Evaluation des traductions automatiques en français des titres de presse japonais.

Raoul Blin

Evaluation des traductions automatiques en français des titres de presse japonais.

Evaluation of machine translations in French of Japanese Newspapers headlines.

Raoul Blin, CNRS-CRLAO[1]

Summary : The primary purpose of this exploratory study is to provide a general assessment of the performances of Japanese-to-French machine translation systems applied to Japanese newspapers headlines. The second purpose is to propose a reasonable quantitative objective to challenge. Japanese newspapers headlines are written using a very particular "sub-language" that differs from standard Japanese. To our knowledge, no system has been evaluated on this type of text for any target language. In this study, we evaluated translations of a small corpus of 350 headlines made using popular translation systems and compared the results with two human translations for each headline. In accordance with a human intuitive evaluation, BLEU scores are very low. They are even lower than a pessimistic estimation. We assume then that it should be easy to obtain better results.

Résumé : Le premier objectif de cette étude exploratoire est de proposer une vue d'ensemble des performances des traducteurs automatiques du japonais vers le français, appliqués à la traduction des titres de presse en ligne. Le second objectif est de proposer un objectif raisonnable à atteindre, compte tenu de l'état de l'art. Les titres de presse japonais sont écrits dans un « sous-langage » différent du japonais standard. A notre connaissance, aucun système n'a été évalué sur ce type de texte, quelle que soit la langue cible. Dans cette étude, nous évaluation les traductions de 350 titres fournis par des systèmes de traduction grand publics et comparons chaque traduction à deux traductions humaines pour chaque article. Conformément à l'intuition, les scores BLEU sont très faibles, et même plus faibles que les estimations les plus pessimistes. Nous montrons qu'il y aurait moyen, à peu de frais, d'obtenir de meilleurs résultats.

Dans ce papier, nous présentons une première évaluation des traductions du japonais vers le français des titres de presse en ligne. En l'absence d'études antérieures comparables, l'objectif est avant tout d'évaluer la situation actuelle, et de débattre des progrès envisageables. Pour cela, nous passons tout d'abord en revue une sélection d'études qui fournissent des résultats sur des tâches apparentées à la nôtre. Ces données permettent de fixer des objectifs chiffrés accessibles. Puis nous évaluons les traductions de quelques systèmes existants sur un petit corpus. Pour finir, nous discuterons des résultats en les comparant aux objectifs avancés dans la première section.

1 Etat de l'art

La traduction automatique directe du japonais vers le français est inexistante. Il existe des services commerciaux gratuits en ligne mais l'examen des traductions produites laisse penser qu'ils utilisent l'anglais comme langue pivot. Les traducteurs génériques comme Moses (Koehn et al. 2007) par exemple n'ont à notre connaissance pas été utilisés pour traduire du japonais vers le français, et a fortiori pour traduire des titres de presse japonais. La cause est certainement l'absence de corpus bilingue aligné. Une autre cause est, peut-être, le manque d'intérêt de la communauté pour ce couple de langues.

La traduction automatique de titres de presse japonais vers d'autres langues n'a pas fait à notre

1 blin@ehess.fr

connaissance l'objet de travaux spécifiques. Et, ce qui est significatif, on ne trouve pas de corpus monolingue ou bilingues de titres de presse, quelle que soit la langue cible. Même le corpus aligné bilingue Reuters (Utiyama and Isahara 2003) de l'agence de presse Reuter ne contient pas de titres à notre connaissance. Ceci est pourtant indispensable car le langage des titres de presse japonais se démarque du reste de la langue écrite et nécessite certainement des corpus indépendants.

La traduction automatique des titres de presse du japonais vers le français est donc un tout nouveau domaine de travail.

Dans un tel contexte, l'absence de *campagne d'évaluation* des performances de traducteurs automatiques direct du japonais vers le français n'est pas surprenant. A défaut de disposer de repères établis, nous allons dans cette revue de l'état de l'art plutôt chercher à esquisser ce que pourraient être les performances de tels traducteurs. Pour cela, nous exploitons des résultats existants obtenus sur des corpus s'approchant du corpus de titres, et sur des couples de langue en relation avec le français et le japonais. Comme les résultats sont obtenus à partir de corpus différents, il n'est pas possibles de les comparer directement entre eux. Par contre, ils sont suffisants pour établir une fourchette de valeurs.

