Il existe des méthodes d'identification automatique des entités nommées, mais celles-ci sont d'une précision limitée, et ne proposent pas de traduction. Et par ailleurs, si de telles ressources existent pour l'anglais, nous n'en disposons pas pour le français.

Toutefois, il nous est apparu qu'une approche relativement simple pouvait aider à réduire ces problèmes avec les entités nommées. Par défaut, lorsqu'un système PORTAGE II reçoit un texte à traduire, il commence par mettre tout le texte en minuscules. C'est cette version du texte qui est traduite, pour ensuite être soumise à un processus de restauration intelligente de la casse ("*truecasing*"). Cette façon de faire, pour surprenante qu'elle puisse paraître, permet une meilleure utilisation des données d'apprentissage : le système apprend ainsi que "*banana*" se traduit par "banane", quelle que soit la forme sous laquelle le mot apparaît : "*banana*", "*Banana*" ou "*BANANA*".

Pourtant, il est parfois bénéfique de créer des systèmes entraînés en "casse véritable", c'est-à-dire qui opèrent directement sur des textes qui n'ont pas été préalablement mis en minuscules [1]. L'intérêt, c'est que les systèmes entraînés de cette façon ont souvent une meilleure capacité à discerner entre l'utilisation d'un mot comme nom propre et comme nom commun : le système peut alors distinguer "rice", la céréale et "Rice", la personne. Toutefois, pour que cette approche soit viable, il faut disposer de suffisamment de données d'entraînement. Il se trouve que c'est possiblement le cas pour les textes du SATJ. C'est pourquoi nous avons décidé de mettre cette stratégie à l'essai.

3.3. Évaluation de la qualité

La contribution relative des différentes composantes décrites dans la section précédente est variable. Elle dépend de plusieurs facteurs, notamment les langues de départ et d'arrivée, la taille des corpus, la complexité du domaine, la répétitivité des textes, etc. Par ailleurs, l'ajout de ces composantes n'est pas sans entraîner des coûts, tant au niveau de l'utilisation de la mémoire que du temps de calcul, aussi bien à l'entraînement qu'à l'utilisation. Il convient donc d'analyser l'impact de l'utilisation de chaque composante, afin de choisir la configuration la plus appropriée.

Pour ce faire, nous avons donc créé des systèmes correspondant à une grande variété de configurations possibles, littéralement des dizaines de systèmes différents, et ce pour chacune des quatre cours desservies par le SATJ. Chacun de ces systèmes a alors été testé sur différents jeux de test, et sa performance mesurée au moyen de la métrique BLEU (voir Annexe B). Partant de cette première analyse, nous avons identifié des configurations qui nous paraissaient prometteuses, que nous avons alors soumises à une évaluation manuelle. Nous détaillons ci-dessous cette procédure.

1. En pratique, les choses ne sont pas si simples : il est parfois nécessaire de mettre le premier mot des phrases en minuscules, et cette décision doit être prise de façon intelligente. Concrètement, on affecte au premier mot de la phrase la casse dans laquelle ce mot a été observé le plus souvent – une méthode que nous appelons "nc1".

Tableau 3 – Jeux de test

	EN→FR		FR→EN	
Cour	phrases	mots	phrases	mots
CAF	543	10 391	163	1413
CF	1494	31 678	1494	32 256
CCI	825	13 372	888	13 300

3.3.1. Évaluation automatique

Nos évaluations reposent sur un jeu de test issu de textes produits et traduits par le SATJ après la collecte des corpus d'entraînement (voir l'Annexe B pour plus de détails sur les méthodes d'évaluation automatiques). Notons qu'aucun document n'a été traduit pour la Cour d'appel de la cour martiale durant la période de collecte des données ; c'est pourquoi nous ne disposions pas de données pour cette cour. La taille des textes est détaillée dans le Tableau 3.

Nous ne reproduisons pas ici tous les résultats obtenus lors de cette phase de l'expérimentation, mais résumons l'essentiel de nos observations.

D'abord, à quelques rares exceptions près, tous les nouveaux systèmes obtiennent des scores BLEU plus élevés que les systèmes présentement en production. C'est notamment vrai des systèmes "de base", c'est-à-dire ceux qui ne diffèrent des systèmes en production que par la procédure de préparation des données (utilisation de AlignFactory – voir la Section 3.1). On peut donc en conclure que la qualité des données produites par cette nouvelle procédure est au moins équivalente à celle de la procédure originale.

Par ailleurs, la stratégie qui consiste à entraîner les systèmes à partir de données en casse véritable semble produire de meilleurs résultats que lorsque les textes sont d'abord réduits en minuscules. Toutefois, on observe qu'il est préférable dans ce cas d'appliquer la méthode "nc1", qui modifie sélectivement la casse du premier mot de chaque phrase dans le texte source.

De toutes les composantes additionnelles proposées à la Section 3.2, ce sont les NNJM (modèles de langue neuronaux) qui produisent l'effet le plus marqué. Les systèmes qui incluent cette composante se retrouvent presque systématiquement en tête de classement. En revanche, l'utilisation d'attributs dispersés (*sparse features*) ne contribuent pas à faire augmenter les scores. Quant aux modèles génériques et aux modèles de langue bruts, ils ont généralement un effet bénéfique, mais celui-ci est le plus souvent minime.

En nous basant sur ces observations, nous avons identifié ce qui nous semblait être la configuration la plus prometteuse :

— Entraînement sur des données en casse véritable, traitement nc1
— Modèle de langue neuronal
— Modèles de langue et de traduction génériques
— Modèle de langue brut

Les scores BLEU obtenus avec ces systèmes (que nous appelons "V2") sur les textes des différentes cours sont rapportés dans le Tableau 4.

Tableau 4 – Scores BLEU des systèmes correspondant à la configuration optimale (V2), comparés aux scores obtenus sur les mêmes textes avec les systèmes présentement en production (V1).

	en→fr		fr→en	
Cour	V1	V2	V1	V2
CAF	39.6	41.0		
CF	38.9	40.9		
CCI	38.3	40.8		

3.3.2. Évaluation manuelle

Même si l'évaluation automatique révèle des différences significatives entre les scores BLEU des nouveaux systèmes et ceux présentement en production, ces différences ne sont pas très grandes, et nous avons tenu à nous assurer qu'elles reflètent effectivement des différences qualitatives dans les traductions. Pour ce faire, nous avons appliqué une méthode d'évaluation comparative, que nous décrivons ici.

Dans les fichiers de test utilisés dans l'évaluation automatique, nous avons identifié les phrases où les systèmes V1 (systèmes présentement en production) et V2 (configuration optimale, tel qu'expliqué ci-dessus) produisaient des phrases différentes. Nous avons alors examiné 500 de ces paires de traduction, et pour chacune, nous avons statué laquelle des deux était la meilleure. Cet examen a été effectué par un membre de l'équipe, à l'aveugle, c'est-à-dire que pour chaque paire, l'origine de la traduction n'était pas identifiée, et l'ordre des traductions était permuté de façon aléatoire.

L'annotateur avait la possibilité de marquer deux traductions comme équivalentes. Lorsque les deux traductions n'étaient pas équivalentes, l'annotateur devait spécifier en regard du quel des aspects ci-dessous l'une des traductions était préférable à l'autre :

Sens : L'un des systèmes rend mieux que l'autre le sens de la phrase source

Syntaxe : L'un des systèmes rend mieux la syntaxe ou les règles de grammaire de langue cible

Casse : les traductions sont équivalentes, mais un des systèmes a mieux rendu la casse

Typographie : les traductions sont équivalentes, mais un des systèmes a commis une erreur typographique (espaces, ponctuation, etc.)

Lorsque la différence entre les deux traductions touchait à plusieurs aspects, un seul aspect était noté, en donnant la préférence à l'aspect qui apparaît le premier dans la liste ci-dessus. Notons que cette annotation n'a été effectuée que sur les traductions de l'anglais vers le français.

Les résultats de cette annotation sont présentés dans le Tableau 5. Cette annotation révèle que les systèmes V1 et V2 se comportent de façon identique en ce qui a trait au respect du sens de la phrase source. Du point de vue de la syntaxe, les systèmes V2 se comportent sensiblement mieux que les V1 (45 contre 27). En pratique, cet avantage se manifeste au niveau des accords grammaticaux, mais également dans l'ordre des mots.

Tableau 5 – Résultats de l'évaluation comparative entre les versions V1 et V2 des systèmes

Meilleur système	Sens	Syntaxe	Casse	Typographie	Total
V1	51	27	3	2	83
V2	51	45	42	33	171
équivalents					246
Total	102	72	45	35	500

Mais les principaux avantages des nouveaux systèmes se situent au niveau du traitement de la casse (42 contre 3) et de la typographie (33 contre 2). Les gains au niveau de la casse s'expliquent principalement par l'entraînement sur des données en casse véritable. Ce résultat est particulièrement intéressant, parce qu'il signale probablement des problèmes systémiques au niveau de notre mécanisme de restitution de la casse. Quant aux gains en typographie, ils sont essentiellement attribuables à une meilleure normalisation de certains caractères (tirets, apostrophes, etc.) dans la préparation des données.

Globalement, le bilan est donc positif pour cette nouvelle génération de systèmes. Toutefois, les gains relativement modestes en regard des problèmes réputés difficiles (problème de syntaxe et de sens) laissent à penser que nous atteignons peut-être les limites de ce qu'on peut attendre des technologies de TA statistique à l'heure actuelle.

4. Conclusions et recommandations

Nous résumons ici les grandes lignes de cette deuxième année de collaboration entre le SATJ et le CNRC. Rappelons que, au cours de cette deuxième année, nous avions initialement prévu travailler sur 1) l'amélioration des systèmes de TA ; 2) l'automatisation du processus d'entraînement ; 3) la mise en production des systèmes ; et 4) le routage automatique des données. Toutefois, vu les circonstances difficiles qui ont marqué l'année 2016, nous avons choisi de nous limiter aux deux premiers objectifs.

Les travaux portant sur l'amélioration des systèmes de TA ont permis d'obtenir des gains mesurables, aussi bien en évaluation automatique qu'en évaluation manuelle. Toutefois, ces gains restent modestes, et n'ont pu être obtenus qu'au prix d'une augmentation substantielle de la mémoire et des temps de calcul requis par les systèmes. On peut probablement affirmer que la qualité obtenue se situe aux limites de ce qu'il est raisonnable d'attendre à l'heure actuelle des systèmes de TA statistiques.

Un des aspects sur lesquels nous avons fait peu de progrès est celui du traitement des entités nommées. Si la stratégie d'entraînement en casse véritable s'est avérée globalement fructueuse (notamment en ce qui concerne le rendu de la casse), il semble qu'elle contribue bien peu à régler le problème pour lequel elle avait été initialement proposée, soit celui de la traduction des noms de personnes, de lieux, d'organismes, etc. Le problème avec ces entités nommées, c'est qu'elles sont souvent très "locales", c'est-à-dire qu'elles n'apparaissent dans un seul document. De ce fait, le système ne les connaît ha-

bituellement pas. À cet égard, et dans le contexte du SATJ, nous croyons qu'il serait très prometteur d'examiner l'avenue des méthodes d'*entraînement incrémental* : cette approche repose sur la mise-à-jour en continu des tables de traduction d'un système de TA, permettant de capitaliser sur les corrections effectuées par le traducteur, pendant le processus de traduction assistée. Nous croyons que le scénario d'utilisation envisagé par le SATJ se prête particulièrement bien à cette approche.

Par ailleurs, nous considérons avoir effectué des progrès importants sur le plan de l'automatisation de l'entraînement des systèmes. D'une part, avec l'utilisation du produit AlignFactory, la gestion des corpus d'entraînement et des mémoires de traduction est grandement simplifiée, et ne requiert plus l'intervention d'un expert. D'autre part, partant de ces données, la création des systèmes PORTAGE II est presque entièrement automatisée. Toutefois, nous n'en sommes probablement pas encore au point où cette opération pourrait être prise en charge par le SATJ. À ce stade, toutefois, il serait entièrement envisageable de confier cette tâche à un tiers parti.

Remerciements

Un grand merci à toute l'équipe des développeurs de PORTAGE II pour leur inestimable expertise et leur soutien indéfectible tout au long du projet : Darlene Stewart, Samuel Larkin et Éric Joanis. Merci aussi à Roland Kuhn et Pierre Charron pour leur appui inconditionnel dans cette aventure. Et, bien entendu, merci à Lucie Langlois, sans quoi rien de tout cela n'aurait été possible. Bon voyage, Lucie !

Références

Chiang, D., K. Knight et W. Wang. 2009, «11,001 new features for statistical machine translation», dans *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 218–226.

Devlin, J., R. Zbib, Z. Huang, T. Lamar, R. M. Schwartz et J. Makhoul. 2014, «Fast and robust neural network joint models for statistical machine translation.», dans *ACL (1)*, Citeseer, p. 1370–1380.

Federico, M., N. Bertoldi, M. Cettolo, M. Negri, M. Turchi, M. Trombetti, A. Cattelan, A. Farina, D. Lupinetti, A. Martines et collab.. 2014, «The Matecat Tool», dans *COLING (Demos)*, p. 129–132.

Macklovitch, E. 2016a, «An Independent Evaluation of Portage Translations for the Courts Administration Service», cahier de recherche, SATJ.

Macklovitch, E. 2016b, «The Field Trial of Portage at the Courts Administration Service», cahier de recherche, SATJ.

Papineni, K., S. Roukos, T. Ward et W.-J. Zhu. 2002, «BLEU : a Method for Automatic Evaluation of Machine Translation», dans *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 311–318.

Simard, M. et P. Isabelle. 2015, «Traduction automatique au Service administratif des tribunaux judiciaires – Livrable 1 : Étude du corpus du SATJ en vue de son utilisation pour la traduction automatique», cahier de recherche, Conseil national de recherches Canada.

Stewart, D., R. Kuhn, E. Joanis et G. Foster. 2014, «Coarse "split and lump" bilingual language models for richer source information in smt», dans *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA). Vancouver, BC, Canada*, vol. 1, p. 28–41.

Watanabe, T., J. Suzuki, H. Tsukada et H. Isozaki. 2007, «Online large-margin training for statistical machine translation», dans *In Proc. of EMNLP*, Citeseer.

Annexe A. La traduction assistée par ordinateur (TAO)

Apparues il y a près de 30 ans, les mémoires de traduction (MT) font, depuis plusieurs années, partie intégrante du paysage dans la plupart des services de traduction. Les archives de traduction sont ainsi systématiquement organisées sous forme de "bi-texte", rendant explicites non seulement les correspondances globales entre des paires de documents mais aussi les correspondances plus fines entre les phrases contenues dans ces paires de documents. L'indexation de ces ressources permet de fouiller ces MT pour y retrouver rapidement des traductions qui auraient été faites antérieurement pour les éléments qu'on s'apprête à traduire ou pour des éléments similaires.

Bien que la TA remonte à une époque bien antérieure (plus de 60 ans), ce n'est que très récemment, avec l'avènement des méthodes de TA dites "statistiques" (TAS), qu'on a commencé à voir apparaître cette technologie sur les postes de travail des traducteurs.

Or, les systèmes de traduction automatique statistique (TAS) se nourrissent exactement des mêmes bi-textes que les MT. Ceci n'est pas surprenant dans la mesure où on peut en fait voir les MT comme le niveau zéro de la TAS : un système de MT offre comme (ébauche de) traduction d'une phrase d'entrée la traduction de la phrase la plus semblable ayant été traduite antérieurement. Généralement, on ne fournit une telle sortie que lorsque l'on a réussi à trouver une phrase déjà traduite qui est suffisamment similaire à la phrase à traduire (disons 70% de recouvrement ou plus). Par conséquent, les MT sont surtout utiles dans les cas où on traduit régulièrement des documents très similaires, comme par exemple des versions différentes du même document. Les MT sont beaucoup moins utiles dans les domaines où le contenu textuel varie au jour le jour, comme dans la presse écrite ou dans le Journal des débats de la Chambre des communes. Les systèmes de TAS, contrairement aux MT, sont capables de décomposer automatiquement les correspondances entre phrases en correspondances plus fines et de recomposer automatiquement une traduction de phrase à partir d'éléments puisés dans plusieurs paires de phrases différentes de la mémoire. Ceci leur permet en principe de s'avérer plus utile que les MT dans les domaines où on observe peu de répétition de phrases. Malheureusement, jusqu'à récemment, la qualité des traductions machine s'était rarement avérée suffisante pour qu'un traducteur puisse en tirer un réel profit. Le succès bien connu de la TA sur les prévisions météo a longtemps constitué l'exception qui confirmait cette règle. L'expérience a montré qu'il existe très peu de domaines où l'on produit de grandes quantités qui sont d'un niveau de simplicité comparable à celui des prévisions météo et qui doivent être traduits.

Ceci dit, ces dernières années les approches statistiques ont permis à la traduction automatique de connaître des progrès très substantiels : la qualité moyenne des traductions machine s'est beaucoup améliorée. Il est maintenant possible d'obtenir des traductions machine relativement bonnes pour des textes qui sont nettement plus complexes que les prévisions météo. C'est pour cette raison qu'un bon nombre de cabinets de traduction au Canada et à l'étranger se sont déjà empressés d'adopter la TAS comme un élément important de la nouvelle boîte à outils du traducteur. Sur la scène canadienne, le système PORTAGE II, fruit de 10 ans de recherche au CNRC, est à l'avant-garde de ce mouvement : il est utilisé quotidiennement par un nombre de plus en plus substantiel de

traducteurs depuis 2010.

Annexe B. Mesures de qualité de la TA

Le succès de la TAS repose en grande partie sur les avancées récentes dans le domaine de l'apprentissage-machine. L'application de ces méthodes au domaine de la traduction a exigé que les chercheurs mettent au point des façons d'évaluer automatiquement la qualité des traductions automatiques. Il existe maintenant une grande variété de méthodes d'évaluation de la TA, mais la plus connue et la plus répandue est sans conteste la méthode BLEU (Papineni et collab., 2002). Nous décrivons brièvement cette méthode ici.

Le score BLEU (Papineni et collab., 2002) n'est rien de plus qu'une mesure du chevauchement entre la traduction produite par la machine et une ou plusieurs traductions humaines de référence. Le chevauchement se mesure par le nombre de *n-grammes* partagés. Par *n-gramme* on entend une suite continue de n mots. Par exemple, la phrase "le petit chat lappe le lait" comprend les unigrammes en (1), les bigrammes en (2), les trigrammes en (3) et les quadrigrammes en (4) :

1. le, petit, chat, lappe, lait.
2. le petit, petit chat, chat lappe, lappe le, le lait.
3. le petit chat, petit chat lappe, chat lappe le, lappe le lait.
4. le petit chat lappe, petit chat lappe le, chat lappe le lait.

Il s'agit donc de compter la proportion des n-grammes de la traduction machine qui apparaissent aussi dans une ou plusieurs traductions de référence. La plupart du temps, on prend en compte l'ensemble formé par tous les n-grammes pour $1 \leq n \leq 4$. La formule complète incorpore certains critères additionnels et le lecteur intéressé pourra consulter Papineni et collab. (2002) ou encore l'entrée pertinente de Wikipédia.

Dans notre cas, il n'y aura généralement qu'une seule traduction de référence, à savoir celle provenant de ce que nous avons nommé ci-dessus le jeu de test extrait du corpus SATJ. (Bien entendu, le jeu de test devra être choisi de manière à assurer qu'il soit représentatif de l'ensemble des textes du domaine concerné.) On fera donc traduire la partie source du jeu de test par la machine et on calculera le score BLEU de la traduction relativement à la traduction de référence que constitue la partie cible du même jeu de test.

Manifestement, ce score BLEU ne fait rien de plus que d'approximer de façon plutôt grossière la qualité des traductions machine. Cette métrique ignore complètement le fait que la (ou les) traduction de référence ne constitue pas nécessairement la seule bonne traduction possible. Toutefois, il est maintenant bien établi que le classement effectué par BLEU sur la qualité relative des traductions du même texte par différents systèmes de TA a une très forte corrélation avec le classement qui serait obtenu par un panel d'experts humains. Il n'est nullement question d'utiliser BLEU pour évaluer des traductions humaines. Mais la valeur de cette métrique pour établir un classement de traductions machine est maintenant reconnue.

On notera toutefois que le problème que nous abordons dans cette étude est quelque peu différent de celui que nous venons de mentionner : pour nous, il ne s'agit pas de classer le niveau de qualité relatif obtenu par différents systèmes sur les mêmes textes mais plutôt de déterminer si un système donné produit des résultats acceptables pour un domaine donné. EN pratique, il n'existe pas de barème précis pour interpréter les scores BLEU. D'une façon générale, pour la traduction entre l'anglais et le français, on remarque que des scores inférieurs à 30 coïncident souvent avec des traductions de qualité médiocre, alors que des scores approchant les 50 vont souvent de pair avec des traductions acceptables.

Annexe C. MateCat

MateCat [2] est un environnement Web de traduction assistée par ordinateur (TAO). L'application MateCat, à laquelle on accède au moyen d'un simple fureteur [3], permet de gérer des projets de traduction de complexité variable, impliquant un ou plusieurs traducteurs et réviseurs. Outre diverses fonctions de gestion du processus de traduction, MateCat fournit un accès au service de mémoire de traduction en-ligne MyMemory [4], de même qu'à différents moteurs de traduction automatique publics (Google Translate, Bing) et privés (DeepLingo, TauYou, Moses, etc.). Lors de la création d'un projet de traduction, l'utilisateur sélectionne un ensemble de ressources auxquelles il souhaite accéder. Pendant la traduction, les propositions provenant de ces différentes ressources sont affichées, et peuvent être récupérées facilement par le traducteur. Une description plus détaillée des fonctionnalités offertes par MateCat dépasse le cadre du présent rapport ; nous invitons le lecteur à faire lui-même l'essai du logiciel.

