La tâche de prédiction de la qualité. Session 5 - Mesure de la qualité en TA

Guillaume Wisniewski, François Yvon

TRALOGY

La tâche de prédiction de la qualité

Guillaume Wisniewski
Université Paris Sud et LIMSI–CNRS
wisniews@limsi.fr

François Yvon
Université Paris Sud et LIMSI–CNRS
yvon@limsi.fr

TRALOGY II - Session 5
Date d'intervention : 18/01/2013

La diffusion des systèmes de traductions automatiques est limitée par leur manque de fiabilité : la qualité de traduction varie beaucoup, parfois de manière imprévisible. Une manière de contourner cette limite serait, pour les systèmes de traduction automatique, de prédire, en même temps que la traduction, une mesure numérique de sa qualité. Cette prédiction de qualité permettrait de fournir une information comparable au pourcentage de correspondance dans une mémoire de traduction, qui est usuellement utilisé en traduction assistée par ordinateur. Cette information pourrait, par exemple, être utilisée par des traductions pour corriger les traductions automatiques plus efficacement en se concentrant sur les phrases les plus problématiques. Dans cette contribution, nous présenterons la manière dont la tâche de prédiction de qualité est généralement abordée par les chercheurs en traduction automatique. Nous décrirons notre contribution à la première campagne internationale d'évaluation de prédiction de qualité et, à partir de cette expérience, mettrons en évidence les difficultés de cette tâche.

http://webcast.in2p3.fr/videos-quality_estimation_measures_for_statistical_machine_translation

Introduction

Durant ces dernières années, les systèmes de traduction automatique (TA) se sont grandement améliorés, en particulier du fait de la maturation d'une nouvelle génération d'outils de TA. Ces nouveaux outils, qui reposent sur l'exploitation, par des techniques d'apprentissage statistique, de vastes corpus de textes bilingues parallèles (pouvant contenir des dizaines de millions de phrases), permettent d'atteindre une qualité moyenne suffisante pour de nombreux usages (voir (Koehn 2010) pour une présentation récente de ces systèmes). Ils présentent l'intérêt de traiter indifféremment tous les types de textes, indépendamment de leur domaine, genre, registre ; dans leurs évolutions les plus récentes, ils s'appliquent également aux énoncés oraux. Il est de surcroît relativement aisé, pour peu que des ressources idoines soient disponibles, d'améliorer encore les performances de ces systèmes généralistes en les spécialisant (par adaptation et enrichissement) à des sous-langages particuliers.

Les usages de la TA évoluent ainsi rapidement, et les traductions artificielles sont de plus en plus utiles, non seulement dans une optique d'assimilation, mais également dans une optique de publication. Il n'est donc pas surprenant que ces outils rencontrent un engouement croissant de la part d'agence de traductions et de traducteurs professionnels (voir, par exemple, (Garcia 2011) et les références citées). Leur dissémination auprès des industriels est toutefois freinée par le manque de fiabilité des traductions automatiques, dont la qualité est susceptible de varier grandement d'une phrase (ou d'un texte) à l'autre, et ce de manière très imprévisible. Le scénario le plus courant d'utilisation de ces traductions automatiques est donc celui dans lequel un traducteur humain reste en charge de l'édition finale et de la publication des traductions, par *post-édition* des propositions de la machine (Roturier 2009).

Dans un contexte d'aide à la traduction ou de post-édition, il semble utile pour les utilisateurs de TA de disposer de mesures automatiques de la qualité des traductions automatiques ainsi produites. Ces mesures de confiance pourraient, par exemple, permettre aux traducteurs de se concentrer sur les segments les moins bien traduits par la machine, ou inversement d'organiser leurs tâches de traduction en fonction inverse des scores de qualité. Lorsque la traduction automatique fournit un premier jet suffisamment bon, il est en fait possible de post-éditer la traduction *sans revenir au texte source*, ce qui semble permettre des gains de productivité substantiels (Mirko Plitt 2010) ; à l'inverse, post-éditer des segments à partir d'une mauvaise traduction demande de repartir du texte source et implique un effort qui s'avère peu rentable par rapport à une traduction produite directement depuis la source (Garcia 2011). D'autres cas d'usage de ces mesures de confiance consisteront par exemple à les utiliser pour sélectionner une traduction automatique parmi plusieurs choix, produits des systèmes différents, qui seraient proposés au correcteur (He et al. 2010) ; ou encore à les exploiter dans des scénarios de traduction interactive (González Rubio, Ortiz Martínez et Casacuberta 2010). Il est finalement notable que de telles mesures de confiance soient aujourd'hui proposées dans des environnements professionnels : c'est en particulier le cas du *trustscore* de SDL (Soricut, Bach et Wang 2012).