La traduction automatique japonais <-> anglais est très active et fournit de nombreux points de repères. Nous ne citerons qu'une seule série de scores, globalement assez élevés au regard de ce que nous obtiendrons dans nos tests. Nous reprenons les résultats présentés dans Neubig & Duh (Neubig and Duh 2014). Ils comparent plusieurs systèmes entraînés et testés sur un corpus de brevets (Goto et al. 2011), qualifié par les auteurs de "référence" dans le domaine de la traduction automatique. Le corpus regroupe environ 3,2 millions de phrases. [2]

Tableau 1 : Scores BLEU (Papineni et al. 2002) pour des traducteurs japonais->anglais (Neubig and Duh 2014)

	bleu-4 (mteval-v13a)	
	Min	Max
Japanese-To-English	0.2941	0.3370

Ces chiffres sont intéressants parce qu'ils sont obtenus à partir d'un genre de textes qui, à défaut d'être identique à celui des titres de presse, n'est du moins pas sans rapport avec eux. Il s'agit d'une prose relativement condensée, même si elle ne l'est certainement pas autant que les titres de presse. Ce sont des textes assez formels, qui recourent abondamment à du vocabulaire sino-japonais. Les performances que l'on peut espérer d'un système de traduction de titres devraient idéalement s'approcher des valeurs minimum annoncées ici.

Reste à estimer les valeurs maximum que l'on peut raisonnablement attendre lorsque la langue cible est le français. Une comparaison entre couples de langues d'Europe (Koehn, Birch, and Steinberger 2009) montre que les meilleurs scores sont obtenus avec des langues morphologiquement et syntaxiquement proches. Une traduction du japonais vers le français, langues non apparentées et très différentes, ne peuvent donc espérer atteindre des scores comparables.

Tableau 2 : Scores pour les traductions sur le corpus JRC-Acquis (Steinberger et al. 2006), pour des langues apparentées au français.

	BLEU

2 Le fait que ce soit entraine sur le corpus même justifierait quand même le "bon" score.

espagnol - français	0.532
portugais - français	0.527

Les scores obtenus annoncés dans Koehn et al. sont globalement très élevés par rapport à l'ensemble des scores observés dans d'autres études. Nous rediscutons de ce point plus bas dans cette section.

Si l'on travaille avec des traducteurs partiellement ou entièrement statistiques, il faut aussi s'attendre à ce que les scores soient meilleurs pour des langues et couples de langues disposant de grands corpus parallèles. Avec un couple de langues plus éloignées mais au moins aussi riches en corpus, les résultats sont un peu inférieurs comme le montrent les résultats de cette même étude de l'anglais vers le français.

Tableau 3 : Scores pour les traductions sur le corpus JRC-Acquis (Steinberger et al. 2006), pour une langue bien renseignée.

	BLEU
anglais - francais	0.501

Les scores annoncés dans (Steinberger et al. 2006) sont globalement très élevés. A titre de comparaison, sur le couple anglais-français, (Bertoldi et al. 2013) obtient avec le IWSLT TED test sets des scores se situant « seulement » entre 02635 et 03951, soit entre environ 25 et 10 points de moins. Si l'on applique cette réduction aux couples de langues apparentées, cela situerait les scores dans une fourchette de 0.287 et 0.432. Le score maximum que l'on peut espérer serait donc aux alentours de 0.43.

Voyons d'autres évaluations de traductions de couples de langues non apparentées et ayant pour langue de départ une langue proche de la langue des titres de presse japonais. Nous nous en tiendrons au chinois et au coréen, que nous choisissons à dessein. En effet, le coréen est réputé la langue (hors langues de Ryukyu) la plus proche du japonais. Le chinois s'en éloigne sensiblement, sauf sur le cas très particulier qui nous intéresse des titres de presse japonais. En effet, les titres de presse japonais recourent massivement au sous-langage sino-japonais, apparenté au chinois, même si il faut rester très prudent et ne pas confondre les deux.

Tableau 4 : Extrait des Résultats de la campagne openMT12, traduction vers l'anglais, Progress test Chinese-To-English Progress Test New Source Subset - "English-True" source ; résultats « Overall » ; (143 documents containing 1537 segments) / Constrained [3]

	bleu-4		NIST		METEOR	
	min	Max	Min	Max	Min	Max
Chinese-To-English	0.2395	0.3156	7.6360	9.0518	0.5224	0.5639
Korean-To-English	0.0679	0.1113	2.2916	5.4745	0.2915	0.4397

Les scores BLEU-4 et NIST(Doddington 2002) sont calculés avec (mteval-v13a [4]). METEOR (Denkowski and Lavie 2014) est calculé avec meteor-0.7.

Le score pour le coréen est étonnamment bas et n'est pas expliqué par les auteurs.