L'accès à l'application MateCat, de même qu'à la plupart des services connexes (MyMemory, Google Translate, Bing, etc.) est gratuit. La plateforme MateCat est en fait le fruit d'un projet de recherche européen (Federico et collab., 2014), et le code source de l'application est disponible en *Open Source*, sous licence LGPL. La version de ce logiciel hébergée à l'adresse `http://matecat.com` est gérée par la firme *Translated.net*, un LSP basé en Italie [5], qui l'utilise comme outil de marketing pour ses services de post-édition. L'intégration de nouveaux moteurs de TA, tels que PORTAGE II, est encouragée.

Avec l'aide de l'équipe de soutien technique de Translated.net, il nous a été relativement aisé d'intégrer PORTAGE II à MateCat. En pratique, MateCat voit PORTAGE II comme un "service Web" : les moteurs PORTAGE II sont hébergés sur un serveur Web, et MateCat y accède via une interface programmatique (API) publique.

2. `http://www.matecat.com`
3. Nos essais ont tous été effectués avec Chrome, le fureteur de Google.
4. `http://mymemory.translated.com`
5. `http://www.translated.net`

Annexe D. Mémoire de traduction MyMemory

Un autre aspect de l'intégration avec MateCat est l'accès au service de mémoire de traduction MyMemory. Comme MateCat, MyMemory est un service Web gratuit. Il est possible d'y accéder de manière autonome, via l'URL `https://mymemory.translated.net`. On peut alors y téléverser ses propres mémoires, en format TMX, et effectuer des recherches ponctuelles, soit sur celles-ci, sur une des mémoires publiques. Mais le principal intérêt est bien sûr d'y accéder par l'intermédiaire d'un outil de TAO, tel que MateCat. Il est toutefois notoire que les principales fonctionnalités de MyMemory sont accessibles via une interface programmatique (API). On peut donc envisager consulter et enrichir les mémoires à partir d'applications tierces.

Bien que nous n'ayons pas testé à fond les fonctionnalités de MyMemory, nous avons pu nous faire une idée du service à travers nos interactions avec MateCat. En autant que nous puissions en juger, MyMemory est tout-à-fait comparable avec la plupart des autres offres commerciales traditionnelles (SDL Trados, MultiTrans, etc.). À la différence de ces dernières, toutefois, du fait qu'on y accède par une interface Web, il n'est pas possible de garantir la confidentialité des informations qu'on y archive. Considérant la nature publique des documents traduits par le SATJ, cet aspect n'est probablement pas un problème. En fait, le SATJ pourrait même considérer la possibilité de permettre un accès public à ses mémoires de traduction.

Dans le premier rapport d'étape du présent projet (Simard et Isabelle, 2015), nous avons souligné le potentiel relativement faible des mémoires de traduction vis-à-vis les textes du SATJ. Malgré tout, il nous semble extrêmement opportun de constituer une mémoire exhaustive des traductions du SATJ, et de maintenir celle-ci à jour de façon systématique. En effet, pour maintenir un niveau de qualité acceptable, il est important de réentraîner périodiquement les systèmes de TA PORTAGE II. Pour ce faire, il est essentiel d'avoir accès à l'ensemble des archives du service. Une mémoire de traduction telle que celle proposée par MyMemory offre un accès simple et robuste à ces données [6].

6. Pour utiliser MyMemory à cette fin, il est essentiel de pouvoir télécharger les mémoires mises-à-jour, en format TMX. Cette fonctionnalité ne semble pas être disponible à travers l'API MyMemory. Toutefois, elle serait possible à travers l'interface utilisateur de MateCat.

Study on How the Translation Bureau's Corpus of Translations Could Be Used in Machine Translation

Pierre Isabelle
Michel Simard
Multilingual Text Processing
National Research Council Canada

1. Introduction

As part of a collaboration on machine translation, the Government of Canada's Translation Bureau (the Bureau) provided the National Research Council (NRC) with a copy of the immense corpus it has created from previous translations. NRC committed to conducting a study on the best ways to use this corpus in machine translation. This document is a preliminary report on the activities carried out by NRC in this study and the provisional findings of the study.

The corpus received from the Bureau (the Bureau corpus) is organized as a bitext, which makes explicit not only the overall matches found in pairs of documents but also more specific matches within sentences contained in the pairs of documents. The Bureau is already using this corpus to provide its translators with a translation memory. The translators now have tools that allow them to search the memory and quickly find previous translations for the elements they will be translating or similar elements.

Statistical machine translation (SMT) systems use exactly the same bitexts as translation memories, which is not surprising, given that translation memories can be considered base-level SMT systems: for input sentences, translation memories offer translations of the most similar previously translated sentences. This output is generally provided only when a previously translated sentence is sufficiently similar to the sentence to be translated (e.g. 70% similarity or higher). As a result, translation memories are especially useful for regularly translated documents that are very similar, e.g. different versions of the same document. Translation memories are much less useful in domains where the textual content varies from day to day, e.g. written press or the Journal of House of Commons debates.

Unlike translation memories, STM systems are able to automatically break down matches into finer matches and recast translations using elements from various pairs of sentences in the memory. This makes them more useful than translation memories in domains where sentences rarely repeat.

Unfortunately, until recently, the quality of machine translations was rarely sufficient to be truly useful to translators. The well known success of machine translation in weather forecasts has long been the

exception to the rule. The experiment showed that there are very few domains in which large quantities of information as simple as weather forecasts are produced and need to be translated.

That being said, in the last few years, statistical approaches have helped machine translation progress substantially: the average quality of machine translations has improved significantly. It is now possible to obtain relatively good machine translations for texts that are much more complex than weather forecasts. That is why a good number of translation firms in Canada and abroad have been quick to use SMT as a key component of the new translator toolbox. In Canada, PORTAGE is spearheading this movement: it has been used daily by an increasingly large number of translators since 2010.

This therefore seems like the right time for the Bureau to closely examine the potential for SMT in its operations. An organization such as the Bureau, however, must exercise caution. The fact that machine translation has progressed sufficiently for people to use it does not mean that it should now be used across the board. The Bureau would be well advised to first try it out in the most conducive situations, and if that proves successful, to expand its use to increasingly more demanding situations. The big question is where to start.

2. Overall approach

How can we determine which situations are most conducive to machine translation? There are at least two things to consider: 1) the technical feasibility of obtaining reasonably good machine translations, and 2) the possibility of successfully implementing new technologies in a particular organization. Although the second point is as important as the first, it is outside the scope of this study. Our attention will be limited to determining how to choose domains of application in which machines can produce relatively good quality translations.

This problem is not new to Canadian research. After the success of machine translation for weather forecasts in the late 1970s, Canadian researchers looked at the possibility of reproducing this success in other domains. A small community of linguistics researchers began studying sub-languages, i.e. relatively limited closed subsets of a particular language used by a specific community of writers to produce standardized texts for a very specific purpose [1]. It is easy to find examples other than weather forecasts: stock market reports, descriptions of sporting events, work descriptions for positions in government organizations, etc.

Research conducted in the 1980s on the notion of sub-languages was based on essentially manual surveys of the relative linguistic complexity of different types of texts at various levels: vocabulary, syntactic constructions, semantic space, etc. Given the size of the Bureau corpus, exhaustive manual surveys of that kind are completely out of scope. We are talking about approximately 2.5 billion words in 174 domains. Strictly speaking, these are more clients (to whom an invoice is sent) than domains of textual content. Below, Table 1 provides some basic statistics on the largest domains in the corpus.

	EN-FR				FR-EN			
Domain	Doc.	Seg.	EN Words	FR Words	Doc.	Seg.	EN Words	FR Words
44	111,326	9.51M	113.10M	138.59M	26,168	2.30M	28.15M	33.87M
349	80,497	5.79M	72.39M	89.59M	10,276	786.72K	10.54M	12.45M
313	23,204	2.23M	44.21M	49.10M	2,287	183.41K	3.65M	3.96M
47	43,166	2.86M	38.90M	48.37M	6,628	465.76K	6.51M	7.86M
10	29,310	2.90M	38.33M	47.31M	2,991	287.67K	3.96M	4.65M
999	44,707	2.62M	37.45M	49.28M	3,226	198.74K	3.30M	4.29M
40	32,628	2.34M	32.23M	39.78M	10,391	774.90K	11.41M	13.37M
75	29,595	2.33M	31.96M	39.02M	9,039	671.29K	9.53M	11.64M
32	33,654	2.13M	31.68M	38.80M	3,466	221.49K	3.58M	4.26M
100	36,065	2.10M	31.42M	38.86M	3,644	230.66K	3.74M	4.41M
146	32,226	2.26M	30.79M	38.92M	4,718	413.48K	6.02M	7.47M
128	18,104	1.70M	22.89M	27.93M	3,267	257.05K	3.71M	4.31M
333	25,158	1.83M	22.76M	27.98M	6,548	389.28K	5.25M	6.35M
555	8,674	1.08M	20.87M	23.88M	2,980	309.84K	6.35M	7.03M
1	15,333	1.36M	19.88M	24.46M	4,977	534.60K	8.01M	9.65M
69	22,944	1.35M	19.35M	23.96M	2,543	167.22K	2.59M	3.14M
330	24,114	1.31M	18.50M	22.59M	2,733	171.21K	2.63M	3.13M
23	9,986	1.24M	18.14M	22.22M	2,471	278.95K	4.02M	4.70M
471	19,792	1.14M	16.30M	19.80M	3,185	171.37K	2.53M	3.04M
227	18,140	1.18M	16.22M	20.00M	2,141	156.59K	2.18M	2.58M

Table 1: Statistics on largest domains (clients)

The size of the various sub-corpora varies greatly, from a few hundred words (e.g. International Centre for Infectious Diseases – 612) to a hundreds of millions of words (e.g. National Defence – 44).

Upon examination, there are clearly links between the various clients and predominant themes. In the best case scenario, the content would be almost perfectly homogeneous. For example, the sub-corpus associated with the Canada Industrial Relations Board (client 184) appears to contain almost exclusively the decisions of this labour relations administrative tribunal. The term sub-language would be appropriate in such cases.

However, the sub-corpora associated with other clients contain much more heterogeneous content. For example, documents randomly extracted from the largest sub-corpus, National Defence (44), seemed to differ greatly: descriptions of training exercises, work descriptions, evaluation questions for job interviews, chapters from a book on the history of World War I, an avionics technical manual, the procedure for drawing up a budget, etc. There are probably some large homogeneous subsets in the National Defence mega-corpus, but we cannot identify these structures using manual methods.

Because we cannot scan the entire Bureau corpus manually, we will use automatic methods. **We will first compare the existing client sub-corpora**. In a later phase of our study, we hope to go beyond this structure in order to identify a) homogeneous sub-corpora within existing client corpora, and/or b) homogeneous groups of such sub-corpora across different clients.

We will make two different types of sub-corpus comparisons: 1) the relative linguistic complexity of these sub-corpora, evaluated using automatic mechanisms, and 2) the relative quality of the machine translations obtained from various sub-corpora, evaluated using automatic mechanisms. It may be surprising how feasibility it is to carry out the second comparison. It is clear that when the aforementioned studies on sub-languages were conducted, such a comparison was completely out of the question. The construction of each machine translation system, tailored to a specific domain, required several of years of work by a team of computational linguistics experts. The purpose of the studies on linguistic complexity was to predict the various domains' relative potential for machine translation: the researchers had to avoid at all cost investing years developing systems that were doomed to fail because the problem was too difficult to solve.

The situation changed significantly with the dawn of SMT: specialized systems are largely developed automatically, and the quality of their translations can also be estimated automatically. So, why not completely eliminate the linguistic complexity component? Would it not be better to simply evaluate the relative quality of machine translations in various domains, given that the desired outcome is to identify the domains in which the best translations are obtained?

Although evaluating the quality of machine translations can largely be done automatically, it still requires significant resources. It involves generating specific PORTAGE models for each domain we want to compare. This requires a fair amount of work in preparing the sub-corpora and a lot of time doing calculations to create the corresponding PORTAGE models.

There is likely a strong correlation between the relative linguistic complexity of the various domains and the relative quality of the machine translations obtained for those domains. If that is true, the measure of relative linguistic complexity could be used as an inexpensive substitute for measuring the relative quality of the machine translations.

3. Measures of lexical richness in various sub-corpora

As indicated above, the studies on sub-languages conducted in the 1980s focussed on analyzing linguistic complexity at various levels: lexical, syntactic, semantic space, etc. Of these levels, the lexical level lends itself most easily to automatic studies. It is relatively easy to compare the size of the vocabularies at play in these domains. The vocabulary of a sub-corpus is directly observable: a few relatively simple rules for segmenting a text into isolated words as well as simple sorting and counting algorithms are all that is required.

It is much more difficult to conduct automatic analyses of syntactic or semantic complexity, as syntactic and semantic structures are not directly observable in texts. To evaluate richness or complexity, they must first be observable. In the case of syntax, this requires parsers capable of automatically assigning each sentence of a text an explicit grammatical structure (generally referred to as a syntax tree). There are currently no parsers sufficiently robust, precise and fast to reliably carry out this task on a corpus as large and heterogeneous as the one we are dealing with. Evaluating semantic complexity is even more difficult, as this notion is still misunderstood.

As a result, our automatic linguistic analyses will be limited to lexical richness, and we will assume a consistent linguistic complexity across the various levels. It would therefore be very natural to speculate that the larger the vocabulary of a domain, the more complex the semantic space.

Various measures relating to the size of the vocabulary can serve as indicators of the relative complexity of a textual domain (in this case, a client's archives). The most well-known indicator of lexical richness is indubitably the type-occurrence ratio (TTR), which compares the size of the vocabulary (i.e. the number of distinct word types) to the size of the entire corpus (i.e. the number of word occurrences). For example, in the sentence *To be, or not to be, that is the question,* there are 10 occurrences but only 8 different word types: *to, be, or, not, that, is, the* and *question*. The ratio of types to occurrences is therefore TTR = 8/10 = 0.8. Clearly, calculating this ratio for a single sentence is not very useful. However, comparing the measures obtained for larger quantities of text (e.g. thousands of sentences) in various domains, provides a fairly accurate indication of the relative lexical richness of these domains.

An alternate measure is the vocabulary's growth rate, P(N), which is calculated as the number of words that appear only once in the corpus, divided by the size of the corpus. P(N) can be interpreted as the probability that the (N+1)th word in the corpus is new (i.e. has never before been seen in the corpus). In the example above, 6 of the 8 observed word types appeared only once in the 10-word sentence. We therefore conclude that an 11th word would have a 6 out of 10 chance of being new. As it turns out, the next word in Hamlet's famous monologue is *Whether*, which is indeed new in this context.

These two indicators vary according to the size of the sample taken into consideration (in practice, they decrease as the size of the corpus increases). This is problematic when the goal is to compare domains of different sizes, which is the case in this study. Researchers examined this problem and tried to develop alternate indicators that were less sensitive to corpus size. Tweedie and Baayen [5] reviewed many of these indicators, to compare their behaviour.

To determine the various domains' potential for machine translation, as we wish to do in this study, there is a simple solution to the problem: measure the lexical richness of random samples of the same size, regardless of the total size of the domains. That is what we did with the Bureau corpus data: for each domain in the corpus, we randomly selected a sufficient number of documents to create a sample of **1 million words**; from this sample, we calculated the values of several indicators proposed in Tweedie and Baayen's article. A few of these values are provided in the table below, for the largest domains in the corpus.

Domain	N	V	V1	TTR	P(N)	R	C	k	a2
313	1,000,470	16,676	6,314	0.016668	0.006311	0.16672	0.703658	0.370235	0.214492
999	1,001,069	17,339	5,970	0.01732	0.005964	0.1733	0.706449	0.371714	0.212463
32	1,000,022	23,690	8,695	0.023689	0.008695	0.2369	0.729093	0.38361	0.196089
555	1,000,429	23,747	8,398	0.023737	0.008394	0.23742	0.729245	0.383697	0.195973
349	1,002,717	24,485	9,209	0.024419	0.009184	0.24452	0.73134	0.384839	0.194425
40	1,004,232	24,686	9,419	0.024582	0.009379	0.24634	0.731851	0.385134	0.194033
330	1,000,011	24,765	8,528	0.024765	0.008528	0.24765	0.732306	0.3853	0.193763
128	1,008,359	25,119	9,232	0.024911	0.009155	0.25015	0.732892	0.385752	0.193223
10	1,000,499	25,579	9,644	0.025566	0.009639	0.25573	0.734621	0.386527	0.192081
75	1,000,364	27,115	10,505	0.027105	0.010501	0.2711	0.738849	0.388749	0.189023
146	1,000,553	27,576	10,292	0.027561	0.010286	0.27568	0.740059	0.389389	0.188144
69	1,002,497	27,698	11,390	0.027629	0.011362	0.27663	0.740274	0.389536	0.187962
47	1,002,664	28,748	10,653	0.028672	0.010625	0.2871	0.742958	0.390951	0.186017
227	1,002,163	28,822	10,705	0.02876	0.010682	0.28791	0.743171	0.391055	0.18587
333	1,000,008	29,601	11,466	0.029601	0.011466	0.29601	0.745217	0.392094	0.184418
23	1,000,410	30,716	12,047	0.030703	0.012042	0.3071	0.747872	0.393497	0.182491
44	1,000,162	31,243	11,241	0.031238	0.011239	0.3124	0.749117	0.394148	0.181593

Table 2: Lexical richness of largest domains; 1M word samples, sorted from lowest TTR to highest TTR

You can see that all the indicators are very closely correlated, i.e. even though the values are not directly comparable, the relative size of the domains is approximately the same. The domain with the lowest TTR (domain 313) also obtained the lowest R and C values, and the highest a2 value, and the inverse is true for the domain with the highest TTR (domain 44). The P(N) indicator yields slightly different results.

For the rest of this study, we will focus on the three largest domains in the corpus, i.e. domains 44 (National Defence), 313 (Immigration and Refugee Board of Canada) and 349 (Public Works and Government Services Canada).

4. Measures of the quality of machine translations for various sub-corpora

The methods used by the SMT research community provide us with an option that was unimaginable at the time of the aforementioned studies on sub-languages: creating specific STM systems for each

domain and measuring their relative performance. This option is possible because **systems can largely be created and evaluated automatically**. If this method is successful, it will then be possible to select the domains where machine translation is most effective before launching application projects.

To create an SMT system for a specific domain, the corpus must be divided into two data sets: training data and test data. The first set is then uploaded to the PORTAGE learning module (in both cases, a set of source language sentences each with a corresponding correct translation), and PORTAGE automatically learns the model for the domain in question.

What is most surprising is that once the SMT system has been automatically created, **the quality of the translations obtained can also be evaluated automatically**, using one or more of the automatic evaluation metrics developed by SMT researchers in recent years. The most well-known of these metrics is the BLEU score.

4.1 The BLEU score

The BLEU score [2] simply measures the overlap between a translation produced by a machine and one or more reference translations produced by a human. The overlap is measured using the number of common n-grams, i.e. sets of *n* consecutive words. For example, the sentence *The small cat drinks the milk* contains the following:

- 1-grams: the, small, cat, drinks, milk
- 2-grams: the small, small cat, cat drinks, drinks the, the milk
- 3-grams: the small cat, small cat drinks, cat drinks the, drinks the milk
- 4-grams: the small cat drinks, small cat drinks the, cat drinks the milk

The number of n-grams in the machine translation that also appear in one or more reference translations is counted. Most of the time, the set created by all n-grams where $1 \leq n \leq 4$ are taken to account. The complete formula includes additional criteria, which are described in [2] or on Wikipedia [3].

In our case, there will usually be only one reference translation—that from the aforementioned test set extracted from the Bureau corpus. We will have a machine translate the source text in the test set and then we will calculate the BLEU score for the translation, in relation to the reference translation, which is the target text in the test set.

Clearly, the BLEU score only roughly approximates the quality of machine translations. This metric completely ignores the fact that reference translations are not necessarily the only good translations possible. However, it is now well established that BLEU scores for the relative quality of translations of the same text by different machine translation systems have a very strong correlation to the rating that would be provided by a panel of human experts. There is no intention of using BLEU to evaluate human translations, but the value of this metric for rating machine translations is now recognized.

It should be noted, however, that the issue we are addressing in this study is somewhat different from the one we just mentioned: we are not trying to rate the relative quality produced by various systems for the same texts but rather the relative difficulty of various types of texts for a single machine.

We assume the following, which seems very plausible to us: the BLEU score obtained for various domains is inversely proportional to the degree of difficulty for the SMT in these domains. However, certain precautions must be taken. We cannot directly compare systems created with training sets of equal size, and the test set must clearly be chosen in such a way that it is representative of all the texts in the domain in question.

4.2 Machine translation experiments

Generally speaking, we expect the lexical richness of a domain to be inversely correlated to automatic translatability: the more limited the vocabulary, the easier it is to automatically translate the texts in that domain, and vice versa.

To verify this hypothesis, we created STM systems for the three largest domains in the corpus. By a stroke of luck, these three domains had very different profiles in terms of lexical richness. As we indicated earlier, domain 44 is not only the largest domain but also one of the domains with the most lexical variety and has one of the highest TTRs and P(N)s. By contrast, domain 313 has one of the lowest TTRs and P(N)s. Domain 349 is in the middle of the pack. In summary, lexically speaking, domain 313 is *poor*, domain 349 is *average* and domain 44 is *rich*.