Le calcul de mesures de confiance, de quelque manière qu'on l'envisage, s'avère toutefois être une tâche remarquablement difficile, presqu'aussi difficile que l'évaluation de la traduction automatique elle-même. C'est ce que nous essayons de montrer dans ce travail, en nous appuyant en particulier sur le développement de telles mesures. Le reste de cet article est organisé de la manière suivante : nous commencerons, dans la section 2, par décrire la tâche d'estimation de confiance et par introduire rapidement le fonctionnement des systèmes de traduction automatique ; puis, dans la section 3, nous présenterons le système de prédiction de confiance que nous avons réalisé dans le cadre de la campagne d'évaluation WMT et évaluerons les performances de celui-ci. Finalement, dans la section 4, nous décrirons plusieurs expériences que nous avons réalisées afin de pouvoir interpréter nos résultats, et qui mettent en évidence la difficulté de la tâche de prédiciton de qualité.

1. L'estimation de confiance

Cette section présente la tâche d'estimation de la confiance telle qu'elle est le plus couramment envisagée dans des travaux portant sur la traduction automatique. Nous commençons par rappeler quelques aspects essentiels du fonctionnement des systèmes de TA statistiques[1] qui justifient dans une certaine mesure l'angle sous lequel ces travaux sont conduits, en particulier en ce qui concerne les grandeurs qui sont utilisées pour prédire la confiance. Nous formalisons ensuite la tâche, en présentant et discutant les diverses instanciations qui en ont été proposées dans la littérature.

1.1 La traduction automatique statistique

1.1.1 Principes de la traduction statistique

Les systèmes de traduction statistique les plus performants effectuent la traduction de phrases isolées par assemblage et recombinaison d'unités de traductions correspondant à des petits groupes de mots (des segments dans la terminologie du domaine).

La première étape du développement de tels systèmes consiste alors à construire des inventaires de segments en langue source, accompagnés d'une ou plusieurs traductions dans la langue cible. Ces inventaires sont produits par analyse statistique de grands ensembles de textes parallèles, l'intuition étant que deux segments sources et cibles qui apparaissent dans des phrases parallèles ont de bonnes chances d'être des traductions mutuelles. Chaque paire de segments ainsi détectée est donc nantie d'une grandeur numérique, qui est d'autant plus forte que la cooccurrence entre segments sources et cibles est grande et que l'association est donc plausible. Le résultat de cet apprentissage est un dictionnaire répertoriant des traductions possibles de segments de taille variable, ainsi que leur plausibilité respective : l'ensemble de ces associations constitue le *modèle de traduction*.

L'élaboration d'une traduction d'une nouvelle phrase source devra alors envisager toutes les manières de la découper en segments disjoints, ainsi que toutes les façons dont chacun de ces segments peut être traduit, pour retenir l'hypothèse de traduction complète qui réalise le meilleur assemblage global à partir de ces bonnes traductions locales. Ce problème combinatoire se trouve grandement complexifié par la nécessité de modéliser des divergences syntaxiques entre langues, qui impliquent en surface des structures de phrases différentes et, par ricochet, des changements de position (des *distortions*) des mots entre les phrases source et cible. Dit autrement, il n'est pas possible de traiter les mots et les segments sources dans l'ordre dans lequel ils se présentent, et la recherche d'une traduction optimale passe également par l'exploration d'un ensemble (lui aussi combinatoirement grand) de réarrangements (*réordonnancements*) possibles. Pour sélectionner la meilleure manière d'assembler les segments en langue cible, d'autres grandeurs numériques, qui permettent d'évaluer les meilleures recombinaisons, doivent alors être utilisées.