L'ensemble des résultats, toutes campagnes confondues, excepté le score extrêmement bas du couple coréen-anglais, convergent vers une valeur que nous situons dans la fourchette de 0.23 -

3 NIST 2012 Open Machine Translation Evaluation (OpenMT12)
Official Release of Results; 2013; http://www.nist.gov/itl/iad/mig/openmt12results.cfm

4 *ftp://jaguar.ncsl.nist.gov/mt/resources/**mteval**-**v13a**.p*

0.31.

L'ensemble des résultats présentés ci-dessus sont obtenus à partir de campagnes avec entraînement. Les systèmes évalués sont en quelques sortes spécialisés: un corpus est divisé en deux, une partie sert à l'entraînement et l'autre au test. Les deux corpus sont donc très proches. Au contraire, les services que nous utilisons dans le présent travail sont généralistes mais nous allons les tester sur des corpus très spécifiques morphologiquement et syntaxiquement, et très différents du corpus généraliste. Compte tenu de ce handicap, il faut s'attendre à ce que les scores obtenus soient inférieurs à ceux existants. Supposons que les valeurs soient inférieures de 10 points. Globalement, les scores observés se situent majoritairement dans la fourchette de 0.2 à 0.3. Le maximum envisageable est de 0.43. Nous nous attendons donc à obtenir des résultats situés dans la fourchette "pessimiste" de 0.1 à 0.2. Dans tous les cas, ils ne devraient pas dépasser 0.33 .

2 Tests

2.1 Les traducteurs testés

Pour des raisons matérielles, nous avons limité nos tests à des services en ligne, qui ne nécessitent pas d'entraînement. En effet, comme nous l'avons indiqué plus haut, il n'existe pas de corpus bilingue japonais-français de taille raisonnable. Nous ne pouvions donc, comme le font de nombreuses études, construire un dispositif par exemple à base de Moses puis l'entraîner de façon satisfaisante. Pour la traduction japonais->français, nous avons utilisé les deux services commerciaux Google Translate[5] et Bing Translator[6], retenus parce qu'ils sont les plus connus et certainement très employés. Nous avons utilisé ces deux services aussi pour une traduction japonais->anglais. Pour cette même traduction japonais->anglais, nous avons en plus eu recours à un troisième service, SYSTRAN[7].

L'usage de ces trois services est libre en ligne. Nous supposons que la version disponible est à chaque fois la plus récente. Les descriptions concernant les moteurs de traduction de ces services commerciaux sont très succintes. A partir des quelques informations disponibles et au regard des sorties de traductions, nous pouvons faire plusieurs suppositions :

- Google Translate[8] et Bing-Translator sont à base de statistiques. Les systèmes sont entraînés sur du corpus "tout venant" (les pages Web collectées par les services associés au service de traduction).

- SYSTRAN est à base de transfert et utilise donc des règles ((Yang and Lange 2003) cité par (Russo 2010)).

- Etant donnée la nature des erreurs, on peut avancer que la traduction du japonais vers le français pour Google Translate fonctionne à l'aide d'une langue pivot, l'anglais. On peut avancer que Bing suit la même stratégie.

Ces services n'acceptent pas sur des textes de grande taille. Au delà de 100 titres, les services ne fonctionnaient plus correctement. Nous avons donc à chaque fois limité le corpus à 400 titres ou phrases, chacun sur une ligne, qu'il a fallu en réalité faire traduire 100 par 100.

2.2 Evaluation des traductions japonais->français des titres

Nous testons un corpus bilingue japonais-français de 358 titres de presse japonais. Ces titres sont

5 *translate.google.com/*
6 *www.**bing**.com/**translator***
7 http://www.systran.fr/lp/traduction-en-ligne/
8 Quelques informations sur la page : http://translate.google.com/about/intl/en_ALL/

collectés sur les versions en ligne de 5 journaux généralistes différents : Asahi[9], Mainichi[10], Nikkei[11], Sankei[12], Yomiuri[13] et Akahata[14]. Il faut ajouter à cela quelques titres du site de la NHK[15]. Ce sont tous des journaux nationaux. Sauf le Akahata, journal du Parti Communiste Japonais de diffusion relativement limitée (exclusivement sur abonnement), tous ces journaux ont des forts tirages (de 4 à 10 millions par jour). A eux seuls, ils représentent une part importante de la presse nationale et peuvent donc être considérés comme représentatifs du langage de la presse.

Le corpus est divisé en trois: un corpus de 216 titres, avec une représentation équilibrée de tous les journaux, un corpus constitué des 71 titres du Mainichi et un corpus de 71 titres du Nikkei.

Ces trois corpus ont été constitués par échantillonnage de la collection de tous les titres édités en ligne par ces journaux, sur la période allant de avril 2011 à juillet 2013. Ce corpus comprend environ 70 000 titres japonais.