The quality that can be expected of an SMT system depends largely on the quantity of data available to train the system. For each of the three selected domains, we therefore created systems with increasingly larger data sets:

- **Small**: approximately 1M source words
- **Medium**: approximately 5M source words
- **Large**: approximately 25M source words

In practice, we did the following for each of the three domains:

1. Extracted, through random sampling, a test set of approximately 2,000 pairs of segments (source texts and translations)
2. Divided the remaining data into four blocks:
    - Block 1: approximately 1M source words
    - Block 2: approximately 4M source words
    - Block 3: approximately 20M source words
    - Block 4: the remaining words
3. Trained three STM English-French systems:
    - Small: using the data from block 1
    - Medium: using the data from blocks 1 and 2

- Large: using the data from blocks 1, 2 and 3
  (We kept block 4 for future experiments.)

4. Used each system to translate the test set segments
5. Evaluated the quality of the translations, using the BLEU metric

All the tests were conducted from English → French.

The results are provided in Table 3.

Domain	1M Words (Small)		5M Words (Medium)		25M Words (Large)	
	Src Words	BLEU	Src Words	BLEU	Src Words	BLEU
313 (poor)	1.01 M	40.3	4.90 M	45.8	25.28 M	50.7
349 (average)	1.00 M	28.0	5.26 M	33.5	27.26 M	42.9
44 (rich)	1.24 M	23.2	6.06 M	27.6	29.48 M	33.2

Table 3: BLEU scores achieved by machine translation systems

It is usually risky to compare BLEU scores obtained for various test sets (remember that each domain corresponds to a test set created from that domain); however, the very significant differences between the systems created for the three domains confirm that lexical richness does indeed have the expected impact on translation quality: the poorer the domain is lexically speaking, the easier it is for a machine translation system to translate texts in that domain. Even though the performance in each domain increases very slightly when more training data is used, **the quality achieved by the smallest system for a lexically poor domain (313) is still superior to the quality we would hope to achieve with 25 times the data in a lexically rich domain (44)**.

Furthermore, there is no precise scale for interpreting BLEU scores. For English to French translation, scores under 30 often coincide with mediocre quality translations, while scores approaching 50 often coincide with acceptable translations. Using these rough figures, the results obtained for domain 44 are not very promising, and the results for domain 349 are only significant when a lot of training data (25M words or more) is available. In contrast, **the results for domain 313 are very encouraging, even with as few as 1 million words**.

5. Link to translation memories

As we said earlier, SMT systems and translation memories use the same data. In addition, it seems that the domains, and more specifically the segments, for which translation memories are useful, are also those in which we can expect the best automatic translations. In practice, machine translation is well-suited to repetitive domains. One might wonder if there is a link between repetition and lexical richness. Intuitively, a lexically poor domain is intrinsically repetitive: because the vocabulary is limited, the same words repeat! But there is no evidence that this repetition applies to segments.

To determine the relationship between lexical richness and repetition, we used the same test sets as above in a translation memory and measured the proportion of the text that had exact and fuzzy matches (similarity of 0.70 to 0.85). For these experiments, rather than using a commercial translation

memory, we opted to use our own algorithm, which gave us better control over the calculation of similarities and coverage [4]. The results are provided in Table 4. The coverage rate is the percentage of text found in the translation memory, measured in segments or in words. **To get an accurate picture of the translation memory's impact, it is better to measure the coverage in terms of words**. Take for example, the domain 313 test, with a large translation memory (25M words): for over 35% of the segments there was a matching segment with a similarity of 0.85 or greater; however, these segments represented only 15% of the words in the text. In other words, repetition is especially relevant in significantly shorter than average segments.

Domain		1M Words (Small)			5M Words (Medium)			25M Words (Large)		
		≥ 0.70	≥ 0.85	Exact	≥ 0.70	≥ 0.85	Exact	≥ 0.70	≥ 0.85	Exact
313 (poor)	words	7.7	5.6	2.8	11.7	9.2	5.4	17.4	15	10.6
	segments	17.6	13.2	8.8	28.7	22.4	15.9	42.4	35.9	28.4
349 (average)	words	2.5	1.9	1.5	4.6	3.6	2.8	11.2	9.2	7.4
	segments	8.8	7.4	6.8	15.5	13.1	11.8	31.0	26.1	23.5
44 (rich)	words	0.7	0.5	0.5	1.7	1.3	0.9	5.4	4.1	3.4
	segments	5.6	4.8	4.7	10.6	9.0	8.4	21.2	16.7	15.6

Table 4: Percentage of covered by translation memory, measured in words and segments, for various domains and degrees of similarity

Here again, there is a fairly clear correlation between lexical richness and repetition: the more lexically rich the domain, the less repetitive it is and the less useful a translation memory is. The coverage rates (measured in words) obtained with the small domain 313 translation memory (1M words) are comparable to those obtained with the domain 44 translation memory, which was 25 times larger.

Moreover, it is interesting that, even under the best conditions (domain 313, 25M word translation memory), the translation memory suggested matches greater than or equal to 0.70 for only 17.4% of the text.

6. Conclusions

We have conducted a preliminary review of the collection of texts in the Bureau's translation archives. Other than the measures relating to the size of the various parts of the archives, we calculated various lexical richness indicators for each domain (client) in the corpus. We then selected three exemplary domains, for which we created English-French machine translation systems, and estimated the expected quality of the translations, using the BLEU metric. We also measured the extent to which using a standard translation memory could be useful for each of these domains.

In these experiments, we maintain that the measures of lexical richness, such as type-occurrence ratio (TTR), which is simply and quickly calculated, are effective in expeditiously identifying promising corpora for machine translation. In practice, lexical richness, the size of the sub-corpora and repetition (as measured by the coverage of a translation memory) seem to all establish clear correlations to machine translation quality (measured using the BLEU score). These finding, however, are based on too little

observation to make categorical generalizations. To do so, we will need to repeat our experiments in several other domains. However, we believe that a prognostic tool could be developed to quickly and inexpensively evaluate the potential for machine translation for a given set of texts.

While the quality of the translations of the most complex texts (domains 44 and 349) seemed questionable, it seems probable that very useful results could be achieved with simpler and more repetitive domains. **Domain 313 (Immigration and Refugee Board of Canada) seems to be a good domain in which to apply machine translation as a tool to help translators**. We believe that further experiments in domains with similar profiles could help us identify other domains with promising potential for machine translation.

Lastly, we suggested that certain sub-corpora (e.g. 44) were too heterogeneous to produce good results using SMT methods. However, within a sub-corpus that large, it is possible to identify subsets sufficiently homogeneous and large to be considered good candidates for machine translation. We believe that automatic clustering methods could be used to identify such subsets. This theory could be the subject of a future study.

References

[1] Kittredge, Richard & John Lehrberger. (1982.) *Sublanguage: Studies of language in restricted semantic domains*. Berlin: Walter de Gruyter.

[2] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation* (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318. *CiteSeerX*: *10.1.1.19.9416*

[3] *https://en.wikipedia.org/wiki/BLEU*

[4] Simard, M., & Fujita, A. (2012, October). A Poor Man's Translation Memory Using Machine Translation Evaluation Metrics. In Proceedings of the 10th Biennial Conference of AMTA.

[5] Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5), 323-352.

Annex – Sample machine translations

Notes

- These samples were randomly selected from the test set: source language segments were first sorted alphabetically and, then, a block of 10 successive segments was selected.
- All the translations were done in lower case, which is why they appear as such below.

Sample 1: Domain 44, large system

Source text	Reference translation	Machine translation
enjoy working with young children	vous aimez travailler avec de jeunes enfants	aime travailler avec les jeunes enfants
ensure that any information that they make available to the public is up-to-date , released within the law (privacy , security of information act) , and respectful of the principle of operational security ;	s' assurer de diffuser au public des renseignements à jour , dans le cadre de la loi (loi sur la protection des renseignements personnels , loi sur la protection de l' information) et respectueux du principe de sécurité opérationnelle ;	s' assurer que toute l' information qu' ils mettent à la disposition du public est à jour , qu' elle est diffusée dans les limites imposées par la loi (loi sur la protection des renseignements personnels , la sécurité de l' information , et en observant le principe de la sécurité opérationnelle ;
enter the quantity of the unit of measure as per the selected purchase unit on the form in the first row of the purchase/recipe unit conversion table .	entrez sur la première ligne de la table de conversion d' unité d' achat/recette la quantité de l' unité de mesure en fonction de l' unité d' achat sélectionnée dans le formulaire .	entrez la quantité de l' unité de mesure conformément à l' unité d' achat sélectionnée sur le formulaire dans la première rangée du tableau de conversion de l' unité purchase/recipe .
enterprise information technology service management	gestion des services en technologie de l' information d' entreprise	gestion des services en technologie de l' information d' entreprise
environment : day or night in any weather condition , and in the full spectrum of operations	environnement : jour et nuit , toute condition météorologique , plein spectre des opérations .	environnement : de jour ou de nuit , dans toutes les conditions météorologiques et dans l' ensemble du spectre des opérations
environmental conditions .	conditions environnementales .	les conditions environnementales .
environmental conditions and available light sources combine to enhance or degrade nvg performance .	la combinaison des conditions environnementales et des sources lumineuses disponibles améliore ou détériore le rendement des lvn .	les conditions environnementales et les sources de lumière à la disposition de combiner , d' améliorer ou de diminuer la performance des lvn .
environmental scan	analyse de la conjoncture	analyse environnementale
equipment acquisition and disposal	acquisition et disposition d' équipement	l' acquisition et l' aliénation de l' équipement
equipment distributed to	équipement distribué à	équipement distribué à

Sample 2: Domain 313, large system

Source text	Reference translation	Machine translation
m. drukarsh – there is an issue of perception when pc/counsel enter room with member and minister 's counsel already there .	m. drukarsh – la question de perception est soulevée lorsque l' intéressé/son conseil entre dans la pièce et que le commissaire et le conseil du ministre y sont déjà .	m. drukarsh – il y a une question de perception lorsque la salle avec le commissaire et le conseil du ministre pc/counsel entrer déjà là .
made in the hearing room to the coordinating member	présentées en salle d' audience au commissaire coordonnateur	dans la salle d' audience pour le commissaire coordonnateur
make reasonable efforts to maintain yourself in such condition that :	faire des efforts raisonnables pour se maintenir dans un état tel :	faire des efforts raisonnables pour se maintenir dans un état tel :
make reasonable efforts to seek and maintain full time employment and immediately report any change in employment to the department .	faire des efforts raisonnables pour obtenir un emploi à temps plein et le conserver , et signaler sans délai tout changement d' emploi à l' agence .	faire des efforts raisonnables pour obtenir un emploi à temps plein et le conserver , et signaler sans délai tout changement d' emploi à l' agence .
many of the discrepancies noted by minister 's counsel fall under a category of what has elsewhere been described as involving a microscopic examination of the evidence .	un grand nombre des divergences signalées par le conseil du ministre relèvent de ce qui a été décrit ailleurs comme un examen microscopique de la preuve .	bon nombre des divergences relevées par le conseil du ministre relève une catégorie de ce qui a été décrit à l' étranger comme impliquant un examen des moindres détails des éléments de preuve .
march , 2006 espinoza investigated by pc careton with another male who (also) was later identified as a la raza member (manix 8154455)	mars 2006 – espinoza et un autre individu qui a (aussi) été plus tard désigné comme membre de la raza ont fait l' objet d' une enquête menée par le capitaine de police careton (manix 8154455)	mars 2006 espinoza enquête par intéressé careton avec un autre homme qui (aussi) a par la suite été identifiée comme la commissaire raza (manix 8154455)
marital status and family composition	état matrimonial et composition de la famille	état matrimonial et composition de la famille .
marital status including common-law relationships ; and	son état matrimonial , y compris ses relations de fait ;	son état matrimonial , y compris ses relations de fait ;
materials :	documentation	documents :
materials establish that the appellant was convicted of a criminal offence in march 2010 as a result of his activities in august 2009 , that is , in the period following the iad 's grant of the stay of removal .	les documents établissent que l' appelant a été déclaré coupable d' une infraction criminelle en mars 2010 en raison d' actes posés en août 2009 , c' est-à-dire après que la sai a prononcé un sursis de la mesure de renvoi .	des documents établissent que l' appelant a été déclaré coupable d' une infraction criminelle en mars 2010 à la suite de ses activités en août 2009 , c' est-à-dire dans la période qui a suivi l' octroi de la sai du sursis à la mesure de renvoi .

Sample 3: Domain 349, large system

Source text	Reference translation	Machine translation
some employees may have little or no garbage of this kind because they choose to eat in the cafeteria , or drink coffee out of a plastic mug , or bring their food in reusable containers .	certains employés ont peu ou pas de déchets parce qu' ils choisissent de manger à la cafétéria ou ont leur propre tasse à café ou qu' ils apportent leur nourriture dans des contenants réutilisables .	certains employés ont peu ou pas de déchets de ce genre parce qu' ils choisissent de manger dans la cafétéria , buvez un café de tasse de plastique , ou apporter leur nourriture dans des contenants réutilisables .
some key statistics	quelques statistiques	certaines statistiques clés
some may prefer to use an ssa , which is also acceptable .	certains préfèrent cependant recourir à une cps , ce qui convient également .	certains membres pourraient préférer utiliser une convention particulière de services , ce qui est également acceptable .
some work may be done at other government of canada locations across the greater toronto area .	certains travaux pourraient devoir être exécutés dans d' autres bureaux du gouvernement du canada dans la région du grand toronto .	certains travaux pourraient être effectués à d' autres emplacements du gouvernement du canada dans l' ensemble de la région du grand toronto .
sonia drapeau for felicity mulgan , csps	sonia drapeau pour felicity mulgan , efpc	sonia drapeau pour felicity mulgan , efpc
sources : centre for disease control and prevention : national health interview survey (nhis) , mediamark research inc. and statistics canada : residential telephone service survey (rtss)	sources d' information : centre for disease control and prevention , national health interview survey (nhis) , mediamark research inc. , et statistique canada , enquête sur le service téléphonique résidentiel (estr)	sources : centre for disease control and prevention : national health interview survey (nhis) , mediamark research inc. et statistique canada : enquête sur le service téléphonique résidentiel (estr)
sources : statistics canada : residential telephone service survey (rtss)	source : statistique canada , enquête sur le service téléphonique résidentiel (estr)	sources : statistique canada : enquête sur le service téléphonique résidentiel (estr)
sow s. 6.11 : the field work is required to commence 3 weeks of the kick-off meeting .	section 6.11 de l' énoncé des travaux : le travail sur le terrain doit commencer trois semaines après la réunion de démarrage .	article 6.11 de l' edt : le travail sur le terrain est nécessaire pour commencer les trois semaines suivant la réunion de lancement .
spar	spsar	spsar
spar / pm / apm	spsar / sgp / sglpsi	spsar / gp / glp

Étude du corpus de traductions du BtB en vue de son utilisation pour la traduction automatique

Pierre Isabelle
Michel Simard
Traitement Multilingue de Textes
Conseil National de Recherches Canada

1. Introduction

Dans le cadre d'une collaboration en cours sur la traduction automatique (TA), le Bureau de la Traduction du gouvernement du Canada (BtB) a confié au CNRC une copie du vaste corpus qu'il a constitué à partir de ses traductions faites antérieurement. Le CNRC s'est engagé à conduire une étude sur les meilleures façons d'exploiter ce corpus en traduction automatique. Le présent document constitue un rapport préliminaire sur les activités conduites par le CNRC dans le cadre de cette étude, et sur les conclusions provisoires qui en émergent.

Le corpus reçu du BtB (ci-après nommé « corpus BtB ») est organisé sous forme de « bi-texte » qui rend explicites non seulement les correspondances globales entre des paires de documents mais aussi les correspondances plus fines entre les phrases contenues dans ces paires de documents. Le BtB exploite déjà ce corpus pour offrir une « mémoire de traduction » (MT) à ses traducteurs. Ceux-ci disposent maintenant d'outils qui leur permettent de fouiller cette mémoire pour y retrouver rapidement des traductions qui auraient été faites antérieurement pour les éléments qu'ils s'apprêtent à traduire ou pour des éléments similaires.

Or, les systèmes de traduction automatique statistique (TAS) se nourrissent exactement des mêmes bi-textes que les MT. Ceci n'est pas surprenant dans la mesure où on peut en fait voir les MT comme le niveau zéro de la TAS : un système de MT offre comme (ébauche de) traduction d'une phrase d'entrée la traduction de la phrase la plus semblable ayant été traduite antérieurement. Généralement, on ne fournit une telle sortie que lorsque l'on a réussi à trouver une phrase déjà traduite qui est suffisamment similaire à la phrase à traduire (disons 70% de recouvrement ou plus). Par conséquent, les MT sont surtout utiles dans les cas où on traduit régulièrement des documents très similaires, comme par exemple des versions différentes du même document. Les MT sont beaucoup moins utiles dans les domaines où le contenu textuel varie au jour le jour, comme dans la presse écrite ou dans le Journal des débats de la Chambre des communes.

Les systèmes de TAS, contrairement aux MT, sont capables de décomposer automatiquement les correspondances entre phrases en correspondances plus fines et de recomposer automatiquement une traduction de phrase à partir d'éléments puisés dans plusieurs paires de phrases différentes de la

mémoire. Ceci leur permet en principe de s'avérer plus utile que les MT dans les domaines où on observe peu de répétition de phrases.

Malheureusement, jusqu'à récemment, la qualité des traductions machine s'était rarement avérée suffisante pour qu'un traducteur puisse en tirer un réel profit. Le succès bien connu de la TA sur les prévisions météo a longtemps constitué l'exception qui confirmait cette règle. L'expérience a montré qu'il existe très peu de domaines où l'on produit de grandes quantités qui sont d'un niveau de simplicité comparable à celui des prévisions météo et qui doivent être traduits.

Ceci dit, ces dernières années les approches statistiques ont permis à la traduction automatique de connaître des progrès très substantiels : la qualité moyenne des traductions machine s'est beaucoup améliorée. Il est maintenant possible d'obtenir des traductions machine relativement bonnes pour des textes qui sont nettement plus complexes que les prévisions météo. C'est pour cette raison qu'un bon nombre de cabinets de traduction au Canada et à l'étranger se sont déjà empressés d'adopter la TAS comme un élément important de la nouvelle boîte à outils du traducteur. Sur la scène canadienne, PORTAGE est à l'avant-garde de ce mouvement : il est utilisé quotidiennement par un nombre de plus en plus substantiel de traducteurs depuis 2010.

Le moment semble donc opportun pour le BtB d'examiner de très près le potentiel de la TAS dans ses opérations. Pour une organisation comme le BtB, la prudence demeure toutefois de mise. Le fait que la TA ait progressé suffisamment pour avoir trouvé des preneurs ne signifie évidemment pas qu'elle soit maintenant applicable de manière universelle. Il paraîtrait bien avisé de chercher d'abord à l'appliquer aux situations les plus propices, puis, en cas de succès, d'étendre la couverture à des situations de plus en plus exigeantes. La question cruciale qui se pose alors est la suivante: par où devrait-on commencer?

2. Démarche générale

Comment identifier les situations les plus propices à la TA? On notera à ce sujet qu'il existe au moins deux ordres de considérations différents : 1) la possibilité technique d'obtenir des traductions machine raisonnablement bonnes; et 2) la possibilité de faire une implantation réussie de nouvelles technologies au sein de d'une organisation particulière. Bien que ce second point soit tout aussi important que le premier, il demeure hors du champ de la présente étude. Notre attention se limitera ici à la question de savoir comment choisir les domaines d'application sur lesquels la machine devrait produire des traductions de d'une qualité relativement bonne.

Cette problématique n'est pas nouvelle sur la scène de la recherche au Canada. Après le succès de la TA des prévisions météo à la fin des années 70, les chercheurs canadiens se sont interrogés sur la possibilité de reproduire ce succès dans d'autres domaines. Une petite communauté de chercheurs en linguistique s'est alors tournée vers l'étude des *sous-langues*, c'est-à-dire de sous-ensembles relativement restreints et fermés d'une langue particulière utilisés par une communauté particulière de rédacteurs pour la production de textes standardisés visant une fin très spécifique [1]. Il est facile de trouver d'autres exemples que celui des prévisions météo: les rapports de marchés boursiers, les descriptions de

matches sportifs, les descriptions de tâches pour les postes au sein d'un organisme gouvernemental, etc.

Les recherches des années 80 sur cette notion de sous-langue reposaient sur des études essentiellement manuelles de la complexité linguistique relative des divers types de textes à différents niveaux : vocabulaire, constructions syntaxiques, espace sémantique, etc. Or, le corpus BtB est d'une ampleur telle que les études manuelles exhaustives de ce type sont complètement hors de portée. On parle d'environ 2.5 milliards de mots répartis entre 174 « domaines » différents. À strictement parler, il s'agit plutôt de clients (à qui la facture est envoyée) que de domaines de contenu textuel. Le Tableau 1 ci-dessous donne quelques statistiques de base pour les domaines les plus volumineux du corpus.