Deux nouveaux types de modèles jouent ici un rôle : les *modèles de distortion*, qui évaluent la plausibilité des déplacements entre source et cible et qui sont appris par analyse de corpus parallèles. Plus importants encore, les *modèles de langue*, estimés à partir d'immenses corpus monolingues, permettent d'évaluer les séquences de mots en langue cible : les séquences pour lesquelles l'évaluation du modèle de langue est grande sont, en général, plus conformes, du moins en surface, aux prescriptions de la syntaxe. Les plus usités de ces modèles de langue sont les modèles *n*-gramme, qui élaborent le score d'une phrase en agrégeant des mesures de conformité calculées sur de petites fenêtres réduites à *n* mots adjacents.

(1) Pour une présentation plus complète, on se reportera à (Lopez 2008 ; Koehn 2010) ou, en langue française, à (Allauzen et Yvon 2011).

Au final, la meilleure traduction d'une phrase cible est produite en explorant un grand ensemble d'alternatives de traductions possibles, correspondant à de multiples manières de segmenter, de réordonner et de traduire les fragments sources. Cette exploration est guidée par les multiples paramètres numériques qui sont associés à ces diverses décisions de segmentation, de réordonnancement et de traduction.

1.1.2 Mesurer la qualité : métriques

Un dernier ingrédient s'avère nécessaire pour développer des systèmes de traduction statistique, à savoir des mesures automatiques de la qualité des traductions ainsi produites. Ces mesures fournissent au système un *feed-back* indispensable pour déterminer les multiples méta-paramètres qui régulent le fonctionnement complet du système et pondèrent l'influence respective des différents modèles qui y sont mobilisés.

Mesurer automatiquement la qualité d'une traduction est malheureusement un problème excessivement ardu qui a fait l'objet de multiples études (voir, par exemple, (Hovy, King et Popescu-Belis 2002) et les références citées dans cet article). Une solution simple et pratique s'est progressivement imposée, consistant à jauger la qualité d'une hypothèse de traduction en la comparant (superficiellement) avec une traduction de référence. Selon cette (courte) vue, une traduction est d'autant meilleure qu'elle ressemble à une traduction correcte. Différentes fonctions de comparaison ont été proposées dans la littérature : ainsi le score BLEU (Papineni et al. 2002) repose (en première approximation) sur le calcul de la proportion de segments de l'hypothèse qui se trouvent également dans la référence : plus cette proportion est élevée, meilleure sera jugée l'hypothèse. La principale innovation introduite par le score METEOR (Banerjee et Lavie 2005) consiste à autoriser des correspondances « floues » entre l'hypothèse et la référence, permettant ainsi de ne pas (trop) pénaliser des différences portant sur des mots synonymes ou morphologiquement apparentés. TER et hTER (Snover et al. 2006) reposent également sur une comparaison de surface entre référence et hypothèse : TER généralise la distance d'édition entre séquences (Wagner et Fischer 1974) en introduisant une nouvelle opération consistant à déplacer des groupes de mots ; hTER s'intéresse à calculer des différences entre une hypothèse de traduction et une traduction correcte déduite par post-édition. Les scores hTER peuvent donc s'interpréter comme le coût (évalué par le nombre de mots ou de segments à transformer ou à déplacer) nécessaire pour transformer la sortie d'un système de TA en une traduction correcte.

La question de savoir à quel point ces métriques automatiques fournissent des approximations réalistes de la qualité de traduction reste largement ouverte et de nouvelles métriques continuent d'être proposées (voir par exemple (Lavie et Przybocki 2009)) ainsi que les articles qui y sont référencés). Il est admis qu'une bonne métrique automatique doit reproduire au mieux les jugements de qualité exprimés par des humains : à cette aune, BLEU ne semble avoir une corrélation acceptable avec les évaluations humaines que lorsque l'on agrège ces évaluations au niveau d'un document complet ; par comparaison, (h)TER et METEOR semblent fournir de meilleures approximations au niveau des phrases.

1.2 L'estimation de confiance

Il est maintenant possible de formaliser le problème de l'estimation de confiance : dans sa forme la plus générale, il consiste à associer, à tout ou partie d'une traduction automatique, une évaluation numérique rendant compte de sa qualité (ou de son utilisabilité ou de toute autre forme d'appréciation que l'on pourrait vouloir porter).