Pour chaque titre du corpus bilingue, deux traductions sont disponibles. Une première est une traduction littérale (« t.littérale »). Elle est aussi proche que possible de la structure du titre original tout en respectant la syntaxe du français et la compréhensibilité par un locuteur français. La seconde traduction (« t.propre ») se veut proche du titre à la française. Cette traduction respecte le titre original sur le fond et en reprend le maximum d'informations.

Exemple	:汚染水見えぬルート	…海と地下水、関連薄く[16]
(litt.)	Eau contaminée, une fuite invisible	... la mer et les eaux souterraines, faible relation.
(« propre »)	Peu de chance que l'eau contaminée ait fui par la mer ou les eaux souterraines	

L'évaluation porte sur les traductions de ces titres japonais par deux services de traduction automatique. Les traductions machine ont été comparées aux deux traductions humaines. Avec les mêmes outils d'évaluation, nous avons en plus mesuré la proximité entre la traduction littérale et la traduction "propre".

Nous proposons trois scores mais les discussions ne porteront que sur BLEU. Les évaluations BLEU et NIST sont obtenues avec mteval-11b [17]. Le score METEOR est obtenu avec meteor-1.5.

Tableau 5 : Evaluation des traductions japonais-français de 358 titres de presse japonais et de la proximité des traductions littérales et traductions propres.

	BLEU		NIST		METEOR	
	t. littérale	t. propre	t. litt.	t. propre	t.litt	t. propre
BING	0.0531	0.0355	2.7755	2.5901	0.1266	0.1030
Google Translate	0.0665	0.0452	3.3773	2.6366	0.1328	0.1140
comp litt-propre	0.3205		6.0279		0.4214	

On relève que l'évaluation de la proximité des deux traductions n'est pas très élevée.

9 http://www.asahi.com
10 http://mainichi.jp
11 http://www.nikkei.jp
12 http://sankei.jp.msn.com
13 http://www.yomiuri.co.jp
14 http://www.jcp.or.jp/akahata/
15 http://www3.nhk.or.jp/news/
16 Yomiuri, 2011/04/01, 22H39
17 ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

2.3 Evaluation des traductions japonais->anglais

Pour contextualiser les performances sur le couple japonais->français, nous évaluons les mêmes services de traduction sur le japonais->anglais.

Nous testons une partie du corpus aligné de Reuter (Utiyama and Isahara 2003), constitué de phrases extraites de dépêches de l'agence de presse Reuters. Nous nous en tenons à 400 phrases, les systèmes en ligne ayant une capacité très limitée, il n'est pas possible de soumettre les plus de 50 000 phrases du corpus.

Les outils d'évaluation sont les mêmes que pour le couple japonais->français. A titre indicatif, nous ajoutons l'évaluation pour le logiciel SYSTRAN, réputé utiliser des règles, ce qui permet de comparer l'approche à base de règle et l'approche statistique de Google Translate et Bing-Translator.

Tableau 6 : Evaluation des traductions japonais-anglais des 400 premières phrases du corpus aligné Reuters

	BLEU	NIST	METEOR
BING	0.1261	3.9732	0.2137
Google Translate	**0.1509**	**5.0037**	**0.2772**
SYSTRAN	0.0626	3.2269	0.1665

3 Analyse

Globalement, les résultats obtenus dans la présente étude sont inférieurs, voire très inférieurs à ceux obtenus dans les études antérieures et présentés dans la section 1. Même les résultats pour les traductions vers l'anglais obtiennent un score inférieur.

Pour la tâche spécifique de traduction des titres japonais vers le français, les scores sont plus bas encore. Ils n'entrent même pas dans la fourchette « pessimiste » (0.1 - 0.2) que nous avions proposé au regard des études existantes.

Il est possible que cela résulte d'un biais dû aux procédure d'évaluation. En effet, nous évaluons les systèmes sur des corpus qui n'ont aucun rapport avec les corpus sur lesquels ils sont entraînés, et qui sont certainement des corpus « tout venant »,collectés sur l'internet. Au contraire, les résultats des études antérieures sont en général obtenus avec des systèmes qui ont été préalablement entraînés sur un corpus similaire à celui sur lequel ils sont testés. En principe, ces derniers sont donc sensiblement avantagés par rapport à ceux de nos test. Il est cependant possible qu'un outil comme Google Translate compense cette faiblesse grâce à des services annexes, comme l'évaluation humaine par les utilisateurs ou même en exploitant leurs mémoires de traduction[18].