	EN-FR				FR-EN			
Domaine	**doc.**	**seg.**	**mots EN**	**mots FR**	**doc.**	**seg.**	**mots EN**	**mots FR**
44	111326	9.51 M	113.10 M	138.59 M	26168	2.30 M	28.15 M	33.87 M
349	80497	5.79 M	72.39 M	89.59 M	10276	786.72 K	10.54 M	12.45 M
313	23204	2.23 M	44.21 M	49.10 M	2287	183.41 K	3.65 M	3.96 M
47	43166	2.86 M	38.90 M	48.37 M	6628	465.76 K	6.51 M	7.86 M
10	29310	2.90 M	38.33 M	47.31 M	2991	287.67 K	3.96 M	4.65 M
999	44707	2.62 M	37.45 M	49.28 M	3226	198.74 K	3.30 M	4.29 M
40	32628	2.34 M	32.23 M	39.78 M	10391	774.90 K	11.41 M	13.37 M
75	29595	2.33 M	31.96 M	39.02 M	9039	671.29 K	9.53 M	11.64 M
32	33654	2.13 M	31.68 M	38.80 M	3466	221.49 K	3.58 M	4.26 M
100	36065	2.10 M	31.42 M	38.86 M	3644	230.66 K	3.74 M	4.41 M
146	32226	2.26 M	30.79 M	38.92 M	4718	413.48 K	6.02 M	7.47 M
128	18104	1.70 M	22.89 M	27.93 M	3267	257.05 K	3.71 M	4.31 M
333	25158	1.83 M	22.76 M	27.98 M	6548	389.28 K	5.25 M	6.35 M
555	8674	1.08 M	20.87 M	23.88 M	2980	309.84 K	6.35 M	7.03 M
1	15333	1.36 M	19.88 M	24.46 M	4977	534.60 K	8.01 M	9.65 M
69	22944	1.35 M	19.35 M	23.96 M	2543	167.22 K	2.59 M	3.14 M
330	24114	1.31 M	18.50 M	22.59 M	2733	171.21 K	2.63 M	3.13 M
23	9986	1.24 M	18.14 M	22.22 M	2471	278.95 K	4.02 M	4.70 M
471	19792	1.14 M	16.30 M	19.80 M	3185	171.37 K	2.53 M	3.04 M
227	18140	1.18 M	16.22 M	20.00 M	2141	156.59 K	2.18 M	2.58 M

Tableau 1: Statistiques sur la taille des domaines (clients) les plus volumineux

On note que la taille des différents sous-corpus est hautement variable : de quelques centaines de mots (e.g. « Centre internationale des maladies infectieuses » - 612) à quelques centaines de millions de mots (« Défense nationale» - 44).

À l'examen, on trouve bien entendu des liens entre les différents clients et les thématiques prédominantes. Dans le meilleur cas, on aura un contenu (presque) parfaitement homogène. Par exemple, le sous-corpus associé au client « Conseil canadien des relations industrielles » (184) paraît contenir presque uniquement les décisions de ce tribunal administratif sur les relations de travail. Le terme de sous-langue paraîtrait approprié dans de tels cas.

Par contre, les sous-corpus associés à d'autres clients ont un contenu nettement plus hétérogène. Par exemple, en extrayant au hasard quelques documents du sous-corpus le plus volumineux, celui du client « Défense nationale» (44), on obtient des documents qui semblent à chaque fois de nature différente : descriptions d'exercices d'entraînement, descriptions de tâches, questions d'évaluation pour entrevues d'embauche, chapitres d'un livre sur l'histoire de la guerre de 1914, manuel technique d'avionique, procédures de préparation du budget, etc. Il est probable toutefois qu'il existe certains grands sous-ensembles homogènes à l'intérieur de ce méga-corpus de la Défense. Mais il ne nous est pas possible de mettre au jour cette structure par des méthodes manuelles.

Puisqu'il s'avère impossible de balayer manuellement l'ensemble du corpus BtB, nous nous tournons donc vers des méthodes automatiques. **Dans un premier temps, nous allons effectuer des comparaisons entre les sous-corpus « clients » existants**. Dans une phase ultérieure de notre étude, nous espérons pouvoir aller au-delà de cette structure de manière à pouvoir identifier : a) des sous-corpus homogènes à l'intérieur des corpus clients existants; et/ou b) des regroupements homogènes de tels sous-corpus à travers des clients différents.

Les comparaisons que nous allons faire entre différents sous-corpus relèveront de deux ordres : 1) complexité linguistique relative de ces sous-corpus, telle qu'évaluée par des mécanismes automatiques; et 2) qualité relative des traductions machine obtenues sur différents sous-corpus, telle qu'évaluée par des mécanismes automatiques. La faisabilité du second volet peut surprendre. Il est clair qu'à l'époque des études de sous-langues mentionnées ci-dessus, une telle démarche était complètement exclue. La construction de chaque système de TA spécialisé pour un domaine particulier exigeait plusieurs années de travail par une équipe experte en linguistique informatique. Le but des études de complexité linguistique était justement de servir de prédicteur du potentiel relatif des différents domaines pour la traduction automatique : il fallait à tout prix éviter d'investir des années d'efforts dans le développement de systèmes condamnés à l'avance par la difficulté trop grande du problème.

Or, la situation a bien changé avec l'avènement de TAS : le développement de systèmes spécialisés s'effectue de manière largement automatique, et il existe également des façons automatiques d'estimer la qualité de leurs traductions. Le lecteur pourra alors se demander pourquoi nous n'abandonnons pas complètement le volet d'évaluation de la complexité linguistique. Ne vaudrait-il pas mieux se contenter d'évaluer la qualité relative des traductions machine dans les différents domaines puisque le résultat souhaité est précisément d'identifier les domaines où on obtiendra les meilleures traductions?

La réponse est la suivante : bien que l'évaluation de la qualité des traductions machine puisse se faire de façon largement automatique, cette démarche exige tout de même des ressources qui sont loin d'être négligeables. Elle suppose en effet que l'on génère des modèles PORTAGE spécifiques pour chacun des domaines que l'on souhaite comparer. Ceci signifie passablement de travaux de préparation de sous-corpus et beaucoup de temps de calcul pour construire le modèle PORTAGE correspondant.

Il nous paraît vraisemblable qu'il existe une forte corrélation entre les mesures de complexité linguistique relative des différents domaines et les mesures de qualité relative des traductions machine

obtenues pour les mêmes domaines. Si c'est le cas, on pourra utiliser les premières comme un substitut peu coûteux aux secondes.

3. Mesures de richesse lexicale sur différents sous-corpus

Comme nous l'avons mentionné plus haut, les études de sous-langues des années 80 se penchaient sur des analyses de complexité linguistique à différents niveaux : lexique, syntaxe, espace sémantique. Or, de ces niveaux, seul celui du lexique se prête facilement à des études automatiques. Il est en effet relativement facile de comparer les domaines entre eux du point de vue de la taille des vocabulaires en jeu. Le vocabulaire d'un sous-corpus est directement observable : quelques règles relativement simples de segmentation d'un texte en mots isolés et des algorithmes simples de tri et de comptage suffisent à la tâche.

Il s'avère beaucoup plus difficile d'effectuer des analyses automatiques de la complexité syntaxique ou sémantique. Les structures syntaxiques et sémantiques ne sont pas directement observables dans les textes. Pour en évaluer la richesse ou la complexité, il faut d'abord les rendre observables. Dans le cas de la syntaxe, ceci suppose la disponibilité de parseurs capables d'attribuer automatiquement à chaque phrase de chaque texte étudié une structure grammaticale explicite (généralement on parle d'un « arbre syntaxique »). Or il n'existe pas à l'heure actuelle de parseurs suffisamment robustes, précis et rapides pour effectuer cette tâche de manière fiable sur un corpus aussi vaste et hétérogène que celui qui nous concerne ici. Et la situation est encore bien pire si on parle d'évaluer la complexité sémantique : cette notion demeure en effet très mal comprise.

Par conséquent, nos analyses linguistiques automatiques vont se limiter à celles qui concernent la richesse du vocabulaire, et nous allons faire l'hypothèse que la complexité linguistique tend à s'accorder entre les différents niveaux. Ainsi, il est très naturel de penser que plus le vocabulaire d'un domaine est vaste, plus son espace sémantique est complexe.

Différentes mesures en lien avec la taille du vocabulaire peuvent servir d'indicateurs de la complexité relative d'un domaine textuel (en l'occurrence, des archives d'un client). L'indice de richesse lexicale le plus connu est sans doute le *ratio type-occurrence* (RTO): cet indice met en rapport la taille du vocabulaire (le nombre de formes de mots distinctes) et la taille du corpus complet (le nombre d'occurrences de mots). Par exemple, dans la phrase:

To be, or not to be, that is the question

on observe 10 occurrences de mots, mais seulement 8 formes différentes: *be, is, not, or, question, that, the, to*. Le rapport des formes sur les occurrences est donc de RTO = 8/10 = 0.8. Évidemment, calculer ce rapport sur une seule phrase ne présente pas grand intérêt. En revanche, en comparant les mesures obtenues sur de plus grandes quantités de texte (par exemple quelques milliers de phrases) de différents domaines, on peut se faire une idée assez précise de la richesse lexicale relative de ces domaines.

Une mesure alternative est le *taux de croissance* du vocabulaire, *P(N)*, qu'on calcule comme le nombre de mots qui n'apparaissent qu'une fois dans le corpus, divisé par la taille du corpus. *P(N)* peut être interprété comme la probabilité que le $(N+1)^{\text{ième}}$ mot d'un corpus donné soit nouveau (c'est-à-dire qu'il n'ait jamais été observé auparavant dans le corpus). Dans l'exemple ci-dessus, 6 des 8 formes de mots observées n'apparaissent qu'une fois dans la phrase de 10 mots: on en conclut que si on devait observer un $11^{\text{ième}}$ mot, celui-ci aurait 6 chances sur 10 d'être nouveau. (Le mot suivant dans le célèbre monologue d'Hamlet est *Whether*, qui s'avère effectivement être nouveau dans ce contexte.)

On se convaincra aisément que ces deux indices varient avec la taille de l'échantillon pris en considération (en pratique, ils diminuent avec la taille du corpus). Cette caractéristique est problématique lorsque le but est de comparer entre eux des domaines de tailles différentes, comme c'est le cas ici. Certains chercheurs se sont penchés sur ce problème, et ont tenté de mettre au point des indices alternatifs, qui soient moins sensibles à la taille du corpus. Tweedie & Baayen [5] passent en revue un grand nombre de ces indices, afin de comparer leur comportement.

Pour évaluer le potentiel de la TA dans différents domaines, comme on souhaite le faire ici, il existe toutefois une solution plus simple à ce problème: mesurer l'indice de richesse lexicale sur des échantillons aléatoires de même taille, quelle que soit la taille totale des domaines. C'est ce que nous avons fait avec les données du corpus BtB: pour chaque domaine du corpus, nous avons sélectionné au hasard un nombre suffisant de documents pour cumuler **un million de mots**; à partir de cet échantillon, nous avons calculé les valeurs de plusieurs des indices proposés dans l'article de Tweedie & Baayen. Nous rapportons dans le tableau ci-dessous quelques-unes de ces valeurs, pour les domaines les plus volumineux du corpus.

Domaine	N	V	V1	RTO	P(N)	R	C	k	a2
313	1000470	16676	6314	0.016668	0.006311	0.16672	0.703658	0.370235	0.214492
999	1001069	17339	5970	0.01732	0.005964	0.1733	0.706449	0.371714	0.212463
32	1000022	23690	8695	0.023689	0.008695	0.2369	0.729093	0.38361	0.196089
555	1000429	23747	8398	0.023737	0.008394	0.23742	0.729245	0.383697	0.195973
349	1002717	24485	9209	0.024419	0.009184	0.24452	0.73134	0.384839	0.194425
40	1004232	24686	9419	0.024582	0.009379	0.24634	0.731851	0.385134	0.194033
330	1000011	24765	8528	0.024765	0.008528	0.24765	0.732306	0.3853	0.193763
128	1008359	25119	9232	0.024911	0.009155	0.25015	0.732892	0.385752	0.193223
10	1000499	25579	9644	0.025566	0.009639	0.25573	0.734621	0.386527	0.192081
75	1000364	27115	10505	0.027105	0.010501	0.2711	0.738849	0.388749	0.189023
146	1000553	27576	10292	0.027561	0.010286	0.27568	0.740059	0.389389	0.188144

69	1002497	27698	11390	0.027629	0.011362	0.27663	0.740274	0.389536	0.187962
47	1002664	28748	10653	0.028672	0.010625	0.2871	0.742958	0.390951	0.186017
227	1002163	28822	10705	0.02876	0.010682	0.28791	0.743171	0.391055	0.18587
333	1000008	29601	11466	0.029601	0.011466	0.29601	0.745217	0.392094	0.184418
23	1000410	30716	12047	0.030703	0.012042	0.3071	0.747872	0.393497	0.182491
44	1000162	31243	11241	0.031238	0.011239	0.3124	0.749117	0.394148	0.181593

Tableau 2: Richesse lexicale des domaines les plus volumineux; échantillons de 1M de mots; triés par ordre croissant de RTO

À l'examen, on observe que tous ces indices sont très corrélés, c'est-à-dire que même si leurs valeurs ne sont pas directement comparables, l'ordre relatif qu'ils imposent aux domaines sont à peu près les mêmes. Le domaine qui obtient le *RTO* le plus faible (313) obtient également les *R* et *C* les plus faibles, et le *a2* le plus élevé, et inversement pour le *RTO* le plus élevé (domaine 44). L'indice P(N) impose quant à lui un ordre légèrement différent.

Pour la suite de cette étude, nous allons concentrer notre attention sur les trois domaines les plus volumineux du corpus, à savoir le 44 (Défense nationale), le 313 (Commission de l'immigration et du statut de réfugié du Canada) et le 349 (Travaux publics et Services gouvernementaux Canada).

4. Mesures de qualité des traductions machine sur différents sous-corpus

Les méthodes mises au point par la communauté de recherche en TAS nous permettent d'envisager une possibilité qui était impensable à l'époque des études de sous-langues mentionnées ci-dessus : celle de construire des systèmes de TAS spécifiques pour chacun d'un certain nombre de domaines différents, puis de mesurer leur performance relative. Cette possibilité découle du fait que **la construction et l'évaluation des systèmes peut se faire de manière largement automatique**. Dans la mesure où cette démarche réussit, il devient dès lors possible de sélectionner les domaines où la TA fonctionne le mieux avant de lancer des projets d'application.

Pour construire un système de TAS pour un domaine spécifique, on doit séparer le corpus en deux jeux de données : données dites d'entraînement, et données de test. Ensuite on fournit au module d'apprentissage de PORTAGE le premier jeu (dans chaque cas un ensemble de phrase en langue source avec chacune sa traduction correcte), puis on laisse PORTAGE apprendre automatiquement un modèle pour le domaine concerné.

Ce qui peut paraître plus surprenant, c'est qu'une fois l'étape de construction automatique terminée, on pourra procéder à une **évaluation automatique de la qualité des traductions** obtenues. À cette fin, on utilise une ou plusieurs des métriques d'évaluation automatique mises au point dans les années récentes par les chercheurs en TAS. La plus connue de ces métriques est le score BLEU.

4.1 Le score BLEU

Ce score BLEU[2] n'est rien de plus qu'une mesure du chevauchement entre la traduction produite par la machine et une ou plusieurs traductions humaines de référence. Le chevauchement se mesure par le nombre de *n*-grammes partagés. Par *n*-gramme on entend une suite continue de *n* mots. Par exemple, la phrase « le petit chat lappe le lait » comprend les unigrammes en (1), les bigrammes en (2), les trigrammes en (3) et les quadrigrammes en (4) :

(1) le, petit, chat, lappe, lait.
(2) le petit, petit chat, chat lappe, lappe le, le lait.
(3) le petit chat, petit chat lappe, chat lappe le, lappe le lait.
(4) le petit chat lappe, petit chat lappe le, chat lappe le lait.

Il s'agit donc de compter la proportion des *n*-grammes de la traduction machine qui apparaissent aussi dans une ou plusieurs traductions de référence. La plupart du temps, on prend en compte l'ensemble formé par tous les n-grammes pour $1 \leq n \leq 4$. La formule complète incorpore certains critères additionnels et le lecteur intéressé pourra consulter [2] ou encore l'entrée pertinente de Wikipédia [3].

Dans notre cas, il n'y aura généralement qu'une seule traduction de référence, à savoir celle provenant de ce que nous avons nommé ci-dessus le jeu de test extrait du corpus BtB. On fera donc traduire la partie source du jeu de test par la machine et on calculera le score BLEU de la traduction relativement à la traduction de référence que constitue la partie cible du même jeu de test.

Manifestement, ce score BLEU ne fait rien de plus que d'approximer de façon plutôt grossière la qualité des traductions machine. Cette métrique ignore complètement le fait que la (ou les) traduction de référence ne constitue pas nécessairement la seule bonne traduction possible. Toutefois, il est maintenant bien établi que le classement effectué par BLEU sur la qualité relative des traductions du même texte par différents systèmes de TA a une très forte corrélation avec le classement qui serait obtenu par un panel d'experts humains. Il n'est nullement question d'utiliser BLEU pour évaluer des traductions humaines. Mais la valeur de cette métrique pour établir un ranking de traductions machine est maintenant reconnue.

On notera toutefois que le problème que nous abordons dans cette étude est quelque peu différent de celui que nous venons de mentionner : pour nous, il ne s'agit pas de classer le niveau de qualité relatif obtenu par différents systèmes sur les mêmes textes mais plutôt de classer le niveau de difficulté relative de différents types de textes pour une même machine.

Nous adoptons l'hypothèse suivante, qui nous paraît très plausible : le score BLEU obtenu sur différents domaines est inversement proportionnel au degré de difficulté de ces domaines pour la TAS. Certaines précautions s'imposent toutefois. On ne pourra comparer directement que des systèmes construits avec des jeux d'apprentissage de taille égale. Et, bien entendu, le jeu de test devra être choisi de manière à assurer qu'il soit représentatif de l'ensemble des textes du domaine concerné.

4.2 Expériences de traduction automatique

D'une façon générale, on s'attend à ce que la richesse lexicale d'un domaine soit inversement corrélée avec sa *traduisibilité automatique*: plus le vocabulaire est restreint, plus il sera facile de traduire automatiquement les textes du domaine, et inversement.

Pour vérifier cette hypothèse, nous avons construit des systèmes de TAS pour les trois domaines les plus volumineux du corpus. Par un hasard heureux, ces trois domaines présentent des profils très différents en termes de richesse lexicale. Comme on l'a souligné plus tôt, le domaine 44 est non seulement le plus volumineux, c'est aussi un des plus variés sur le plan lexical, et il obtient des indices RTO et P(N) parmi les plus élevés. À l'opposé, le domaine 313 est parmi ceux pour lesquels ces indices sont les plus faibles. Quant au domaine 349, on le retrouve à peu près au milieu de l'échelle. Pour résumer, on peut dire que, lexicalement parlant, le 313 est *pauvre*, le 349 est *moyen* et le 44 est *riche*.

Comme on le sait, la qualité qu'on peut attendre d'un système de TAS dépend en grande partie de la quantité de données disponible pour l'apprentissage du système. C'est pourquoi, pour chacun des trois domaines retenus, nous avons produit des systèmes à partir de jeux de données de taille croissante:

- « **Petit** »: environ 1 million de mots en langue source
- « **Moyen** »: Environ 5 millions de mots
- « **Grand** »: Environ 25 millions de mots

En pratique, pour chacun des trois domaines, nous avons:

1. Extrait par échantillonnage aléatoire un *jeu de test* d'environ 2000 paires de segments, texte en langue source et traduction
2. Partitionné les données restantes en quatre blocs:
    - bloc 1: environ 1M mots source
    - bloc 2: environ 4M mots source
    - bloc 3: environ 20M mots source
    - bloc 4: le restant
3. Entraîné trois systèmes de TAS anglais-français:
    - « Petit »: avec les données du bloc 1
    - « Moyen »: avec les données des blocs 1 et 2
    - « Grand »: avec les données des blocs 1, 2 et 3
      (nous mettons de côté le bloc 4 pour des expériences futures)
4. Traduit les segments du jeu de test avec chacun de ces systèmes
5. Évalué la qualité de la traduction résultante avec la métrique BLEU

Tous les tests ont été effectués dans la direction anglais → français.

Nous rapportons les résultats obtenus dans le Tableau 3.

Domaine	1M mots (petit)	5M mots (moyen)	25M mots (grand)

	mots src	BLEU	mots src	BLEU	mots src	BLEU
313 (pauvre)	1.01 M	40.3	4.90 M	45.8	25.28 M	50.7
349 (moyen)	1.00 M	28.0	5.26 M	33.5	27.26 M	42.9
44 (riche)	1.24 M	23.2	6.06 M	27.6	29.48 M	33.2

Tableau 3: Scores BLEU produits par les systèmes de traduction automatique.

Il est en général hasardeux de comparer entre eux des scores BLEU obtenus sur des jeux de test différents (rappelons qu'à chaque domaine correspond un jeu de test issu du même domaine). Malgré tout, les différences très marquées que l'on peut observer entre les systèmes produits pour les différents domaines confirment que la richesse lexicale a bien l'impact direct qu'on peut attendre sur la qualité des traductions: moins un domaine est riche lexicalement, plus il est facile à traduire par un système de TA. Et même si la performance obtenue pour chaque domaine augmente très nettement lorsqu'on utilise plus de données d'apprentissage, **la qualité obtenue par le plus petit système pour un domaine lexicalement pauvre (313) reste supérieure à ce qu'on peut espérer avec 25 fois plus de données d'un domaine lexicalement riche (44)**.