Supposons que l'on puisse représenter chaque énoncé (ou fragment d'énoncé) par un ensemble de descripteurs numériques (noté x) et que l'on note y la mesure de confiance étudiée ; supposons également que l'on dispose d'un ensemble d'exemples de couples $\{(x_i, y_i)\}$, c'est-à-dire d'énoncés dont la qualité est connue. Dans ce cadre, qui est le plus usuellement considéré,

l'estimation de confiance se formalise comme un problème d'apprentissage automatique, visant à inférer à partir d'exemples une relation fonctionnelle $y\hat{} = f(x)$, où f appartient à une classe connue (par exemple les fonctions linéaires), et est choisie de façon que les prédictions $y\hat{}i = f(xi)$ soient les plus proches possibles des valeurs réelles yi .

Ce cadre générique se décline de multiples manières selon (i) les unités dont on évalue la confiance (des mots aux documents), (ii) les descripteurs utilisés pour représenter ces unités, et (iii) les mesures de qualité utilisées. La figure 1, tirée de (Bach, Huang et Al-Onaizan 2011), illustre l'utilisation, à des fins de visualisation, de mesures de confiance au niveau des mots : sur ce graphe, on peut voir que trois scores de confiance au niveau des mots sont représentés par des tailles et des couleurs différentes.

Concernant le choix des unités au niveau desquelles la confiance est calculée, les unités les plus souvent considérées sont le mot et la phrase, et c'est également le point de vue que nous détaillons ci-dessous, même si des tentatives ont également été faites pour évaluer la qualité au niveau des segments (Blatz et al. 2004 ; Gispert et al. 2012) ou au niveau du document entier (Soricut et Echihabi 2010).

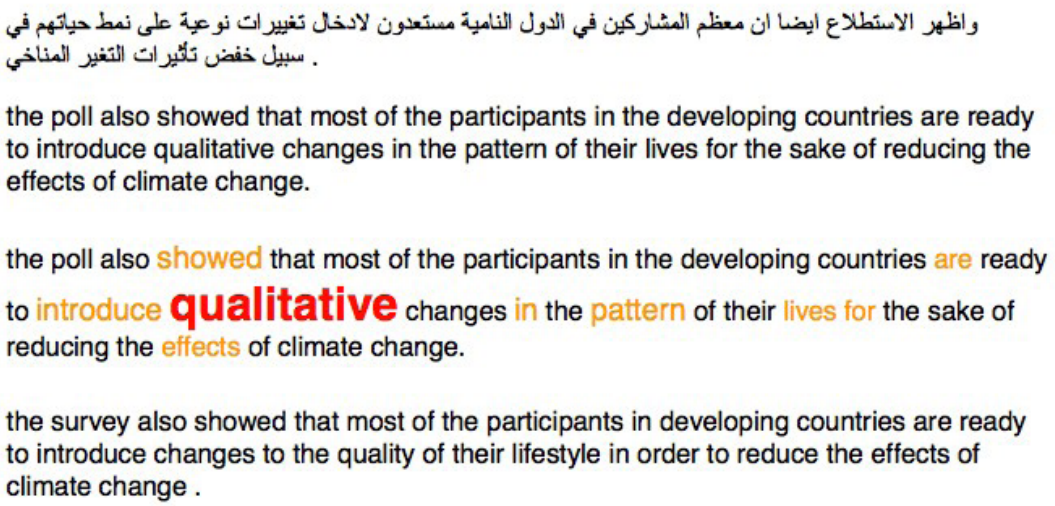

Figure 1 – Visualisation de la confiance au niveau des mots

La figure représente de haut en bas la phrase source, la traduction automatique, la même traduction automatique dans laquelle les mots « suspects » sont mis en évidence par un code couleur, finalement la traduction après post-édition. Extrait de (Bach, Huang et Al-Onaizan 2011).

Concernant le choix des fonctions caractéristiques, on distingue en général entre les caractéristiques *internes*, qui sont dérivées des grandeurs que manipule le système de traduction pour construire sa meilleure hypothèse, et les caractéristiques *externes*, correspondant à des descripteurs et qui sont calculées indépendamment de tout système. Deux arguments plaident pour la préférence donnée, dans la suite (comme dans la littérature scientifique), aux seconds types de descripteurs : (i) ils peuvent être calculés pour toute hypothèse de traduction[2], sans référence à la manière dont elle a été produite ; (ii) les scores internes ne sont généralement pas très fiables : s'ils l'étaient, ils permettraient certainement de produire de meilleures hypothèses.