Notre tâche de traduction vers le français porte exclusivement sur les titres. Si les systèmes testés sont entraînés sur des corpus « tout venant », sachant que les titres de presse n'occupe qu'une très petite part, leur faible score n'est pas étonnant.

Il existe une autre explication aux faibles scores obtenus pour la traduction vers le français. L'examen qualitatif des résultats et des erreurs montrent que de toute évidence, les systèmes testés utilisent l'anglais comme langue pivot. Les risques d'erreurs s'en trouvent augmentés.

Il est à noter que la traduction vers l'anglais entre dans la fourchette pessimiste. Cela tend à légitimer cette fourchette.

Le traducteur Google Translate est systématiquement meilleur que Bing-translator, pour toutes les tâches.

La comparaison entre Google Translate et Bing-Translator d'un coté, et Systran est a priori défavorable à ce dernier. Si l'on considère Systran comme représentatif d'une approche mixte régulée et statistique de la traduction automatique, alors l'approche tout statistique se révèlerait

18 https://support.google.com/translate/toolkit/answer/147863?hl=fr

préférable.

Toutefois, il faut relativiser ce résultat car il est peut être biaisé. En effet, le corpus aligné Reuters est disponible en ligne et aura pu servir de corpus d'entraînement aux deux traducteurs statistiques. Cela pourrait justifier leur meilleur score.

4 Conclusion et perspectives

Nous avons présenté quelques résultats permettant de grossièrement situer où en était la traduction automatique de titres de presse du japonais vers le français. Nous avons pris comme références des traducteurs automatiques grand public et utilisé des outils d'évaluation traditionnels en traduction automatique (scores BLEU, NIST, METEOR). Les résultats obtenus sont très inférieurs à tout ce qui se fait actuellement et même en deça de de ce que l'on pourrait attendre. Ils sont cohérents avec une estimation humaine "intuitive" qui est que ces traducteurs fournissent de mauvaises traductions pour les titres japonais vers le français.

Les outils d'évaluations ne permettent pas de localiser les difficultés. On peut néanmoins faire l'hypothèse suivante. Les titres de presses japonais recourent à un sous-langage partiellement original. Les systèmes actuels de traduction sont tous massivement à base de statistiques et d'apprentissage automatique. Leurs performances dépendent donc de la qualité du corpus d'entraînement et de sa proximité avec le corpus test. Or on peut supposer que les services testés ici sont entraînés sur des corpus où le sous-langage des titres de presse n'est que très faiblement représenté, d'où leur difficulté à le traiter.

Les scores sont si bas qu'on peut espérer, à peu de frais, parvenir à élaborer un dispositif capable d'atteindre une fourchette de valeurs minimales que nous estimons entre 01 et 0.2 et dont

5 Références

Bertoldi, Nicola, M.Amin Farajian, Prashant Mathur, Nicholas Ruiz, and Marcello Federico
2013 FBK's Machine Translation Systems for the IWSLT 2013 Evaluation Campaign. *In* In Proceedings of the International Workshop on Spoken Language Translation 2013. Heidelberg, Germany.

Denkowski, Michael, and Alon Lavie
2014 Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *In* Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.

Doddington, George
2002 Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. *In* Proceedings of the Second International Conference on Human Language Technology Research Pp. 138–145. HLT '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=1289189.1289273.

Goto, Isao, Lu Bin, Po Chow Ka, Eiichiro Sumita, and Benjamin K. Tsou
2011 Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. *In* Proceedings of NTCIR Pp. 559–57?

Koehn, Philipp, Alexandra Birch, and Ralf Steinberger
2009 462 Machine Translation Systems for Europe. *In* MT Summit XII.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, et al.
2007 Moses: Open Source Toolkit for Statistical Machine Translation. *In* Annual Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic.

Neubig, Graham, and Kevin Duh
2014 On the Elements of an Accurate Tree-to-String Machine Translation System. *In* The 52nd

Annual Meeting of the Association for Computational Linguistics. Baltimore, USA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu
2002 BLEU: A Method for Automatic Evaluation of Machine Translation. *In* Proc. ACL Pp. 311–318.

Russo, Lorenza
2010 La Traduction Automatique Des Pronoms Litiques. Quelle Approche Pour Quels Résultats? *In* TALN 2010. Montréal.

Steinberger, Ralf, B Pouliquen, A Widiger, et al.
2006 The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *In* LREC.

Utiyama, Masao, and Hitoshi Isahara
2003 Reliable Measures for Aligning Japanese-English News Articles and Sentences. *In* ACL-2003 Pp. 72–79.

Yang, J., and E. Lange
2003 Going Live on the Internetdans. *In* Computers and Translation. A Translator's Guide. John Benjamins Publishing Company. Pp. 191–210.