Par ailleurs, il n'existe pas de barème précis pour interpréter les scores BLEU. D'une façon générale, pour la traduction entre l'anglais et le français, on remarque que des scores inférieurs à 30 coïncident souvent avec des traductions de qualité médiocre, alors que des scores approchant les 50 vont souvent de pair avec des traductions acceptables. Si on se fie sur cette grille très approximative, les systèmes obtenus pour le domaine 44 ne sont pas très prometteurs, alors que pour le 349, les résultats ne sont intéressants qu'avec beaucoup de données d'apprentissage (25M mots ou plus). À l'opposé, **le domaine 313 donne des résultats très encourageants, même avec aussi peu qu'un million de mots**.

5. Lien avec les mémoires de traduction

Nous l'avons dit, TAS et mémoire de traduction (MT) se nourrissent des mêmes données. En outre, il semble que les domaines, et plus spécifiquement les segments pour lesquels les MT sont productives sont également ceux pour lesquels on peut espérer les meilleures traductions automatiques. En pratique, la MT est bien adaptée aux domaines « répétitifs ». On peut se demander s'il existe un lien entre répétitivité et richesse lexicale. Intuitivement, un domaine lexicalement pauvre est intrinsèquement répétitif: le vocabulaire étant restreint, ce sont donc les mêmes mots qui se répètent! Mais rien ne prouve que cette répétitivité s'étende également aux segments.

Pour voir quel lien existe entre la richesse lexicale et la répétitivité, nous avons soumis les mêmes jeux de test que ci-dessus à une mémoire de traduction et nous avons mesuré pour quelle proportion du texte on pouvait trouver des correspondances exactes ou floues (taux de similarité ≥ 0.70 et ≥ 0.85). Pour ces expériences, plutôt que d'avoir recours à un système de MT commercial, nous avons opté pour un algorithme maison qui nous permet un meilleur contrôle sur le calcul des similarités et des couvertures obtenues [4]. Les résultats obtenus sont rapportés au Tableau 4. Les taux de couverture expriment un pourcentage du texte mesuré soit en nombre de segments, soit en nombre de mots. Nous soumettons ici que **pour se faire une idée juste de l'impact de la MT, il est préférable de mesurer la couverture en termes de nombre de mots**. Par exemple, pour le test du domaine 313, avec une

«grande » MT (25M mots), c'est plus de 35% des segments qui trouvent un segment similaire à 0.85 ou plus; toutefois, ces segments ne représentent que 15% des mots du texte. En d'autres termes, la répétition concerne surtout des segments nettement plus courts que la moyenne.

Domaine		1M mots (petit)			5M mots (moyen)			25M mots (grand)		
		≥ 0.70	≥ 0.85	exact	≥ 0.70	≥ 0.85	exact	≥ 0.70	≥ 0.85	exact
313 (pauvre)	mots	7.7	5.6	2.8	11.7	9.2	5.4	17.4	15	10.6
	segments	17.6	13.2	8.8	28.7	22.4	15.9	42.4	35.9	28.4
349 (moyen	mots	2.5	1.9	1.5	4.6	3.6	2.8	11.2	9.2	7.4
	segments	8.8	7.4	6.8	15.5	13.1	11.8	31.0	26.1	23.5
44 (riche)	mots	0.7	0.5	0.5	1.7	1.3	0.9	5.4	4.1	3.4
	segments	5.6	4.8	4.7	10.6	9.0	8.4	21.2	16.7	15.6

Tableau 4: Pourcentage du texte couvert par la mémoire de traduction, mesuré en termes de mots et de segments, pour différents domaines et taux de similarité

Ici encore, on observe une corrélation assez claire entre richesse lexicale et répétitivité: plus le domaine est riche lexicalement, moins il est répétitif, et donc moins l'utilisation d'une mémoire de traduction sera fructueuse. Les taux de couverture (mesurés en mots) observés avec une petite MT du domaine 313 (1M mots) sont comparables avec ceux qu'on obtient avec une MT 25 fois plus grande du domaine 44.

Par ailleurs, il est intéressant de noter que, même dans les meilleures conditions (domaine 313, MT de 25M mots), la MT ne fournit des propositions de match ≥ 0.70 que pour 17.4% du texte.

6. Conclusions

Nous avons effectué un examen préliminaire de la collection de textes constituant les archives de traduction du BtB. Outre des mesures relatives à la taille des différentes parties de l'archive, nous avons calculé pour chaque domaine (client) du corpus différents indices visant à en mesurer la richesse lexicale. Nous avons ensuite sélectionné trois domaines exemplaires, pour lesquels nous avons produit des systèmes de traduction automatique anglais-français, et estimé la qualité des traductions qu'on pouvait en espérer, au moyen de la métrique automatique BLEU. Nous avons également mesuré dans quelle mesure l'usage d'une mémoire de traduction standard pourrait être fructueux pour chacun de ces domaines.

De ces expériences, nous retenons que les mesures de richesse lexicale telles que le *ratio type-occurrence* (RTO), dont le calcul est simple et rapide, s'avèrent efficace pour identifier rapidement des corpus prometteurs du point de vue de la TA. En pratique, la richesse lexicale, la taille des sous-corpus et la répétitivité des textes (telle que mesurée par la couverture d'une mémoire de traduction) semblent toutes afficher des corrélations claires avec la qualité de la TA (mesurée avec BLEU). Ces conclusions reposent toutefois sur trop peu d'observations pour qu'on puisse encore les généraliser avec assurance. Pour cela, il sera nécessaire de répéter nos expériences sur plusieurs autres domaines. À terme, nous

croyons qu'il serait possible de mettre au point un outil de pronostic qui permettrait d'évaluer rapidement et à peu de frais le potentiel de la TA pour un ensemble de textes donné.

Par ailleurs, alors que la qualité des traductions obtenues pour les textes les plus complexes (44 et 349) paraît incertaine, il semble probable qu'on puisse obtenir des résultats très intéressants pour certains domaines plus simples et répétitifs. **Le domaine 313 (Commission de l'immigration et du statut de réfugié du Canada) apparait comme un bon exemple d'application de la TA comme outil d'aide pour les traducteurs**. Nous croyons que des expériences plus poussées sur des domaines offrant des profils similaires pourront révéler d'autres domaines dont le potentiel à cet égard est prometteur.

Finalement, nous avons suggéré que certains sous-corpus (par exemple, le 44) sont trop hétérogènes pour offrir une bonne prise aux méthodes de TAS. Toutefois, à l'intérieur de sous-corpus aussi volumineux, il est possible que l'on puisse identifier des sous-ensembles suffisamment homogènes et volumineux pour être considérés comme de bons candidats à la TA. Nous pensons qu'il serait possible d'utiliser des méthodes de *partionnement* automatique (en anglais: *clustering*) pour identifier de tels sous-ensembles. Cette question pourrait faire l'objet d'un volet de recherche futur.

Références

[1] Kittredge, Richard & John Lehrberger. (1982.) *Sublanguage: Studies of language in restricted semantic domains*. Berlin: Walter de Gruyter.

[2] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation* (PDF). ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318. *CiteSeerX*: *10.1.1.19.9416*

[3] *https://en.wikipedia.org/wiki/BLEU*

[4] Simard, M., & Fujita, A. (2012, October). A Poor Man's Translation Memory Using Machine Translation Evaluation Metrics. In Proceedings of the 10th Biennial Conference of AMTA.

[5] Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5), 323-352.

Annexe - Exemples de traductions automatiques

Notes:

- Ces exemples ont été choisis aléatoirement dans les jeu de test : les segments en langue source ont d'abord été triés en ordre alphabétique, puis une tranche de 10 segments successifs a été prélevée.
- Toutes les traductions ont été effectuées en casse minuscule, c'est pourquoi elles apparaissent sous cette forme ici.

Exemple 1: Domain 44, « Grand » système

Texte source	Traduction de référence	Traduction automatique
enjoy working with young children	vous aimez travailler avec de jeunes enfants	aime travailler avec les jeunes enfants
ensure that any information that they make available to the public is up-to-date , released within the law (privacy , security of information act) , and respectful of the principle of operational security ;	s' assurer de diffuser au public des renseignements à jour , dans le cadre de la loi (loi sur la protection des renseignements personnels , loi sur la protection de l' information) et respectueux du principe de sécurité opérationnelle ;	s' assurer que toute l' information qu' ils mettent à la disposition du public est à jour , qu' elle est diffusée dans les limites imposées par la loi (loi sur la protection des renseignements personnels , la sécurité de l' information , et en observant le principe de la sécurité opérationnelle ;
enter the quantity of the unit of measure as per the selected purchase unit on the form in the first row of the purchase/recipe unit conversion table .	entrez sur la première ligne de la table de conversion d' unité d' achat/recette la quantité de l' unité de mesure en fonction de l' unité d' achat sélectionnée dans le formulaire .	entrez la quantité de l' unité de mesure conformément à l' unité d' achat sélectionnée sur le formulaire dans la première rangée du tableau de conversion de l' unité purchase/recipe .
enterprise information technology service management	gestion des services en technologie de l' information d' entreprise	gestion des services en technologie de l' information d' entreprise
environment : day or night in any weather condition , and in the full spectrum of operations	environnement : jour et nuit , toute condition météorologique , plein spectre des opérations .	environnement : de jour ou de nuit , dans toutes les conditions météorologiques et dans l' ensemble du spectre des opérations
environmental conditions .	conditions environnementales .	les conditions environnementales .
environmental conditions and available light sources combine to enhance or degrade nvg performance .	la combinaison des conditions environnementales et des sources lumineuses disponibles améliore ou détériore le rendement des lvn .	les conditions environnementales et les sources de lumière à la disposition de combiner , d' améliorer ou de diminuer la performance des lvn .
environmental scan	analyse de la conjoncture	analyse environnementale
equipment acquisition and disposal	acquisition et disposition d' équipement	l' acquisition et l' aliénation de l' équipement
equipment distributed to	équipement distribué à	équipement distribué à

Exemple 2: Domaine 313, « Grand » système

Texte source	Traduction de référence	Traduction automatique
m. drukarsh – there is an issue of perception when pc/counsel enter room with member and minister 's counsel already there .	m. drukarsh – la question de perception est soulevée lorsque l' intéressé/son conseil entre dans la pièce et que le commissaire et le conseil du ministre y sont déjà .	m. drukarsh – il y a une question de perception lorsque la salle avec le commissaire et le conseil du ministre pc/counsel entrer déjà là .
made in the hearing room to the coordinating member	présentées en salle d' audience au commissaire coordonnateur	dans la salle d' audience pour le commissaire coordonnateur
make reasonable efforts to maintain yourself in such condition that :	faire des efforts raisonnables pour se maintenir dans un état tel :	faire des efforts raisonnables pour se maintenir dans un état tel :
make reasonable efforts to seek and maintain full time employment and immediately report any change in employment to the department .	faire des efforts raisonnables pour obtenir un emploi à temps plein et le conserver , et signaler sans délai tout changement d' emploi à l' agence .	faire des efforts raisonnables pour obtenir un emploi à temps plein et le conserver , et signaler sans délai tout changement d' emploi à l' agence .
many of the discrepancies noted by minister 's counsel fall under a category of what has elsewhere been described as involving a microscopic examination of the evidence .	un grand nombre des divergences signalées par le conseil du ministre relèvent de ce qui a été décrit ailleurs comme un examen microscopique de la preuve .	bon nombre des divergences relevées par le conseil du ministre relève une catégorie de ce qui a été décrit à l' étranger comme impliquant un examen des moindres détails des éléments de preuve .
march , 2006 espinoza investigated by pc careton with another male who (also) was later identified as a la raza member (manix 8154455)	mars 2006 – espinoza et un autre individu qui a (aussi) été plus tard désigné comme membre de la raza ont fait l' objet d' une enquête menée par le capitaine de police careton (manix 8154455)	mars 2006 espinoza enquête par intéressé careton avec un autre homme qui (aussi) a par la suite été identifiée comme la commissaire raza (manix 8154455)
marital status and family composition	état matrimonial et composition de la famille	état matrimonial et composition de la famille .
marital status including common-law relationships ; and	son état matrimonial , y compris ses relations de fait ;	son état matrimonial , y compris ses relations de fait ;
materials :	documentation	documents :
materials establish that the appellant was convicted of a criminal offence in march 2010 as a result of his activities in august 2009 , that is , in the period following the iad 's grant of the stay of removal .	les documents établissent que l' appelant a été déclaré coupable d' une infraction criminelle en mars 2010 en raison d' actes posés en août 2009 , c' est-à-dire après que la sai a prononcé un sursis de la mesure de renvoi .	des documents établissent que l' appelant a été déclaré coupable d' une infraction criminelle en mars 2010 à la suite de ses activités en août 2009 , c' est-à-dire dans la période qui a suivi l' octroi de la sai du sursis à la mesure de renvoi .

Exemple 3: Domain 349, « Grand » système

Texte source	Traduction de référence	Traduction automatique
some employees may have little or no garbage of this kind because they choose to eat in the cafeteria , or drink coffee out of a plastic mug , or bring their food in reusable containers .	certains employés ont peu ou pas de déchets parce qu' ils choisissent de manger à la cafétéria ou ont leur propre tasse à café ou qu' ils apportent leur nourriture dans des contenants réutilisables .	certains employés ont peu ou pas de déchets de ce genre parce qu' ils choisissent de manger dans la cafétéria , buvez un café de tasse de plastique , ou apporter leur nourriture dans des contenants réutilisables .
some key statistics	quelques statistiques	certaines statistiques clés
some may prefer to use an ssa , which is also acceptable .	certains préfèrent cependant recourir à une cps , ce qui convient également .	certains membres pourraient préférer utiliser une convention particulière de services , ce qui est également acceptable .
some work may be done at other government of canada locations across the greater toronto area .	certains travaux pourraient devoir être exécutés dans d' autres bureaux du gouvernement du canada dans la région du grand toronto .	certains travaux pourraient être effectués à d' autres emplacements du gouvernement du canada dans l' ensemble de la région du grand toronto .
sonia drapeau for felicity mulgan , csps	sonia drapeau pour felicity mulgan , efpc	sonia drapeau pour felicity mulgan , efpc
sources : centre for disease control and prevention : national health interview survey (nhis) , mediamark research inc. and statistics canada : residential telephone service survey (rtss)	sources d' information : centre for disease control and prevention , national health interview survey (nhis) , mediamark research inc. , et statistique canada , enquête sur le service téléphonique résidentiel (estr)	sources : centre for disease control and prevention : national health interview survey (nhis) , mediamark research inc. et statistique canada : enquête sur le service téléphonique résidentiel (estr)
sources : statistics canada : residential telephone service survey (rtss)	source : statistique canada , enquête sur le service téléphonique résidentiel (estr)	sources : statistique canada : enquête sur le service téléphonique résidentiel (estr)
sow s. 6.11 : the field work is required to commence 3 weeks of the kick-off meeting .	section 6.11 de l' énoncé des travaux : le travail sur le terrain doit commencer trois semaines après la réunion de démarrage .	article 6.11 de l' edt : le travail sur le terrain est nécessaire pour commencer les trois semaines suivant la réunion de lancement .
spar	spsar	spsar
spar / pm / apm	spsar / sgp / sglpsi	spsar / gp / glp

A challenge set approach to evaluating machine translation

Isabelle, Pierre; Cherry, Colin; Foster, George

Publisher's version / Version de l'éditeur:

Computer Science, 2017-08-27

A Challenge Set Approach to Evaluating Machine Translation

Pierre Isabelle and **Colin Cherry** and **George Foster**
National Research Council Canada
first.last@nrc.gc.ca

Abstract

Neural machine translation represents an exciting leap forward in translation quality. But what longstanding weaknesses does it resolve, and which remain? We address these questions with a challenge set approach to translation evaluation and error analysis. A challenge set consists of a small set of sentences, each hand-designed to probe a system's capacity to bridge a particular structural divergence between languages. To exemplify this approach, we present an English-French challenge set, and use it to analyze phrase-based and neural systems. The resulting analysis provides not only a more fine-grained picture of the strengths of neural systems, but also insight into which linguistic phenomena remain out of reach.

1 Introduction

The advent of neural techniques in machine translation (MT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014) has led to profound improvements in MT quality. For "easy" language pairs such as English/French or English/Spanish in particular, neural (NMT) systems are much closer to human performance than previous statistical techniques (Wu et al., 2016). This puts pressure on automatic evaluation metrics such as BLEU (Papineni et al., 2002), which exploit surface-matching heuristics that are relatively insensitive to subtle differences. As NMT continues to improve, these metrics will inevitably lose their effectiveness. Another challenge posed by NMT systems is their opacity: while it was usually clear which phenomena were ill-handled by previous statistical systems—and why—these questions are more difficult to answer for NMT.

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

Figure 1: Example challenge set question.

We propose a new evaluation methodology centered around a *challenge set* of difficult examples that are designed using expert linguistic knowledge to probe an MT system's capabilities. This methodology is complementary to the standard practice of randomly selecting a test set from "real text," which remains necessary in order to predict performance on new text. By concentrating on difficult examples, a challenge set is intended to provide a stronger signal to developers. Although we believe that the general approach is compatible with automatic metrics, we used manual evaluation for the work presented here. Our challenge set consists of short sentences that each focus on one particular phenomenon, which makes it easy to collect reliable manual assessments of MT output by asking direct yes-no questions. An example is shown in Figure 1.

We generated a challenge set for English to French translation by canvassing areas of linguistic divergence between the two language pairs, especially those where errors would be made visible by French morphology. Example choice was also partly motivated by extensive knowledge of the weaknesses of phrase-based MT (PBMT). Neither of these characteristics is essential to our method, however, which we envisage evolving as NMT progresses. We used our challenge set to evaluate in-house PBMT and NMT systems as well as Google's GNMT system.

In addition to proposing the novel idea of a challenge set evaluation, our contribution includes our annotated English-French challenge set, which we provide in an appendix and will make available in a separate machine-readable file. We also supply further evidence that NMT is systematically better than PBMT, even when BLEU score differences are small. Finally, we give an analysis of the challenges that remain to be solved in NMT, an area that has received little attention thus far.

2 Related Work

A number of recent papers have evaluated NMT using broad performance metrics. The WMT 2016 News Translation Task (Bojar et al., 2016) evaluated submitted systems according to both BLEU and human judgments. NMT systems were submitted to 9 of the 12 translation directions, winning 4 of these and tying for first or second in the other 5, according to the official human ranking. Since then, controlled comparisons have used BLEU to show that NMT outperforms strong PBMT systems on 30 translation directions from the United Nations Parallel Corpus (Junczys-Dowmunt et al., 2016a), and on the IWSLT English-Arabic tasks (Durrani et al., 2016). These evaluations indicate that NMT performs better on average than previous technologies, but they do not help us understand what aspects of the translation have improved.

Some groups have conducted more detailed error analyses. Bentivogli et al. (2016) carried out a number of experiments on IWSLT 2015 English-German evaluation data, where they compare machine outputs to professional post-edits in order to automatically detect a number of error categories. Compared to PBMT, NMT required less post-editing effort over-all, with substantial improvements in lexical, morphological and word order errors. NMT consistently out-performed PBMT, but its performance degraded faster as sentence length increased. Later, Toral and Sánchez-Cartagena (2017) conducted a similar study, examining the outputs of competition-grade systems for the 9 WMT 2016 directions that included NMT competitors. They reached similar conclusions regarding morphological inflection and word order, but found an even greater degradation in NMT performance as sentence length increased, perhaps due to these systems' use of subword units.

Most recently, Sennrich (2016) proposed an approach to perform targeted evaluations of NMT through the use of contrastive translation pairs. This method introduces a particular type of error automatically in reference sentences, and then checks whether the NMT system's conditional probability model prefers the original reference or the corrupted version. Using this technique, they are able to determine that a recently-proposed character-based model improves generalization on unseen words, but at the cost of introducing new grammatical errors.

Our approach differs from these studies in a number of ways. First, whereas others have analyzed sentences drawn from an existing bitext, we conduct our study on sentences that are manually constructed to exhibit canonical examples of specific linguistic phenomena. This challenge set methodology allows us to emphasize the difficult cases in an otherwise "easy" language pair. These sentences are designed to allow us to dive deep into phenomena of interest, and do a much finer-grained analysis of the strengths of NMT than has come before. However, this strategy also necessitates that we work on many fewer sentences. We leverage the small size of our challenge set to manually evaluate whether the system's actual output correctly handles our phenomena of interest. Manual evaluation side-steps some of the pitfalls that can come with Sennrich (2016)'s contrastive pairs, as a ranking of two contrastive sentences may not necessarily reflect whether the error in question will occur in the system's actual output.

3 Challenge Set Evaluation

Our challenge set is meant to measure the ability of MT systems to deal with some of the more difficult problems that arise in translating English into French. This particular language pair happened to be most convenient for us, but similar sets can be built for any language pair.

One aspect of MT performance that we aimed to exclude from our evaluation is robustness to sparse data. To control for this, when crafting source and reference sentences, we chose words that occurred at least 100 times in the training corpus described in section 4.1.[1]

[1] With three principled exceptions: *boeuf* (87 occurrences) and *spilt* (58 occurrences)—both part of idiomatic phrases—and *guitared* (0 occurrences).

3.1 Building the Challenge Set

The challenging aspect of the test set we are presenting stems from the fact that the source English sentences have been chosen so that their closest French equivalent will be *structurally divergent* from the source in some crucial way. Translational divergences have been extensively studied in the past – see for example (Vinay and Darbelnet, 1958; Dorr, 1994). We expect the level of difficulty of an MT test set to correlate well with its density in divergence phenomena. We classify divergence phenomena into three main types: morpho-syntactic, lexico-syntactic and purely syntactic divergences.