(2) Y compris pour des traductions humaines !

1.2.1 Prédire la confiance au niveau des mots

En dépit des efforts notamment de (Blatz et al. 2004 ; Ueffing et Ney 2005 ; Raybaud, Langlois et Smaïli 2011 ; Bach, Huang et Al-Onaizan 2011), la prédiction de la confiance au niveau des mots d'une hypothèse de traduction reste un exercice particulièrement difficile. Cette tâche demande d'étiqueter chaque mot d'une hypothèse comme étant ou non correct[3] --- en faisant en particulier abstraction de la validité ou non des mots qui l'entourent.

Une solution pratique (Blatz et al. 2003) consiste à aligner automatiquement l'hypothèse et la référence et à considérer comme corrects les mots qui sont alignés avec eux-mêmes ; un critère plus faible consiste à considérer comme correct tout mot qui apparaît dans la référence, indépendamment de sa position.

1.2.2 Prédire la confiance au niveau des phrases

Par comparaison, prédire la confiance au niveau des phrases semble un problème mieux posé et c'est le cadre que nous étudions plus particulièrement dans les sections suivantes. À ce niveau, il est en particulier possible de considérer de multiples variantes du problème, selon l'interprétation que l'on donne à y. L'approche la plus simple (Blatz et al. 2003) consiste à supposer que y ne prend que deux valeurs : « bon » et « mauvais », ces deux catégories étant, par exemple, déduites de mesures de qualité automatiques[4]. On retrouve un problème classique d'apprentissage automatique, l'apprentissage supervisé d'une classification binaire, qui peut être résolu par de nombreux outils standard (régression logistique, séparateurs à vaste marge, etc (Cornuéjols et Miclet 2002)). Un jeu d'étiquettes plus fin est considéré dans (Specia et al. 2009), qui ordonne les traductions depuis celles qui sont complètement inutilisables à celles qui sont complètement correctes, en distinguant deux niveaux de correction intermédiaires pour les traductions demandant beaucoup ou, au contraire, peu de révisions.

Pour s'affranchir du caractère potentiellement subjectif des étiquettes représentant la qualité d'une phrase, il est également possible d'envisager de prédire des quantités moins sujettes à caution : (Specia 2011) étudie des scénarios dans lesquels y est égal au score hTER ou encore au temps nécessaire pour corriger l'hypothèse de traduction ; voir également (Denkowski et Lavie 2012) pour une discussion sur la manière de noter *l'utilisabilité* d'une pré-traduction. On notera que dans cette configuration, le niveau de qualité n'est plus représenté par une variable discrète, mais par une grandeur continue, ce qui implique d'utiliser des méthodes statistiques de régression, plutôt que de classification.

Il existe naturellement bien d'autres manières d'envisager la tâche : une variante mineure considérera ainsi que les étiquettes de qualité définissent un ordre total et utilisent des techniques de régression ordinale. Une modification plus substantielle de la tâche consiste à s'affranchir complètement de la qualification de la qualité *absolue* d'une phrase et à ne considérer que la qualité *relative* des phrases d'un document. Vue sous cet angle, l'estimation de qualité vise à *ordonner* les phrases de la plus à la moins utile --- simulant un scénario dans lequel le « bugdet » de temps consacré à la post-édition est réduit et où les phrases les moins correctes doivent être traitées en priorité. Ce problème d'ordonnancement peut être alors abordé par des techniques d'apprentissage de fonction de reclassement (voir, par exemple, (Amini 2011)).

Dans la suite, nous présentons les résultats de deux études qui portent sur l'estimation de confiance au niveau des phrases et qui permettront de mieux comprendre le problème et d'en saisir les difficultés.

(3) (Bach, Huang et Al-Onaizan 2011) utilise un jeu d'étiquettes plus riche.
(4) Sont considérées comme « bonnes »les z% meilleures traductions au sens de la métrique, pour un choix approprié du seuil z et comme mauvaises toutes les autres.

