangé), et bien que la traduction du NMT soit tout à fait fluide, le sens de cette dernière n'a rien à voir avec le sens de la phrase source (mauvaise adéquation).

L'exemple 15, concernant la catégorie expression de l'heure, montre qu'aucun des deux systèmes ne traduit correctement l'expression de l'heure (utilisation de "hour" pour le SMT et "time" pour le NMT à la place de "o'clock").

La troisième catégorie qui est l'expression de nombres, montre que les deux systèmes produisent la même erreur (oubli de "and" après les centaines dans l'exemple 16).

L'exemple 17 montre une erreur de la catégorie expression des températures pour lesquelles les deux systèmes ne fournissent pas une traduction adaptée, et indique à nouveau que l'adéquation de la traduction du NMT n'est pas bonne.

La catégorie verbe pronominal illustrée par l'exemple 18 montre que les deux systèmes commettent le même type d'erreurs bien que les traductions soient différentes.

Enfin, l'exemple 19 indique que les deux systèmes ne savent pas traduire les comparatifs, de supériorité dans ce cas. Ces erreurs sont plus difficiles à expliquer et une recherche sur les patrons morphosyntaxiques correspondant à ces catégories à travers le corpus permettrait de trouver une raison.

La table 6 ci-dessous comporte des exemples représentatifs des erreurs du type Mots Incorrects, Sens, Mauvais Choix Lexical.

#	Source	SMT	NMT
20	les adolescents japonais aiment les jeux vidéos .	the adolescents japanese love electronic vidéos .	japanese teenagers are interested in fashion .
21	aurons-nous du temps libre pendant le voyage ?	will we spare time for the trip ?	do we have time time during the trip ?
22	quelle boîte de nuit nous plairait ?	what night we like this ?	what kind of night would you like ?

TABLE 6 : Exemples de phrases communes au SMT et NMT pour lesquelles une annotation d'erreurs du type Mots Incorrects, Sens, Mauvais Choix Lexical a été associée

L'exemple 20 montre que le SMT n'a pas traduit correctement le terme jeux vidéos, en commettant deux erreurs, un mauvais choix lexical plus une non traduction puisque vidéos est laissé tel quel. L'hypothèse de traduction du NMT quant à elle, ne respecte pas le sens de la phrase source, donc pas en adéquation avec la source, bien que tout à fait correcte grammaticalement et donc fluide. Ceci corrobore le fait que le NMT est réputé plus fluide mais moins en adéquation avec la source.

L'exemple 21 montre encore une tendance au bégaiement du NMT, alors que le SMT ne fournit pas la traduction correspondant au sens de la phrase source et surtout n'est pas fluide puisque le verbe est absent. Nous avons fréquemment relevé ce comportement (bégaiement) du NMT pour d'autres types d'erreurs que nous ne traitons pas dans cet article. Comportement qui, comme nous l'avons déjà vu, est inhérent au fonctionnement du NMT.

Enfin, dans l'exemple 22, nous voyons que les deux systèmes produisent deux hypothèses de traduction très différentes, mais avec la même erreur concernant la traduction de l'expression idiomatique "boite de nuit" ce qui rejoint les travaux précédents d'Isabelle et al. (2017). De plus, il est à noter encore une fois que l'hypothèse de traduction du NMT est plus fluide que celle du SMT.

7 Conclusion

Après un état de l'art des travaux de comparaison des systèmes de TA neuronaux et des systèmes de TA probabilistes, nous avons présenté notre étude portant sur la comparaison de deux systèmes de TA maison, l'un neuronal et l'autre probabiliste. Cette étude indique que pour le domaine du tourisme et sur des phrases relativement courtes, les deux systèmes donnent des résultats équivalents bien que le système de TA neuronal semble être légèrement plus performant en ayant un total de phrases annotées inférieur de 26 à celui pour le système de TA probabiliste.

Nous rejoignons les travaux précédents sur les systèmes de TA neuronaux, disant qu'ils offrent une meilleure fluidité tout en fournissant une moins bonne adéquation. Certains exemples montrent que, bien que l'hypothèse de traduction du système neuronal soit fausse (pas en adéquation), c'est-à-dire que le sens de la phrase source n'est pas maintenu, elle n'en est pas néanmoins fluide, en ce sens qu'elle est grammaticalement correcte.

Une étude plus approfondie sur plus de phrases mérite d'être menée afin de voir si la tendance relevée ici se révèle juste. Il est également à noter que la jeunesse de notre système de TA neuronal peut expliquer cette faible démarcation entre les deux systèmes. D'autres études seront à mener lorsqu'il sera plus mature.

Même si elle porte sur une approche linguistique, la typologie de Vilar et al. (2006) regroupe sous un même type d'erreur des occurrences lexicales, morphologiques et syntaxiques très diverses, dont le seul type d'erreur ne rend pas compte. Nous envisageons donc de travailler sur une nouvelle typologie d'erreurs se rapprochant des "challenge set" des travaux d'Isabelle et al. (2017).

Remerciements

Nous remercions le Pôle Grenoble Cognition pour le financement obtenu afin de conduire cette étude.

Références

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, Ł., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Aziz, W., de Sousa, S. C. M., Specia, L., (2012). PET: a Tool for Post-editing and Assessing Machine Translation. Dans LREC/EAMT pages 3982-3987

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*, pages 3104–3112, San Diego, California, USA.

Bentivogli, L., Bisazza, A., Cettolp, M. (2016). Neural versus Phrase-Based Machine Translation Quality: A Case Study. CoRR, abs/1608.04631

Bérard, A., Pietquin, O., Besacier, L., Servan, C. (2016) Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. *NIPS Workshop on end-to-end learning for speech and audio processing*, Dec 2016, Barcelona, Spain. 2016. ⟨hal-01408086⟩

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Find- ings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics

Brunet-Manquat, F., Esperança-Rodier, E., (2018) ACCOLÉ : Annotation Collaborative d'erreurs de traduction pour COrpus aLignÉs. TALN 2018, Démonstration, à paraître.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, Y., Lohar, P., Way, A., Miceli Barone, A. V., and Gialama, M. (2017). A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In Proceedings of Machine Translation Summit XVI, Nagoya, Japan

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 577–585, Montréal, Canada.

Esperanca-Rodier, E., Rossi, C., Berard, A., Besacier, L., (2017) Evaluation of NMT and SMT Systems: A Study on Uses and Perceptions dans *Proceedings of the 39th Conference Translating and the Computer*, pages 11–24, London, UK, November 16-17, 2017. © 2017 AsLing

Isabelle, P., Cherry, C., Foster, G., (2017). A Challenge Set Approach to Evaluating Machine Translation dans Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2486-2496, Copenhagen, Denmark, September 2017, Association for Computational Linguistics.

Kingma, D. and Ba, J. (2015). Adam: A Method for Stochastic Optimization dans International Conference on Learning Representations, May 7-9, 2015, San Diego.
Ithaca, NY: arXiv.org

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., and Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Associa- tion for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems (NIPS 2014)*, pages 3104–3112, Montréal, Canada.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, M. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*, Boston, US-MA.

Servan, C., Bérard, A., Elloumi, Z., Blanchon, H., Besacier, L., (2016). Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources? *COLING 2016*, Dec 2016, Osaka, Japan. 26th International Conference on Computational Linguistics (COLING 2016), 2016

Toral, A. and Sanchez-Cartagena, V. 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*, pages 1063–1073, Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

Vilar, D. Xu, J., D'Haro L. F., et al., (2006). Error analysis of statistical machine translation output. In : Proceedings of LREC. 2006. p. 697-702.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.