Morpho-syntactic divergences

In some languages, word morphology (e.g. inflections) carries more grammatical information than in others. When translating a word towards the richer language, there is a need to recover additional grammatically-relevant information from the context of the target language word. Note that we only include in our set cases where the relevant information is available in the *linguistic* context.[2]

We lack the space to describe all the subtypes of morpho-syntactic divergences that appear in our challenge set, but illustrate through representative examples. One particularly important case is that of *subject-verb agreement*. French verbs typically have more than 30 different inflected forms, while English verbs typically have 4 or 5. As a result, English verb forms strongly underspecify their French counterparts. Much of the missing information must be filled in through forced agreement in person, number and gender with the grammatical subject of the verb. But extracting these parameters can prove difficult. For example, the agreement features of a coordinated noun phrase are a complex function of the coordinated elements: a) the gender is feminine if all conjuncts are feminine, otherwise masculine wins; b) the conjunct with the smallest person (p1<p2<p3) wins; and c) the number is always plural when the coordination is "et" but the case is more complex with "ou".

[2]The so-called Winograd Schema Challenges (https://en.wikipedia.org/wiki/Winograd_Schema_Challenge) often involve cases where common-sense reasoning is required to correctly choose between two potential antecedent phrases for a pronoun. Such cases become En → Fr translation challenges when the pronoun in the source sentence is *they* and its alternative antecedents happen to have different grammatical genders in French: *they* → *ils/elles*.

A second example of morpho-syntactic divergence between English and French is the more explicit marking of the *subjunctive mood* in French subordinate clauses. In the following example, the verb "partiez", unlike its English counterpart, is marked as subjunctive:

> He demanded that you leave immediately. → Il a exigé que vous *partiez* immédiatement.

When translating an English verb within a subordinate clause, the context must be examined for possible subjunctive triggers. Typically these are specific lexical items found in a governing position with respect to the subordinate clause: verbs such as "exiger que", adjectives such as "regrettable que" or subordinate conjunctions such as "à condition que".

Lexico-syntactic divergences

Syntactically governing words such as verbs tend to impose specific requirements on their complements: they *subcategorize* for complements of a certain syntactic type. But a source language governor and its target language counterpart can diverge on their respective requirements. The translation of such words must then trigger adjustments in the target language complement pattern. We can only examine here a few of the subtypes instantiated in our challenge set.

A good example is *argument switching*. This refers to the situation where the translation of a source verb V_s as V_t is correct but only provided the arguments (usually the subject and the object) are flipped around. The translation of "to miss" as "manquer à" is such a case:

> John misses Mary → Mary *manque à* John.

Failing to perform the switch results in a severe case of mistranslation.

A second example of lexico-syntactic divergence is that of "crossing movement" verbs. Consider the following example:

> Terry swam across the river → Terry *a traversé* la rivière *à la nage*.

The French translation could be glossed as, "Terry crossed the river by swimming." A literal translation such as "Terry a nagé à travers la rivière," is ruled out.

Syntactic divergences

Some syntactic divergences are not relative to the presence of a particular lexical item but rather stem from differences in the basic set of available syntactic patterns. Source-language instances of structures that do not exist in the target language must be mapped onto equivalent structures. Here are some of the subtypes appearing in our challenge set.

The position of French pronouns is a major case of divergence from English. French is basically an SVO language just like English but it departs from that canonical order when post-verbal complements are pronominalized: the pronouns must then be *cliticized*, that is phonetically attached to the verb, in this case to the left side of the verb.

> He gave Mary a book. → Il a donné un livre à Marie.
>
> He gave$_i$ it$_j$ to her$_k$. → Il *le*$_j$ *lui*$_k$ a donné$_i$.

Another example of syntactic divergence between English and French is that of *stranded prepositions*. When forming a relative clause or a question on a prepositional phrase, English can leave the preposition stranded, fronting only the pronominalized object of that preposition. In French, the preposition needs to be fronted alongside its object:

> The girl whom$_i$ he was dancing with$_j$ is rich. → La fille *avec*$_j$ *qui*$_i$ il dansait est riche.

A final example of syntactic divergence is the use of the so-called *middle voice*. While English uses the passive voice in subjectless generic statements, French tends to prefer the use of a special pronominal construction where the pronoun "se" has no real referent:

> Caviar is eaten with bread. → Le caviar *se mange* avec du pain.

This completes our exemplification of morpho-syntactic, lexico-syntactic and purely syntactic divergences. Our actual test set includes several more subcategories of each type. The ability of MT systems to deal with each such subcategory is then tested using at least three different test sentences. We use short test sentences so as to keep the targeted divergence in focus. The 108 sentences that constitute our current challenge set can be found in Appendix B.

3.2 Evaluation Methodology

Given the very small size of our challenge set, it is easy to perform a human evaluation of the respective outputs of a handful of different systems. The obvious advantage is that the assessment is then absolute instead of relative to one or a few reference translations.

The intent of each challenge sentence is to test one and only one system capability, namely that of coping correctly with the particular associated divergence subtype. As illustrated in Figure 1, we provide annotators with a question that specifies the divergence phenomenon currently being tested, along with a reference translation with the areas of divergence highlighted. As a result, judgments become straightforward: was the targeted divergence correctly bridged, yes or no?[3] There is no need to mentally average over a number of different aspects of the test sentence as one does when rating the global translation quality of a sentence, e.g. on a 5-point scale. However, we acknowledge that measuring translation performance on complex sentences exhibiting many different phenomena remains crucial. We see our approach as being complementary to evaluations of overall translation quality.

One consequence of our divergence-focused approach is that faulty translations will be judged as successes when the faults lie outside of the targeted divergence zone. However, this problem is mitigated by our use of short test sentences.

4 Machine Translation Systems

We trained state-of-the-art neural and phrase-based systems for English-French translation on data from the WMT 2014 evaluation.

4.1 Data

We used the LIUM shared-task subset of the WMT 2014 corpora,[4] retaining the provided tokenization and corpus organization, but mapping characters to lowercase. Table 1 gives corpus statistics.

[3] Sometimes the system produces a translation that circumvents the divergence issue. For example, it may dodge a divergence involving adverbs by reformulating the translation to use an adjective instead. In these rare cases, we instruct our annotators to abstain from making a judgment, regardless of whether the translation is correct or not.

[4] http://www.statmt.org/wmt14/translation-task.html http://www-lium.univ-lemans.fr/~schwenk/nnmt-shared-task

corpus	lines	en words	fr words
train	12.1M	304M	348M
mono	15.9M	—	406M
dev	6003	138K	155K
test	3003	71K	81K

Table 1: Corpus statistics. The WMT12/13 eval sets are used for dev, and the WMT14 eval set is used for test.

4.2 Phrase-based systems

To ensure a competitive PBMT baseline, we performed phrase extraction using both IBM4 and HMM alignments with a phrase-length limit of 7; after frequency pruning, the resulting phrase table contained 516M entries. For each extracted phrase pair, we collected statistics for the hierarchical reordering model of Galley and Manning (2008).

We trained an NNJM model (Devlin et al., 2014) on the HMM-aligned training corpus, with input and output vocabulary sizes of 64K and 32K. Words not in the vocabulary were mapped to one of 100 mkcls classes. We trained for 60 epochs of 20K x 128 minibatches, yielding a final dev-set perplexity of 6.88.

Our set of log-linear features consisted of forward and backward Kneser-Ney smoothed phrase probabilities and HMM lexical probabilities (4 features); hierarchical reordering probabilities (6); the NNJM probability (1); a set of sparse features as described by Cherry (2013) (10,386); word-count and distortion penalties (2); and 5-gram language models trained on the French half of the training corpus and the French monolingual corpus (2). Tuning was carried out using batch lattice MIRA (Cherry and Foster, 2012). Decoding used the cube-pruning algorithm of Huang and Chiang (2007), with a distortion limit of 7.

We include two phrase-based systems in our comparison: PBMT-1 has data conditions that exactly match those of the NMT system, in that it does not use the language model trained on the French monolingual corpus, while PBMT-2 uses both language models.

4.3 Neural systems

To build our NMT system, we used the Nematus toolkit,[5] which implements a single-layer neural sequence-to-sequence architecture with attention (Bahdanau et al., 2015) and gated recurrent units (Cho et al., 2014). We used 512-dimensional word embeddings with source and target vocabulary sizes of 90K, and 1024-dimensional state vectors. The model contains 172M parameters.

We preprocessed the data using a BPE model learned from source and target corpora (Sennrich et al., 2016). Sentences longer than 50 words were discarded. Training used the Adadelta algorithm (Zeiler, 2012), with a minibatch size of 100 and gradients clipped to 1.0. It ran for 5 epochs, writing a checkpoint model every 30K minibatches. Following Junczys-Dowmunt et al. (2016b), we averaged the parameters from the last 8 checkpoints. To decode, we used the AmuNMT decoder (Junczys-Dowmunt et al., 2016a) with a beam size of 4.

While our primary results will focus on the above PBMT and NMT systems, where we can describe replicable configurations, we have also evaluated Google's production system,[6] which has recently moved to NMT (Wu et al., 2016). Notably, the "GNMT" system uses (at least) 8 encoder and 8 decoder layers, compared to our 1 layer for each, and it is trained on corpora that are "two to three decimal orders of magnitudes bigger than the WMT." The evaluated outputs were downloaded in December 2016.

5 Experiments

The 108-sentence English-French challenge set presented in Appendix B was submitted to the four MT systems described in section 4: PBMT-1, PBMT-2, NMT, and GNMT. We employed three bilingual native speakers of French who had no prior knowledge of the challenge set. They rated each translated sentence as either a success or a failure according to the protocol described in section 3.2. For example, the 26 sentences of the subcategories S1-S5 of Appendix B are all about different cases of subject-verb agreement. The corresponding translations were judged successful if and only if the translated verb correctly agrees with the translated subject.

The different system outputs for each source sentence were grouped together to reduce the burden on the annotators. That is, in figure 1, annotators were asked to answer the question for each of four outputs, rather than just one as shown. The outputs were listed in random order, without identification. Questions were also presented in ran-

[5] https://github.com/rsennrich/nematus

[6] https://translate.google.com

dom order to each annotator. Appendix A contains the instructions shown to the annotators.

5.1 Quantitative comparison

Table 2 summarizes our results in terms of percentage of successful translations, globally and over each main type of divergence. For comparison with traditional metrics, we also include BLEU scores measured on the WMT 2014 test set.

As we can see, the two PBMT systems fare very poorly on our challenge set, especially in the morpho-syntactic and purely syntactic types. Their relatively better handling of lexico-syntactic cases probably reflects the fact that PBMT systems are naturally more attuned to lexical cues than to morphology or syntax. The two NMT systems are clear winners in all three categories. The GNMT system is best overall with a success rate of 68%, likely due to the data and architectural factors mentioned in section 4.3.[7]

WMT BLEU scores correlate poorly with challenge-set performance. The large gap of 2.3 BLEU points between PBMT-1 and PBMT-2 corresponds to only a 1% gain on the challenge set, while the small gap of 0.4 BLEU between PBMT-2 and NMT corresponds to a 21% gain.

Inter-annotator agreement (final column in table 2) is excellent overall, with all three annotators agreeing on almost 90% of system outputs. Syntactic divergences appear to be somewhat harder to judge than other categories.

5.2 Qualitative assessment of NMT

We now turn to an analysis of the strengths and weaknesses of neural MT through the microscope of our divergence categorization system, hoping that this may help focus future research on key issues. In this discussion we ignore the results obtained by PBMT-2 and compare: a) the results obtained by PBMT-1 to those of NMT, both systems having been trained on the same dataset; and b) the results of these two systems with those of Google NMT which was trained on a much larger dataset.

In the remainder of the present section we will reference the sentences of our challenge set using the subcategory-based numbering scheme S1-S26 as assigned in Appendix B.

[7] We cannot offer a full comparison with the pre-NMT Google system. However, in October 2016 we ran a smaller 35-sentence version of our challenge set on both the Google system and our PBMT-1 system. The Google system only got 4 of those examples right (11.4%) while our PBMT-1 got 6 (17.1%).

Strengths of neural MT

Overall, both neural MT systems do much better than PBMT-1 at bridging divergences. Their dramatic advantage on morpho-syntactic divergences (a jump from 16% to 72% in the case of our two local systems) results from achievements such as the following:

- The subject's head noun agreement features get correctly passed to the verb phrase across intervening noun phrase complements (sentences S1a-c).
- Subject agreement marks appear to be correctly distributed to each element of a coordinated verb phrase (S3a-c).
- Much of the calculus that is at stake in determining the agreement features of a subject noun phrase (cf. our relevant description in section 3.1) appears to be correctly captured in the 12 translations of S4.
- Most instances of the difficult case of past participle agreement after the "avoir" auxiliary are correctly handled (S5b-e).

The NMT systems are also better at handling lexico-syntactic divergences. For example:

- They can perform the required restructuring of English double object constructions (sentences S8a-S8c).
- They can discriminate between an NP complement and a sentential complement starting with an NP: cf. *to know NP* versus *to know NP is VP* (S11b-e)
- They often correctly restructure English NP-to-VP complements (S12a-c).

Finally, NMT systems also turn out to better handle purely syntactic divergences. For example:

- The differences in yes-no question syntax is correctly bridged (S17a-c).
- English pronouns in verb complement position are often correctly cliticized, that is, moved before the main verb and case-inflected correctly (S23a-e).
- The Google NMT system manages to correctly translate tag questions (S18a-c), most cases of the "inalienable possession" construction (S25a-e), zero relative pronouns (S26a-c) and constructions with stranded prepositions (S19a-f).

Divergence type	PBMT-1	PBMT-2	NMT	Google NMT	Agreement
Morpho-syntactic	16%	16%	72%	65%	94%
Lexico-syntactic	42%	46%	52%	62%	94%
Syntactic	33%	33%	40%	75%	81%
Overall	31%	32%	53%	68%	89%
WMT BLEU	34.2	36.5	36.9	—	—

Table 2: Summary performance statistics for each system under study, including challenge set success rate grouped by linguistic category, as well as BLEU scores on the WMT 2014 test set. The final column gives the proportion of system outputs on which all three annotators agreed.

The large gap observed between the results of the in-house and Google NMT systems indicates that current neural MT systems are extremely data hungry. But given enough data, they can successfully tackle some challenges that are often thought of as extremely difficult. A case in point is that of stranded prepositions, in which English and French happen to diverge in their handling of the celebrated "WH-movement" long-distance dependencies. Specifically, in the French translation, the preposition must be repatriated with its fronted WH object no matter how far on the left it happens to be.

Weaknesses of neural MT

In spite of its clear edge over PBMT, NMT is not without some serious shortcomings. Some of them have been mentioned already, such as the tendency of system output to degrade with sentence length. By design this particular problem could not be observed with our challenge set. But many others get highlighted by an analysis of our results. Globally, we note that even using a staggering quantity of data and a highly sophisticated NMT model, the Google system fails to reach the 70% mark on our challenge set. Thus, there is ample room for improvement. The fine-grained error categorization associated with the challenge set makes it possible to single out precise areas where more research is needed. A first analysis of our results yields the following observations.

Incomplete generalizations. In several cases, while partial results might suggest that NMT has correctly captured a basic generalization about linguistic data, further instances reveal that this is not fully the case. Here are some examples:

- The calculus governing the agreement features of coordinated noun phrases (see section 3.1) appears to be handled correctly most of the time. However unlike our NMT system, the Google NMT system gets into difficulty with mixed-person subjects (sentences S4d1-3).
- While some subjunctive mood triggers are correctly captured (e.g. "demander que" and "malheureux que"), others such as the very common subordinate conjunction *provided that* $\rightarrow$ *à condition que* are getting missed (sentence S6a).
- The NMT systems often appear to have successfully captured the semantic relation that ties together the two nouns of an English noun compound, thereby giving rise to the correct preposition in the French translation $N_1\ N_2 \rightarrow N_2\ Prep\ N_1$. However, some cases that one might think of as easy are being missed. For example, the Google translation of "steak knife" (sentence S14c) fails to convey that this is a knife intended to cut steak; similarly, the Google translation of "paper filter" (sentence S14i) suggests the filter is intended to filter paper rather being made of it.
- The so-called French "inalienable possession" construction arises when an agent performs an action on one of her body parts, e.g. *I brushed my teeth*. In such cases the French translation normally follows a pattern that can be glossed as *He brushed the teeth to himself*. In our dataset, the Google system gets this right for examples in the first and third persons (sentences S25a,b) but fails to do the same with the example in the second person (sentence S25c).

Then there are also phenomena that current NMT systems, even with massive amounts of data, appear to be completely missing:

- *Idioms*. While PBMT-1 produces an acceptable translation for half of the idiomatic ex-

pressions of S15 and S16, the local NMT system misses them all and the Google system does just slightly better. It looks as if NMT systems lack sufficient capacity for raw memorization.

- *Control verbs*. Two different classes of verbs can govern a subject NP, an object NP plus an infinitival complement. With verbs of the "object-control" class (e.g. "persuade"), the object of the verb is understood as the semantic subject of the infinitive. But with those of the subject class (e.g. "promise"), it is rather the subject of the verb which plays that semantic role. None of the systems tested here appear to get a grip on subject control cases, as evidenced by the lack of correct feminine agreement on the French adjectives in sentences S2b-d.
- *Argument switching verbs*. All systems tested here mistranslate sentences S7a-c by failing to perform the required argument switch: *NP_1 misses NP_2 → NP_2 manque à NP_1*.
- *Crossing movement verbs*. None of the systems managed to correctly restructure the regular manner-of-movement verbs e.g. *swim across X → traverser X à la nage* in sentences S10a-c, let alone the even harder example S10d, in which the word "guitar" is spontaneously recast as a manner-of-movement verb.
- Middle voice. None of the systems tested here were able to recast the English "generic passive" of S21a-c into the expected French "middle voice" pronominal construction.

6 Conclusions

We have presented a radically different kind of evaluation for machine translation systems: the use of challenge sets designed to stress-test MT systems on "hard" linguistic material, while providing a fine-grained linguistic classification of their successes and failures. This approach is not meant to replace our community's traditional evaluation tools but to supplement them.

Our proposed error categorization scheme makes it possible to bring to light different strengths and weaknesses of PBMT and neural MT. With the exception of idiom processing, in all cases where a clear difference was observed it turned out to be in favor of neural MT. A key factor in NMT's superiority appears to be its ability to overcome many limitations of n-gram language modeling. This is clearly at play in dealing with subject-verb agreement, double-object verbs, overlapping subcategorization frames and last but not least, the pinnacle of Chomskyan linguistics, WH-movement (in this case, stranded prepositions).

But our challenge set also brings to light some important shortcomings of current neural MT, regardless of the massive amounts of training data it may have been fed. As may have been already known or suspected, NMT systems struggle with the translation of idiomatic phrases. Perhaps more interestingly, we notice that neural MT's impressive generalizations still seem somewhat brittle. For example, the NMT system can appear to have mastered the rules governing subject-verb agreement or inalienable possession in French, only to trip over a rather obvious instantiation of those rules. Probing where these boundaries are, and how they relate to the neural system's training data and architecture is an obvious next step.

7 Future Work

It is our hope that the insights derived from our challenge set evaluation will help inspire future MT research, and call attention to the fact that even "easy" language pairs like English-French still have many linguistic issues left to be resolved. But there are also several ways to improve and expand upon our challenge set approach itself.

First, though our human judgments of output sentences allowed us to precisely assess the phenomena of interest, this approach is not scalable to large sets, and requires access to native speakers in order to replicate the evaluation. It would be interesting to see whether similar scores could be achieved through automatic means. The existence of human judgments for this set provides a gold-standard by which proposed automatic judgments may be meta-evaluated.

Second, the construction of such a challenge set is as much an art as a science, and requires in-depth knowledge of the structural divergences between the two languages of interest. A method to automatically create such a challenge set for a new language pair would be extremely useful. One could imagine approaches that search for divergences, indicated by atypical output configurations, or perhaps by a system's inability to repro-

duce a reference from its own training data. Localizing a divergence within a difficult sentence pair would be another useful subtask.

Finally, and perhaps most interestingly, we would like to explore how to train an MT system to improve its performance on these divergence phenomena. This could take the form of designing a curriculum to demonstrate a particular divergence to the machine, or altering the network structure to more easily capture such generalizations.

Acknowledgments

We would like to thank Cyril Goutte, Eric Joanis and Michel Simard, who graciously spent the time required to rate the output of four different MT systems on our challenge sentences. We also thank Roland Kuhn for valuable discussions, and comments on an earlier version of the paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*. San Diego, USA. http://arxiv.org/abs/1409.0473.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 257–267. https://aclweb.org/anthology/D16-1025.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198. http://www.aclweb.org/anthology/W16-2301.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 22–31. http://www.aclweb.org/anthology/N13-1003.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, pages 427–436. http://www.aclweb.org/anthology/N12-1047.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734. http://www.aclweb.org/anthology/D14-1179.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 1370–1380. http://www.aclweb.org/anthology/P14-1129.

Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics* 20:4.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2016. QCRI machine translation systems for IWSLT 16. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, pages 848–856. http://www.aclweb.org/anthology/D08-1089.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 144–151. http://www.aclweb.org/anthology/P07-1019.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016a. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, Washington.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. 2016b. The amu-uedin submission to

the wmt16 news translation task: Attention-based nmt models as feature functions in phrase-based smt. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 319–325. http://www.aclweb.org/anthology/W16-2316.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 1700–1709. http://www.aclweb.org/anthology/D13-1176.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rico Sennrich. 2016. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. *CoRR* abs/1612.04629. http://arxiv.org/abs/1612.04629.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pages 3104–3112.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus statistical machine translation for 9 language directions. In *Proceedings of the The 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais*, volume 1. Didier, Paris.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144. http://arxiv.org/abs/1609.08144.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR* abs/1212.5701. http://arxiv.org/abs/1212.5701.

A Instructions to Annotators

The following instructions were provided to annotators:

You will be presented with 108 short English sentences and the French translations produced for them by each of four different machine translation systems. You will not be asked to provide an overall rating for the machine-translated sentences. Rather, you will be asked to determine whether or not a highly specific aspect of the English sentence is correctly rendered in each of the different translations. Each English sentence will be accompanied with a yes-no question which precisely specifies the targeted element for the associated translations. For example, you may be asked to determine whether or not the main verb phrase of the translation is in correct grammatical agreement with its subject.

In order to facilitate this process, each English sentence will also be provided with a French reference (human) translation in which the particular elements that support a yes answer (in our example, the correctly agreeing verb phrase) will be highlighted. Your answer should be "yes" if the question can be answered positively and "no" otherwise. Note that this means that any translation error which is unrelated to the question at hand should be disregarded. Using the same example: as long as the verb phrase agrees correctly with its subject, it does not matter whether or not the verb is correctly chosen, is in the right tense, etc. And of course, it does not matter if unrelated parts of the translation are wrong.

In most cases you should be able to quickly determine a positive or negative answer. However, there may be cases in which the system has come up with a translation that just does not contain the phenomenon targeted by the associated question. In such cases, and only in such cases, you should choose "not applicable" regardless of whether or not the translation is correct.

B Challenge Set

We include a rendering of our challenge set in the pages that follow, along with system output for the PBMT-1, NMT and Google systems. Sentences are grouped by linguistic category and subcategory. For convenience, we also include a reference translation, which is a manually-crafted translation that is designed to be the most straightforward solution to the divergence problem at hand. Needless to say, this reference translation is seldom the only acceptable solution to the targeted divergence problem. Our judges were provided these references, but were instructed to use their knowledge of French to judge whether the divergence was correctly bridged, regardless of the translation's similarity to the reference.

In all translations, the locus of the targeted divergence is highlighted in boldface and it is specifically on that portion that our annotators were asked to provide a judgment. For each system output, we provide a summary of our annotator's judgments on its handling of the phenomenon of interest. We label the translation with a ✓ if two or more annotators judged the divergence to be correctly bridged, and with an ✗ otherwise.

We also release a machine-readable version of this same data, including all of the individual judgments, in the hope that others will find interesting new uses for it.

Morpho-Syntactic

S-V agreement, across distractors

Is subject-verb agrement correct? (Possible interference from distractors between the subject's head and the verb).

S1a	Source	The repeated calls from his mother **should** have alerted us.
	Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
	PBMT-1	Les appels répétés de sa mère aurait dû nous a alertés. ✗
	NMT	Les appels répétés de sa mère devraient nous avoir alertés. ✓
	Google	Les appels répétés de sa mère auraient dû nous alerter. ✓
S1b	Source	The sudden noise in the upper rooms **should** have alerted us.
	Ref	Le bruit soudain dans les chambres supérieures **aurait** dû nous alerter.
	PBMT-1	Le bruit soudain dans les chambres supérieures auraient dû nous a alertés. ✗
	NMT	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
	Google	Le bruit soudain dans les chambres supérieures devrait nous avoir alerté. ✓
S1c	Source	Their repeated failures to report the problem **should** have alerted us.
	Ref	Leurs échecs répétés à signaler le problème **auraient** dû nous alerter.
	PBMT-1	Leurs échecs répétés de signaler le problème aurait dû nous a alertés. ✗
	NMT	Leurs échecs répétés pour signaler le problème devraient nous avoir alertés. ✓
	Google	Leur échec répété à signaler le problème aurait dû nous alerter. ✓

S-V agreement, through control verbs

Does the flagged adjective agree correctly with its subject? (Subject-control versus object-control verbs).

S2a	Source	She asked her brother not to be **arrogant**.
	Ref	Elle a demandé à son frère de ne pas se montrer **arrogant**.
	PBMT-1	Elle a demandé à son frère de ne pas être arrogant. ✓
	NMT	Elle a demandé à son frère de ne pas être arrogant. ✓
	Google	Elle a demandé à son frère de ne pas être arrogant. ✓
S2b	Source	She promised her brother not to be **arrogant**.
	Ref	Elle a promis à son frère de ne pas être **arrogante**.
	PBMT-1	Elle a promis son frère à ne pas être arrogant. ✗
	NMT	Elle a promis à son frère de ne pas être arrogant. ✗
	Google	Elle a promis à son frère de ne pas être arrogant. ✗
S2c	Source	She promised her doctor to remain **active** after retiring.
	Ref	Elle a promis à son médecin de demeurer **active** après s'être retirée.
	PBMT-1	Elle a promis son médecin pour demeurer actif après sa retraite. ✗
	NMT	Elle a promis à son médecin de rester actif après sa retraite. ✗
	Google	Elle a promis à son médecin de rester actif après sa retraite. ✗
S2d	Source	My mother promised my father to be more **prudent** on the road.
	Ref	Ma mère a promis à mon père d'être plus **prudente** sur la route.
	PBMT-1	Ma mère, mon père a promis d'être plus prudent sur la route. ✗
	NMT	Ma mère a promis à mon père d'être plus prudent sur la route. ✗
	Google	Ma mère a promis à mon père d'être plus prudent sur la route. ✗

S-V agreement, coordinated targets

Do the marked verbs/adjective agree correctly with their subject? (Agreement distribution over coordinated predicates)

S3a	Source	The woman was very **tall** and extremely **strong**.
	Ref	La femme était très **grande** et extrêmement **forte**.
	PBMT-1	La femme était très gentil et extrêmement forte. ✗
	NMT	La femme était très haute et extrêmement forte. ✓
	Google	La femme était très grande et extrêmement forte. ✓
S3b	Source	Their politicians were more **ignorant** than **stupid**.
	Ref	Leurs politiciens étaient plus **ignorants** que **stupides**.
	PBMT-1	Les politiciens étaient plus ignorants que stupide. ✗
	NMT	Leurs politiciens étaient plus ignorants que stupides. ✓
	Google	Leurs politiciens étaient plus ignorants que stupides. ✓
S3c	Source	We **shouted** an insult and **left** abruptly.
	Ref	Nous **avons** lancé une insulte et nous **sommes** partis brusquement.
	PBMT-1	Nous avons crié une insulte et a quitté abruptement. ✗
	NMT	Nous avons crié une insulte et nous avons laissé brusquement. ✓
	Google	Nous avons crié une insulte et nous sommes partis brusquement. ✓

S-V agreement, feature calculus on coordinated source

Do the marked verbs/adjective agree correctly with their subject? (Masculine singular ET masculine singular yields masculine plural).

S4a1	Source	The cat and the dog **should** be **watched**.
	Ref	Le chat et le chien **devraient** être **surveillés**.
	PBMT-1	Le chat et le chien doit être regardée. ✗
	NMT	Le chat et le chien doivent être regardés. ✓
	Google	Le chat et le chien doivent être surveillés. ✓
S4a2	Source	My father and my brother **will** be **happy** tomorrow.
	Ref	Mon père et mon frère **seront heureux** demain.
	PBMT-1	Mon père et mon frère sera heureux de demain. ✗
	NMT	Mon père et mon frère seront heureux demain. ✓
	Google	Mon père et mon frère seront heureux demain. ✓
S4a3	Source	My book and my pencil **could** be **stolen**.
	Ref	Mon livre et mon crayon **pourraient** être **volés**.
	PBMT-1	Mon livre et mon crayon pourrait être volé. ✗
	NMT	Mon livre et mon crayon pourraient être volés. ✓
	Google	Mon livre et mon crayon pourraient être volés. ✓

Do the marked verbs/adjectives agree correctly with their subject? (Feminine singular ET feminine singular yields feminine plural).

S4b1	Source	The cow and the hen **must** be **fed**.
	Ref	La vache et la poule **doivent** être **nourries**.
	PBMT-1	La vache et de la poule doivent être nourris. ✗
	NMT	La vache et la poule doivent être alimentées. ✓
	Google	La vache et la poule doivent être nourries. ✓

S4b2	Source	My mother and my sister **will be happy** tomorrow.
	Ref	Ma mère et ma sœur **seront heureuses** demain.
	PBMT-1	Ma mère et ma sœur sera heureux de demain. ✗
	NMT	Ma mère et ma sœur seront heureuses demain. ✓
	Google	Ma mère et ma sœur seront heureuses demain. ✓
S4b3	Source	My shoes and my socks **will** be **found**.
	Ref	Mes chaussures et mes chaussettes **seront retrouvées**.
	PBMT-1	Mes chaussures et mes chaussettes sera trouvé. ✗
	NMT	Mes chaussures et mes chaussettes seront trouvées. ✓
	Google	Mes chaussures et mes chaussettes seront trouvées. ✓
Do the marked verbs/adjectives agree correctly with their subject? (Masculine singular ET feminine singular yields masculine plural.)		
S4c1	Source	The dog and the cow **are nervous**.
	Ref	Le chien et la vache **sont nerveux**.
	PBMT-1	Le chien et la vache sont nerveux. ✓
	NMT	Le chien et la vache sont nerveux. ✓
	Google	Le chien et la vache sont nerveux. ✓
S4c2	Source	My father and my mother will be happy tomorrow.
	Ref	Mon père et ma mère **seront heureux** demain.
	PBMT-1	Mon père et ma mère se fera un plaisir de demain. ✗
	NMT	Mon père et ma mère seront heureux demain. ✓
	Google	Mon père et ma mère seront heureux demain. ✓
S4c3	Source	My refrigerator and my kitchen table **were stolen**.
	Ref	Mon réfrigérateur et ma table de cuisine **ont** été **volés**.
	PBMT-1	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
	NMT	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
	Google	Mon réfrigérateur et ma table de cuisine ont été volés. ✓
Do the marked verbs/adjectives agree correctly with their subject? (Smallest coordinated grammatical person wins.)		
S4d1	Source	Paul and I **could** easily be **convinced** to join you.
	Ref	Paul et moi **pourrions** facilement être **convaincus** de se joindre à vous.
	PBMT-1	Paul et je pourrais facilement être persuadée de se joindre à vous. ✗
	NMT	Paul et moi avons facilement pu être convaincus de vous rejoindre. ✓
	Google	Paul et moi pourrait facilement être convaincu de vous rejoindre. ✗
S4d2	Source	You and he **could** be **surprised** by her findings.
	Ref	Vous et lui **pourriez** être **surpris** par ses découvertes.
	PBMT-1	Vous et qu'il pouvait être surpris par ses conclusions. ✗
	NMT	Vous et lui pourriez être surpris par ses conclusions. ✓
	Google	Vous et lui pourrait être surpris par ses découvertes. ✗

S4d3	Source	We and they **are** on different courses.
	Ref	Nous et eux **sommes** sur des trajectoires différentes.
	PBMT-1	Nous et ils sont en cours de différents. ✗
	NMT	Nous et nous sommes sur des parcours différents. ✗
	Google	Nous et ils sont sur des parcours différents. ✗

S-V agreement, past participles

Are the agreement marks of the flagged participles the correct ones? (Past participle placed after auxiliary AVOIR agrees with verb object iff object precedes auxiliary. Otherwise participle is in masculine singular form).

S5a	Source	The woman who **saw** a mouse in the corridor is charming.
	Ref	La femme qui a **vu** une souris dans le couloir est charmante.
	PBMT-1	La femme qui a vu une souris dans le couloir est charmante. ✓
	NMT	La femme qui a vu une souris dans le couloir est charmante. ✓
	Google	La femme qui a vu une souris dans le couloir est charmante. ✓
S5b	Source	The woman that your brother **saw** in the corridor is charming.
	Ref	La femme que votre frère a **vue** dans le couloir est charmante.
	PBMT-1	La femme que ton frère a vu dans le couloir est charmante. ✗
	NMT	La femme que votre frère a vu dans le corridor est charmante. ✗
	Google	La femme que votre frère a vue dans le couloir est charmante. ✓
S5c	Source	The house that John has **visited** is crumbling.
	Ref	La maison que John a **visitée** tombe en ruines.
	PBMT-1	La maison que John a visité est en train de s'écrouler. ✗
	NMT	La maison que John a visitée est en train de s'effondrer. ✓
	Google	La maison que John a visité est en ruine. ✗
S5d	Source	John sold the car that he had **won** in a lottery.
	Ref	John a vendu la voiture qu'il avait **gagnée** dans une loterie.
	PBMT-1	John a vendu la voiture qu'il avait gagné à la loterie. ✗
	NMT	John a vendu la voiture qu'il avait gagnée dans une loterie. ✓
	Google	John a vendu la voiture qu'il avait gagnée dans une loterie. ✓

Subjunctive mood

Is the flagged verb in the correct mood? (Certain triggering verbs, adjectives or subordinate conjunctions, induce the subjunctive mood in the subordinate clause that they govern).

S6a	Source	He will come provided that you **come** too.
	Ref	Il viendra à condition que vous **veniez** aussi.
	PBMT-1	Il viendra à condition que vous venez aussi. ✗
	NMT	Il viendra lui aussi que vous le faites. ✗
	Google	Il viendra à condition que vous venez aussi. ✗
S6b	Source	It is unfortunate that he is not **coming** either.
	Ref	Il est malheureux qu'il ne **vienne** pas non plus.
	PBMT-1	Il est regrettable qu'il n'est pas non plus à venir. ✗
	NMT	Il est regrettable qu'il ne soit pas non plus. ✗
	Google	Il est malheureux qu'il ne vienne pas non plus. ✓

S6c	Source	I requested that families not **be** separated.
	Ref	J'ai demandé que les familles ne **soient** pas séparées.
	PBMT-1	J'ai demandé que les familles ne soient pas séparées. ✓
	NMT	J'ai demandé que les familles ne soient pas séparées. ✓
	Google	J'ai demandé que les familles ne soient pas séparées. ✓

Lexico-Syntactic

Argument switch

Are the experiencer and the object of the "missing" situation correctly preserved in the French translation? (Argument switch).

S7a	Source	**Mary** sorely misses **Jim**.
	Ref	**Jim** manque cruellement **à Mary**.
	PBMT-1	Marie manque cruellement de Jim. ✗
	NMT	Mary a lamentablement manqué de Jim. ✗
	Google	Mary manque cruellement à Jim. ✗
S7b	Source	**My sister** is really missing **New York.**
	Ref	**New York** manque beaucoup **à ma sœur**.
	PBMT-1	Ma sœur est vraiment absent de New York. ✗
	NMT	Ma sœur est vraiment manquante à New York. ✗
	Google	Ma sœur manque vraiment New York. ✗
S7c	Source	What **he** misses most is **his dog**.
	Ref	Ce qui **lui** manque le plus, c'est **son chien**.
	PBMT-1	Ce qu'il manque le plus, c'est son chien. ✗
	NMT	Ce qu'il manque le plus, c'est son chien. ✗
	Google	Ce qu'il manque le plus, c'est son chien. ✗

Double-object verbs

Are "gift" and "recipient" arguments correctly rendered in French? (English double-object constructions)

S8a	Source	John gave **his wonderful wife** a nice present.
	Ref	John a donné un beau présent **à sa merveilleuse épouse**.
	PBMT-1	John a donné sa merveilleuse femme un beau cadeau. ✗
	NMT	John a donné à sa merveilleuse femme un beau cadeau. ✓
	Google	John a donné à son épouse merveilleuse un présent gentil. ✓
S8b	Source	John told **the kids** a nice story.
	Ref	John a raconté une belle histoire **aux enfants**.
	PBMT-1	John a dit aux enfants une belle histoire. ✓
	NMT	John a dit aux enfants une belle histoire. ✓
	Google	John a raconté aux enfants une belle histoire. ✓
S8c	Source	John sent **his mother** a nice postcard.
	Ref	John a envoyé une belle carte postale **à sa mère**.
	PBMT-1	John a envoyé sa mère une carte postale de nice. ✗
	NMT	John a envoyé sa mère une carte postale de nice. ✗
	Google	John envoya à sa mère une belle carte postale. ✓

Fail to

Is the meaning of "fail to" correctly rendered in the French translation?		
S9a	Source	John **failed to** see the relevance of this point.
	Ref	John **n'a pas** vu la pertinence de ce point.
	PBMT-1	John a omis de voir la pertinence de ce point. ✗
	NMT	John n'a pas vu la pertinence de ce point. ✓
	Google	John a omis de voir la pertinence de ce point. ✗
S9b	Source	He failed to respond.
	Ref	Il **n'a pas répondu**.
	PBMT-1	Il n'a pas réussi à répondre. ✓
	NMT	Il n'a pas répondu. ✓
	Google	Il n'a pas répondu. ✓
S9c	Source	Those who fail to comply with this requirement will be penalized.
	Ref	Ceux qui **ne** se conforment **pas** à cette exigence seront pénalisés.
	PBMT-1	Ceux qui ne se conforment pas à cette obligation seront pénalisés. ✓
	NMT	Ceux qui ne se conforment pas à cette obligation seront pénalisés. ✓
	Google	Ceux qui ne respectent pas cette exigence seront pénalisés. ✓

Manner-of-movement verbs

Is the movement action expressed in the English source correctly rendered in French? (Manner-of-movement verbs with path argument may need to be rephrased in French).		
S10a	Source	John would like to **swim across** the river.
	Ref	John aimerait **traverser** la rivière **à la nage**.
	PBMT-1	John aimerait nager dans la rivière. ✗
	NMT	John aimerait nager à travers la rivière. ✗
	Google	John aimerait nager à travers la rivière. ✗
S10b	Source	They **ran into** the room.
	Ref	Ils **sont entrés** dans la chambre **à la course**.
	PBMT-1	Ils ont couru dans la chambre. ✗
	NMT	Ils ont couru dans la pièce. ✗
	Google	Ils coururent dans la pièce. ✗
S10c	Source	The man **ran out of** the park.
	Ref	L'homme **est sorti du** parc **en courant**.
	PBMT-1	L'homme a manqué du parc. ✗
	NMT	L'homme s'enfuit du parc. ✗
	Google	L'homme sortit du parc. ✗
Hard example featuring spontaneous noun-to-verb derivation ("nonce verb").		
S10d	Source	John **guitared his way** to San Francisco.
	Ref	John **s'est rendu** jusqu'à San Francisco **en jouant de la guitare**.
	PBMT-1	John guitared son chemin à San Francisco. ✗
	NMT	John guitared sa route à San Francisco. ✗
	Google	John a guité son chemin à San Francisco. ✗

Overlapping subcat frames

Is the French verb for "know" correctly chosen? (Choice between "savoir"/"connaître" depends on syntactic nature of its object)

S11a	Source	Paul **knows** that this is a fact.
	Ref	Paul **sait que** c'est un fait.
	PBMT-1	Paul sait que c'est un fait. ✓
	NMT	Paul sait que c'est un fait. ✓
	Google	Paul sait que c'est un fait. ✓
S11b	Source	Paul **knows** this story.
	Ref	Paul **connaît** cette histoire.
	PBMT-1	Paul connaît cette histoire. ✓
	NMT	Paul connaît cette histoire. ✓
	Google	Paul connaît cette histoire. ✓
S11c	Source	Paul **knows** this story is hard to believe.
	Ref	Paul **sait que** cette histoire est difficile à croire.
	PBMT-1	Paul connaît cette histoire est difficile à croire. ✗
	NMT	Paul sait que cette histoire est difficile à croire. ✓
	Google	Paul sait que cette histoire est difficile à croire. ✓
S11d	Source	He **knows** my sister will not take it.
	Ref	Il **sait que** ma soeur ne le prendra pas.
	PBMT-1	Il sait que ma soeur ne prendra pas. ✓
	NMT	Il sait que ma soeur ne le prendra pas. ✓
	Google	Il sait que ma soeur ne le prendra pas. ✓
S11e	Source	My sister **knows** your son is reliable.
	Ref	Ma sœur **sait que** votre fils est fiable.
	PBMT-1	Ma soeur connaît votre fils est fiable. ✗
	NMT	Ma sœur sait que votre fils est fiable. ✓
	Google	Ma sœur sait que votre fils est fiable. ✓

NP to VP

Is the English "NP to VP" complement correctly rendred in the French translation? (Sometimes one needs to translate this structure as a finite clause).

S12a	Source	John believes **Bill to be dishonest**.
	Ref	John croit **que Bill est malhonnête**.
	PBMT-1	John estime que le projet de loi soit malhonnête. ✓
	NMT	John croit que le projet de loi est malhonnête. ✓
	Google	John croit que Bill est malhonnête. ✓
S12b	Source	He liked **his father to tell him stories**.
	Ref	Il aimait **que son père lui raconte des histoires**.
	PBMT-1	Il aimait son père pour lui raconter des histoires. ✗
	NMT	Il aimait son père pour lui raconter des histoires. ✗
	Google	Il aimait son père à lui raconter des histoires. ✗

S12c	Source	She wanted **her mother to let her go**.
	Ref	Elle voulait **que sa mère la laisse partir**.
	PBMT-1	Elle voulait que sa mère de lui laisser aller. ✗
	NMT	Elle voulait que sa mère la laisse faire. ✓
	Google	Elle voulait que sa mère la laisse partir. ✓

Factitives

Is the English verb correctly rendered in the French translation? (Agentive use of some French verbs require embedding under "faire").