2. Construction d'un système de CE

Nous allons, dans cette section, décrire un système de prédiction de la qualité pour la traduction automatique que nous avons développé dans le cadre de la campagne d'évaluation WMT. Cette campagne d'évaluation, organisée chaque année depuis 2006, a pour objectif principal d'évaluer les avancées récentes dans le développement des systèmes de traduction automatique (Callison-Burch et al. 2012). Elle propose depuis 2011 une tâche de prédiction de la qualité. Nous commencerons par présenter les données sur lesquelles la campagne est fondée, puis décrirons notre contribution et finirons par analyser les informations utiles pour discriminer les bonnes traductions des mauvaises.

2.1 La campagne WMT

Comme il est d'usage pour les campagnes d'évaluation, les organisateurs de WMT ont fourni un corpus de textes annotés pour entraîner les systèmes et tester leurs performances ; ils ont défini un protocole d'évaluation permettant de comparer les performances des différents systèmes.

Le corpus fourni est composé de 2 254 phrases en espagnol issues de différents journaux européens accompagnées d'une traduction automatique depuis l'anglais, produite par un système statistique état de l'art (Moses (Koehn, Hoang et al. 2007)), ainsi que d'une estimation de la qualité de cette traduction. Cette estimation est une moyenne pondérée de trois notes données par des traducteurs professionnels spécialement formés à la tâche. Ces notes sont des entiers compris entre 1 et 5 : 5 indique une très bonne traduction qui ne nécessite que peu, voire aucune correction, tandis que 1 indique une traduction tellement mauvaise qu'il est plus efficace de ne pas la considérer. Par rapport à la typologie introduite dans la section précédente, il s'agit donc d'une tâche de prédiction de confiance au niveau des phrases.

Plusieurs mesures ont été proposées par les organisateurs pour évaluer les performances d'un système. Par souci de simplicité, nous ne considérons dans ce travail que la mesure dont l'interprétation est la plus simple, l'erreur absolue moyenne (MAE pour *mean absolute error*). Le MAE d'un système est évalué sur le corpus de test par :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \qquad (1)$$

où *n* est le nombre d'exemples, *yi* est la véritable note de la ie phrase du corpus et *y^i* la note prédite. Le MAE peut s'interpréter comme la différence moyenne entre la note prédite et la note réelle : plus le MAE est petit, meilleures sont les performances d'un système.

2.2 Un système de prédiction de la confiance

Nous avons utilisé comme modèle d'apprentissage un simple modèle de régression linéaire régularisée *(ridge regression)* qui détermine l'étiquette comme une moyenne pondérée des caractéristiques, les coefficients associés à chaque caractéristiques étant déterminés automatiquement à partir du corpus d'apprentissage. Le choix d'une méthode de régression est naturelle dans la mesure où les étiquettes à prédire sont des réels ; le choix d'un modèle linéaire s'explique par la facilité et la rapidité de l'apprentissage.

De nombreuses caractéristiques ont été proposées pour la tâche d'estimation de confiance, ainsi que pour des tâches proches comme la prédiction de la lisibilité (Kanungo et Orr 2009) ou le développement de système capable d'évaluer automatiquement des rédactions (Burstein et al. 1998). Par exemple, le rapport final de l'atelier organisé en 2004 sur l'estimation de confiance pour la traduction automatique (Blatz et al. 2004) liste 91 caractéristiques et les travaux récents dans ce domaine comme (Felice et Specia 2012 ; Rubino et al. 2012) a augmenté de manière significative le nombre de caractéristiques considérées. Ces caractéristiques sont soit monolingues (c'est-à-dire qu'elles considèrent la phrase source indépendamment de la phrase cible et vice versa), soit bilingues (c'est-à-dire qu'elles considèrent des *alignements* entre la phrase source et la phrase cible) ; elles peuvent porter sur les mots individuels ou, au contraire, sur des groupes de mots, voire la phrase entière. Certaines ne portent que sur des informations de surface (comme la longueur des phrases), alors que d'autres sont fondées sur des analyses linguistiques sophistiquées de la phrase telle que l'analyse morpho-syntaxique de la phrase ou son analyse en dépendances.

Les organisateurs de la campagne WMT ont fourni, en plus des corpus et des méthodes d'évaluation, un ensemble de 17 caractéristiques de base que nous appellerons *baseline* dans la suite de ce travail. Dans nos expériences, nous avons enrichi ces caractéristiques *baseline* de différentes manières et avons considéré dans nos expériences un total de 107 caractéristiques qui peuvent être regroupés en cinq catégories principales :

1. des mesures de la qualité de l'« association » entre la phrase source et la phrase cible telles que, par exemple, des caractéristiques dérivées des scores IBM 1 ;

2. des mesures de la grammaticalité et de la fluidité de la phrase cible telles que les caractéristiques fondées sur les scores de modèles de langue ;

3. des caractéristiques de surface extraites essentiellement de la phrase source telles que le nombre de mots, le nombre de mots inconnus ou le nombre de mots qui ne sont pas alignés ;

4. des caractéristiques syntaxiques simples comme le nombre de noms, de verbes, de mots outils, etc. ;

5. des caractéristiques extraites du décodeur, qui dérivent des informations utilisées de manière interne par le système pour sélectionner la traduction produite.

Dans nos expériences, plusieurs sous-ensembles de ces caractéristiques ont été considérés. Le premier sous-ensemble, appelé ALL considère la totalité de ces caractéristiques ; le second, appelé BASELINE, ne considère que les caractéristiques fournies par les organisateurs de la tâche ; enfin, le troisième, SYNTAX, regroupe uniquement des caractéristiques syntaxiques dérivées essentiellement d'une annotation morpho-syntaxique des phrases sources et cibles.

2.3 Résultats obtenus

Le Tableau 1 résume les résultats obtenus dans les différentes configurations que nous avons considérées. Nous avons reporté à la fois la performance obtenue sur un corpus de test regroupant 20 % des données de la campagne et sur un corpus d'apprentissage regroupant le reste des données. Dans les deux cas, ces erreurs ont été estimées par une méthode de *bootstraping* et les intervalles de confiance à 95 % sont reportés.

Table 1 – Performance de différents systèmes de prédiction de la qualité en traduction automatique

Caractéristiques	Apprentissage	Test
BASELINE	0,621 – 0,622	0,624 – 0,631
ALL	0,495 – 0,496	0,522 – 0,540
SYNTAX	0,549 – 0,550	0,652 – 0,660

Ces résultats confirment la validité de nos choix : notre système « complet » obtient de meilleures performances que le système BASELINE. À notre connaissance, ces performances sont les meilleures reportées à ce jour sur ce corpus de données. Il apparaît également que les caractéristiques syntaxiques n'ont que peu de valeur, dans la mesure où elles obtiennent des performances plus mauvaises que celles du système le plus simple. Cette conclusion doit toutefois être prise avec précaution : les caractéristiques syntaxiques apportent une information utile quand on les conjoint à d'autres ; il est possible, par ailleurs, que des caractéristiques syntaxiques plus « sophistiquées » obtiennent de meilleures performances.

De manière générale, on peut dire que les performances obtenues (avec un MAE de plus de 0,5 dans le meilleur des cas) sont plutôt mauvaises et il n'est pas assuré qu'elles permettent une utilisation du système d'estimation de la confiance dans un *workflow* de post-édition réel. Nous présenterons une analyse des causes possibles de ces contre-performances relatives dans la section 4.

2.4 Analyse des caractéristiques pertinentes

Nous discutons, dans cette section, de l'identification des caractéristiques qui sont les plus utiles pour déterminer la qualité d'une traduction. Dans ces expériences, nous avons considéré cinq groupes de caractéristiques qui semblent jouer un rôle important :

1. les caractéristiques dérivées du modèle IBM 1, qui est un modèle statistique simple capable de prédire à quel point deux groupes de mots sont traduction l'un de l'autre ; ces caractéristiques seront notées ibm1 dans la suite ;

2. des caractéristiques dérivées d'un modèle de langue « classique » qui permettent de mesurer la grammaticalité d'une phrase (lm) ;

3. des caractéristiques dérivées d'un modèle de langue neuronal, par le truchement de représentation dans un espace continu, qui permettent de mieux généraliser les régularités observées à de nouvelles données (soul) ;

4. des caractéristiques dérivées d'un modèle de langue estimé sur les étiquettes morpho-syntaxiques (poslm) ;

5. des caractéristiques fondées sur les comptes d'étiquettes morpho-syntaxiques dans la phrase source et la phrase cible (poscount).

Ces deux derniers groupes sont composés à partir des informations syntaxiques les plus simples que l'on peut facilement et précisément déterminer automatiquement.