S13a	Source	John **cooked** a big chicken.
	Ref	John a **fait cuire** un gros poulet.
	PBMT-1	John cuit un gros poulet. ✗
	NMT	John cuit un gros poulet. ✗
	Google	John a fait cuire un gros poulet. ✓
S13b	Source	John **melted** a lot of ice.
	Ref	John a **fait fondre** beaucoup de glace.
	PBMT-1	John fondu a lot of ice. ✗
	NMT	John a fondu beaucoup de glace. ✗
	Google	John a fondu beaucoup de glace. ✗
S13c	Source	She likes to **grow** flowers.
	Ref	Elle aime **faire pousser** des fleurs.
	PBMT-1	Elle aime à se développer des fleurs. ✗
	NMT	Elle aime à cultiver des fleurs. ✓
	Google	Elle aime faire pousser des fleurs. ✓

Noun Compounds

Is the English nominal compound rendered with the right preposition in the French translation?

S14a	Source	Use the meat knife.
	Ref	Utilisez le couteau **à** viande.
	PBMT-1	Utilisez le couteau de viande. ✗
	NMT	Utilisez le couteau à viande. ✓
	Google	Utilisez le couteau à viande. ✓
S14b	Source	Use the butter knife.
	Ref	Utilisez le couteau **à** beurre.
	PBMT-1	Utilisez le couteau à beurre. ✓
	NMT	Utilisez le couteau au beurre. ✗
	Google	Utilisez le couteau à beurre. ✓
S14c	Source	Use the steak knife.
	Ref	Utilisez le couteau **à** steak.
	PBMT-1	Utilisez le steak couteau. ✗
	NMT	Utilisez le couteau à steak. ✓
	Google	Utilisez le couteau de steak. ✗

S14d	Source	Clean the water filter.
	Ref	Nettoyez le filtre **à** eau.
	PBMT-1	Nettoyez le filtre à eau. ✓
	NMT	Nettoyez le filtre à eau. ✓
	Google	Nettoyez le filtre à eau. ✓
S14e	Source	Clean the juice filter.
	Ref	Nettoyez le filtre **à** jus.
	PBMT-1	Nettoyez le filtre de jus. ✗
	NMT	Nettoyez le filtre de jus. ✗
	Google	Nettoyez le filtre à jus. ✓
S14f	Source	Clean the tea filter.
	Ref	Nettoyez le filtre **à** thé.
	PBMT-1	Nettoyez le filtre à thé. ✓
	NMT	Nettoyez le filtre de thé. ✗
	Google	Nettoyez le filtre à thé. ✓
S14g	Source	Clean the cloth filter.
	Ref	Nettoyez le filtre **en** tissu.
	PBMT-1	Nettoyez le filtre en tissu. ✓
	NMT	Nettoyez le filtre en tissu. ✓
	Google	Nettoyez le filtre en tissu. ✓
S14h	Source	Clean the metal filter.
	Ref	Nettoyez le filtre **en** métal.
	PBMT-1	Nettoyez le filtre en métal. ✓
	NMT	Nettoyez le filtre en métal. ✓
	Google	Nettoyez le filtre métallique. ✓
S14i	Source	Clean the paper filter.
	Ref	Nettoyez le filtre **en** papier.
	PBMT-1	Nettoyez le filtre en papier. ✓
	NMT	Nettoyez le filtre en papier. ✓
	Google	Nettoyez le filtre à papier. ✗

Common idioms

Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?		
S15a	Source	Stop **beating around the bush**.
	Ref	Cessez de **tourner autour du pot**.
	PBMT-1	Cesser de battre la campagne. ✗
	NMT	Arrêtez de battre autour de la brousse. ✗
	Google	Arrêter de tourner autour du pot. ✓

S15b	Source	You are **putting the cart before the horse**.
	Ref	Vous **mettez la charrue devant les bœufs**.
	PBMT-1	Vous pouvez mettre la charrue avant les bœufs. ✓
	NMT	Vous mettez la charrue avant le cheval. ✗
	Google	Vous mettez le chariot devant le cheval. ✗
S15c	Source	His comment proved to be **the straw that broke the camel's back**.
	Ref	Son commentaire s'est avéré être **la goutte d'eau qui a fait déborder le vase**.
	PBMT-1	Son commentaire s'est révélé être la goutte d'eau qui fait déborder le vase. ✓
	NMT	Son commentaire s'est avéré être la paille qui a brisé le dos du chameau. ✗
	Google	Son commentaire s'est avéré être la paille qui a cassé le dos du chameau. ✗
S15d	Source	His argument really **hit the nail on the head**.
	Ref	Son argument a vraiment **fait mouche**.
	PBMT-1	Son argument a vraiment mis le doigt dessus. ✓
	NMT	Son argument a vraiment frappé le clou sur la tête. ✗
	Google	Son argument a vraiment frappé le clou sur la tête. ✗
S15e	Source	It's **no use crying over spilt milk**.
	Ref	**Ce qui est fait est fait**.
	PBMT-1	Ce n'est pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait haché. ✗
	Google	Ce qui est fait est fait. ✓
S15f	Source	It is **no use crying over spilt milk**.
	Ref	**Ce qui est fait est fait**.
	PBMT-1	Il ne suffit pas de pleurer sur le lait répandu. ✗
	NMT	Il ne sert à rien de pleurer sur le lait écrémé. ✗
	Google	Il est inutile de pleurer sur le lait répandu. ✗

Syntactically flexible idioms

Is the English idiomatic expression correctly rendered with a suitable French idiomatic expression?		
S16a	Source	The cart has been put before the horse.
	Ref	La **charrue a été mise devant les bœufs**.
	PBMT-1	On met la charrue devant le cheval. ✗
	NMT	Le chariot a été mis avant le cheval. ✗
	Google	Le chariot a été mis devant le cheval. ✗
S16b	Source	With this argument, **the nail has been hit on the head**.
	Ref	Avec cet argument, **la cause est entendue**.
	PBMT-1	Avec cette argument, l'ongle a été frappée à la tête. ✗
	NMT	Avec cet argument, l'ongle a été touché à la tête. ✗
	Google	Avec cet argument, le clou a été frappé sur la tête. ✗

Syntactic

Yes-no question syntax

Is the English question correctly rendered as a French question?		
S17a	Source	**Have the kids** ever watched that movie?
	Ref	**Les enfants ont-ils** déjà vu ce film?
	PBMT-1	Les enfants jamais regardé ce film? ✗
	NMT	Les enfants ont-ils déjà regardé ce film? ✓
	Google	Les enfants ont-ils déjà regardé ce film? ✓
S17b	Source	**Hasn't your boss denied you** a promotion?
	Ref	**Votre patron ne vous a-t-il pas refusé** une promotion?
	PBMT-1	N'a pas nié votre patron vous un promotion? ✗
	NMT	Est-ce que votre patron vous a refusé une promotion? ✓
	Google	Votre patron ne vous a-t-il pas refusé une promotion? ✓
S17c	Source	**Shouldn't I attend** this meeting?
	Ref	**Ne devrais-je pas assister** à cette réunion?
	PBMT-1	Ne devrais-je pas assister à cette réunion? ✓
	NMT	Est-ce que je ne devrais pas assister à cette réunion? ✓
	Google	Ne devrais-je pas assister à cette réunion? ✓

Tag questions

Is the English "tag question" element correctly rendered in the translation?		
S18a	Source	Mary looked really happy tonight, **didn't she**?
	Ref	Mary avait l'air vraiment heureuse ce soir, **n'est-ce pas**?
	PBMT-1	Marie a regardé vraiment heureux de ce soir, n'est-ce pas elle? ✗
	NMT	Mary s'est montrée vraiment heureuse ce soir, ne l'a pas fait? ✗
	Google	Mary avait l'air vraiment heureuse ce soir, n'est-ce pas? ✓
S18b	Source	We should not do that again, **should we**?
	Ref	Nous ne devrions pas refaire cela, **n'est-ce pas**?
	PBMT-1	Nous ne devrions pas faire qu'une fois encore, faut-il? ✗
	NMT	Nous ne devrions pas le faire encore, si nous? ✗
	Google	Nous ne devrions pas recommencer, n'est-ce pas? ✓
S18c	Source	She was perfect tonight, **was she not**?
	Ref	Elle était parfaite ce soir, **n'est-ce pas**?
	PBMT-1	Elle était parfait ce soir, elle n'était pas? ✗
	NMT	Elle était parfaite ce soir, n'était-elle pas? ✗
	Google	Elle était parfaite ce soir, n'est-ce pas? ✓

WH-MVT and stranded preps

Is the dangling preposition of the English sentence correctly placed in the French translation?		
S19a	Source	The guy **that** she is going out **with** is handsome.
	Ref	Le type **avec qui** elle sort est beau.
	PBMT-1	Le mec qu'elle va sortir avec est beau. ✗
	NMT	Le mec qu'elle sort avec est beau. ✗
	Google	Le mec avec qui elle sort est beau. ✓

S19b	Source	**Whom** is she going out **with** these days?
	Ref	**Avec qui** sort-elle ces jours-ci?
	PBMT-1	Qu'est-ce qu'elle allait sortir avec ces jours? ✗
	NMT	À qui s'adresse ces jours-ci? ✗
	Google	Avec qui sort-elle de nos jours? ✓
S19c	Source	The girl **that** he has been talking **about** is smart.
	Ref	La fille **dont** il a parlé est brillante.
	PBMT-1	La jeune fille qu'il a parlé est intelligent. ✗
	NMT	La fille qu'il a parlé est intelligente. ✗
	Google	La fille dont il a parlé est intelligente. ✓
S19d	Source	**Who** was he talking **to** when you left?
	Ref	**À qui** parlait-il au moment où tu est parti?
	PBMT-1	Qui est lui parler quand vous avez quitté? ✗
	NMT	Qui a-t-il parlé à quand vous avez quitté? ✗
	Google	Avec qui il parlait quand vous êtes parti? ✓
S19e	Source	The city **that** he is arriving **from** is dangerous.
	Ref	La ville **d'où** il arrive est dangereuse.
	PBMT-1	La ville qu'il est arrivé de est dangereuse. ✗
	NMT	La ville qu'il est en train d'arriver est dangereuse. ✗
	Google	La ville d'où il vient est dangereuse. ✓
S19f	Source	**Where** is he arriving **from**?
	Ref	**D'où** arrive-t-il?
	PBMT-1	Où est-il arrivé? ✗
	NMT	De quoi s'agit-il? ✗
	Google	D'où vient-il? ✓

Adverb-triggered inversion

Is the adverb-triggered subject-verb inversion in the English sentence correctly rendered in the French translation?

S20a	Source	Rarely **did the dog** run.
	Ref	Rarement **le chien courait-il**.
	PBMT-1	Rarement le chien courir. ✗
	NMT	Il est rare que le chien marche. ✗
	Google	Rarement le chien courir. ✗
S20b	Source	Never before **had she been** so unhappy.
	Ref	Jamais encore **n'avait-elle** été aussi malheureuse.
	PBMT-1	Jamais auparavant, si elle avait été si malheureux. ✗
	NMT	Jamais auparavant n'avait été si malheureuse. ✗
	Google	Jamais elle n'avait été aussi malheureuse. ✓

S20c	Source	Nowhere **were the birds** so colorful.
	Ref	Nulle part **les oiseaux n'étaient** si colorés.
	PBMT-1	Nulle part les oiseaux de façon colorée. ✗
	NMT	Les oiseaux ne sont pas si colorés. ✗
	Google	Nulle part les oiseaux étaient si colorés. ✗

Middle voice

Is the generic statement made in the English sentence correctly and naturally rendered in the French translation?

S21a	Source	Soup **is eaten** with a large spoon.
	Ref	La soupe **se mange** avec une grande cuillère
	PBMT-1	La soupe est mangé avec une grande cuillère. ✗
	NMT	La soupe est consommée avec une grosse cuillère. ✗
	Google	La soupe est consommée avec une grande cuillère. ✗
S21b	Source	Masonry **is cut** using a diamond blade.
	Ref	La maçonnerie **se coupe** avec une lame à diamant.
	PBMT-1	La maçonnerie est coupé à l'aide d'une lame de diamant. ✗
	NMT	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
	Google	La maçonnerie est coupée à l'aide d'une lame de diamant. ✗
S21c	Source	Champagne **is drunk** in a glass called a flûte.
	Ref	Le champagne **se boit** dans un verre appelé flûte.
	PBMT-1	Le champagne est ivre dans un verre appelé une flûte. ✗
	NMT	Le champagne est ivre dans un verre appelé flûte. ✗
	Google	Le Champagne est bu dans un verre appelé flûte. ✗

Fronted "should"

Fronted "should" is interpreted as a conditional subordinator. It is normally translated as "si" with imperfect tense.

S22a	Source	**Should** Paul leave, I would be sad.
	Ref	**Si** Paul **devait** s'en aller, je serais triste.
	PBMT-1	Si le congé de Paul, je serais triste. ✗
	NMT	Si Paul quitte, je serais triste. ✗
	Google	Si Paul s'en allait, je serais triste. ✓
S22b	Source	Should he become president, she would be promoted immediately.
	Ref	**S'**il devait devenir président, elle recevrait immédiatement une promotion.
	PBMT-1	S'il devait devenir président, elle serait encouragée immédiatement. ✓
	NMT	S'il devait devenir président, elle serait immédiatement promue. ✓
	Google	Devrait-il devenir président, elle serait immédiatement promue. ✗
S22c	Source	**Should** he fall, he would get up again immediately.
	Ref	**S'** il venait à tomber, il se relèverait immédiatement.
	PBMT-1	S'il devait tomber, il allait se lever immédiatement de nouveau. ✓
	NMT	S'il tombe, il serait de nouveau immédiatement. ✗
	Google	S'il tombe, il se lèvera immédiatement. ✗

Clitic pronouns

Are the English pronouns correctly rendered in the French translations?		
S23a	Source	She had a lot of money but he did not have **any**.
	Ref	Elle avait beaucoup d'argent mais il n'**en** avait pas.
	PBMT-1	Elle avait beaucoup d'argent mais il n'en avait pas. ✓
	NMT	Elle avait beaucoup d'argent, mais il n'a pas eu d'argent. ✓
	Google	Elle avait beaucoup d'argent mais il n'en avait pas. ✓
S23b	Source	He did not talk **to them** very often.
	Ref	Il ne **leur** parlait pas très souvent.
	PBMT-1	Il n'a pas leur parler très souvent. ✗
	NMT	Il ne leur a pas parlé très souvent. ✓
	Google	Il ne leur parlait pas très souvent. ✓
S23c	Source	The men are watching **each other**.
	Ref	Les hommes **se** surveillent **l'un l'autre**
	PBMT-1	Les hommes se regardent les uns les autres. ✓
	NMT	Les hommes se regardent les uns les autres. ✓
	Google	Les hommes se regardent. ✗
S23d	Source	He gave **it** to the man.
	Ref	Il **le** donna à l'homme.
	PBMT-1	Il a donné à l'homme. ✗
	NMT	Il l'a donné à l'homme. ✓
	Google	Il le donna à l'homme. ✓
S23e	Source	He did not give **it** to **her**.
	Ref	Il ne **le lui** a pas donné.
	PBMT-1	Il ne lui donner. ✗
	NMT	Il ne l'a pas donné à elle. ✗
	Google	Il ne lui a pas donné. ✗

Ordinal placement

Is the relative order of the ordinals and numerals correct in the French tranlation?		
S24a	Source	The **first four** men were exhausted.
	Ref	Les **quatre premiers** hommes étaient tous épuisés.
	PBMT-1	Les quatre premiers hommes étaient épuisés. ✓
	NMT	Les quatre premiers hommes ont été épuisés. ✓
	Google	Les quatre premiers hommes étaient épuisés. ✓
S24b	Source	The **last three** candidates were eliminated.
	Ref	Les **trois derniers** candidats ont été éliminés.
	PBMT-1	Les trois derniers candidats ont été éliminés. ✓
	NMT	Les trois derniers candidats ont été éliminés. ✓
	Google	Les trois derniers candidats ont été éliminés. ✓

S24c	Source	The **other two** guys left without paying.
	Ref	Les **deux autres** types sont partis sans payer.
	PBMT-1	Les deux autres mecs ont laissé sans payer. ✓
	NMT	Les deux autres gars à gauche sans payer. ✓
	Google	Les deux autres gars sont partis sans payer. ✓

Inalienable possession

Is the French translation correct and natural both in: a) its use of a particular determiner on the body part noun; and b) the presence or absence of a reflexive pronoun before the verb?

S25a	Source	He washed **his** hands.
	Ref	Ils **s'**est lavé **les** mains.
	PBMT-1	Ils se lavait les mains. ✓
	NMT	Il a lavé ses mains. ✗
	Google	Ils se lava les mains. ✓
S25b	Source	I brushed **my** teeth.
	Ref	Je **me** suis brossé **les** dents.
	PBMT-1	J'ai brossé mes dents. ✗
	NMT	J'ai brossé mes dents. ✗
	Google	Je me suis brossé les dents. ✓
S25c	Source	You brushed **your** teeth.
	Ref	Tu **t'**es brossé **les** dents
	PBMT-1	Vous avez brossé vos dents. ✗
	NMT	vous avez brossé vos dents. ✗
	Google	Tu as brossé les dents. ✗
S25d	Source	I raised **my** hand.
	Ref	J'ai levé **la** main.
	PBMT-1	J'ai levé la main. ✓
	NMT	J'ai soulevé mamain. ✗
	Google	Je levai la main. ✓
S25e	Source	He turned **his** head.
	Ref	Il a tourné **la** tête.
	PBMT-1	Il a transformé sa tête. ✗
	NMT	Il a tourné sa tête. ✗
	Google	Il tourna la tête. ✓
S25f	Source	He raised his eyes to heaven.
	Ref	Il leva **les** yeux au ciel.
	PBMT-1	Il a évoqué les yeux au ciel. ✓
	NMT	Il a levé les yeux sur le ciel. ✓
	Google	Il leva les yeux au ciel. ✓

Zero REL PRO

Is the English zero relative pronoun correctly translated as a non-zero one in the French translation?

S26a	Source	The strangers the woman saw were working.
	Ref	Les inconnus **que** la femme vit travaillaient.
	PBMT-1	Les étrangers la femme vit travaillaient. ✗
	NMT	Les inconnus de la femme ont travaillé. ✗
	Google	Les étrangers que la femme vit travaillaient. ✓
S26b	Source	The man your sister hates is evil.
	Ref	L'homme **que** votre sœur déteste est méchant.
	PBMT-1	L'homme ta soeur hait est le mal. ✗
	NMT	L'homme que ta soeur est le mal est le mal. ✓
	Google	L'homme que votre sœur hait est méchant. ✓
S26c	Source	The girl my friend was talking about is gone.
	Ref	La fille **dont** mon ami parlait est partie.
	PBMT-1	La jeune fille mon ami a parlé a disparu. ✗
	NMT	La petite fille de mon ami était révolue. ✗
	Google	La fille dont mon ami parlait est partie. ✓

A Challenge Set Approach to Evaluating Machine Translation

Pierre Isabelle and Colin Cherry
National Research Council Canada
first.last@nrc-cnrc.gc.ca

George Foster
Google*
fosterg@google.com

Abstract

Neural machine translation represents an exciting leap forward in translation quality. But what longstanding weaknesses does it resolve, and which remain? We address these questions with a challenge set approach to translation evaluation and error analysis. A challenge set consists of a small set of sentences, each hand-designed to probe a system's capacity to bridge a particular structural divergence between languages. To exemplify this approach, we present an English–French challenge set, and use it to analyze phrase-based and neural systems. The resulting analysis provides not only a more fine-grained picture of the strengths of neural systems, but also insight into which linguistic phenomena remain out of reach.

1 Introduction

The advent of neural techniques in machine translation (MT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014) has led to profound improvements in MT quality. For "easy" language pairs such as English/French or English/Spanish in particular, neural (NMT) systems are much closer to human performance than previous statistical techniques (Wu et al., 2016). This puts pressure on automatic evaluation metrics such as BLEU (Papineni et al., 2002), which exploit surface-matching heuristics that are relatively insensitive to subtle differences. As NMT continues to improve, these metrics will inevitably lose their effectiveness. Another challenge posed by NMT systems is their opacity: while it was usually clear which phenomena were ill-handled by previous statistical systems—and why—these questions are more difficult to answer for NMT.

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

Figure 1: Example challenge set question.

We propose a new evaluation methodology centered around a *challenge set* of difficult examples that are designed using expert linguistic knowledge to probe an MT system's capabilities. This methodology is complementary to the standard practice of randomly selecting a test set from "real text," which remains necessary in order to predict performance on new text. By concentrating on difficult examples, a challenge set is intended to provide a stronger signal to developers. Although we believe that the general approach is compatible with automatic metrics, we used manual evaluation for the work presented here. Our challenge set consists of short sentences that each focus on one particular phenomenon, which makes it easy to collect reliable manual assessments of MT output by asking direct yes-no questions. An example is shown in Figure 1.

We generated a challenge set for English to French translation by canvassing areas of linguistic divergence between the two language pairs, especially those where errors would be made visible by French morphology. Example choice was also partly motivated by extensive knowledge of the weaknesses of phrase-based MT (PBMT). Neither of these characteristics is essential to our method, however, which we envisage evolving as NMT progresses. We used our challenge set to evalu-

*Work performed while at NRC.