En utilisant ces différents groupes de caractéristiques, deux séries d'expériences ont été menées. Dans une première série, chaque groupe de caractéristiques a été successivement supprimé de l'ensemble des caractéristiques considérées dans les expériences décrites dans la

section précédente ; dans la seconde, les groupes de caractéristiques ont été successivement ajoutés aux caractéristiques *baseline.*

Nos résultats sont résumés dans le Tableau 2. Ces résultats montrent clairement que les informations les plus utiles pour prédire la qualité d'une traduction sont celles fournies par les modèles de langage neuronaux et, dans une moindre mesure, le modèle IBM 1 : ajouter ces caractéristiques aux caractéristiques *baseline* ou les enlever de l'ensemble des caractéristiques que nous avons considérées entraîne les plus fortes modifications de score.

Ces expériences confirment également les résultats obtenus par de nombreux participants à la campagne d'évaluation : les caractéristiques les plus pertinentes sont issues de modèles statistiques très simples et toutes les tentatives d'utiliser des caractéristiques plus linguistiques se sont soldées par des échecs. Une raison, assez naturelle, pourrait être que, les phrases traduites sont souvent non grammaticales, ce qui rend difficile toute analyse linguistique un peu poussée.

Table 2 – Résultat de l'analyse de l'importance des différentes caractéristiques

	Train	Test	Variation
Baseline	0,589 – 0,591	0,589 – 0,591	
+ibm1	0,579 – 0,582	0,591 – 0,599	(=)
+poscount	0,580 – 0,583	0,596 – 0,604	(=)
+poslm	0,587 – 0,589	0,599 – 0,607	(=)
+soul	0,540 – 0,544	0,563 – 0,571	(++)
+lm	0,585 – 0,587	0,596 – 0,604	(=)
+poslm,poscount	0,579 – 0,583	0,597 – 0,605	(=)
All	0,526 – 0,528	0,536 – 0,543	
-ibm1	0,537 – 0,543	0,554 – 0,561	(–)
-poscount	0,533 – 0,535	0,540 – 0,548	(=)
-poslm	0,531 – 0,533	0,540 – 0,547	(=)
-soul	0,582 – 0,585	0,595 – 0,603	(– –)
-lm	0,527 – 0,530	0,536 – 0,544	(=)
-poslm,poscount	0,537 – 0,539	0,544 – 0,552	(–)

3. Difficultés de la tâche

Nous allons, dans cette section, décrire plusieurs observations que nous avons effectuées pour mieux comprendre les résultats obtenus par notre système de prédiction de la qualité. Nous commencerons par critiquer la manière dont la tâche de prédiction de qualité a été définie dans le cadre de la campagne WMT, avant de parler du problème plus général de l'accord inter-annotateur lors de post-édition de traductions.

3.1 Limites de la tâche WMT

Comme l'on signalé les organisateurs, malgré l'attention toute particulière qui a été portée pour garantir la qualité des données, l'accord inter-annotateur est beaucoup plus bas que celui

qui est généralement observé dans les tâches de Traitement Automatique des Langues : le κ de Cohen pondéré varie entre 0,380 et 0,483 suivant la paire d'annotateurs considérée ; le coefficient de Fleiss est de 0,257. Cela peut s'expliquer par le fait que, comme l'illustre la Figure 2, les différents annotateurs humains ont utilisé l'échelle des scores de manière remarquablement différente. En particulier, le second annotateur a une tendance claire à donner des scores plus « moyens » que les deux autres annotateurs et la variance de son score est plutôt petite. Au contraire, les deux autres annotateurs ont plutôt tendance à privilégier les notes extrêmes décrivant les traductions comme étant soit bonnes soit mauvaises, mais rarement comme « moyennes ». C'est pourquoi, la variance de leurs scores est plus grande. Des résultats similaires ont été observés lors de l'étude de l'évaluation des sorties d'un système de traduction automatique au niveau des phrases (Koehn et Monz 2006).

Pour limiter l'impact de ce faible accord inter-annotateurs, les organisateurs de la campagne WMT ont décidé de définir la qualité d'une traduction (c'est-à-dire le score à prédire) comme une moyenne *pondérée* des trois scores humains en « donnant un poids plus important aux juges ayant la plus grande déviation standard » (Callison-Burch et al. 2012). Ce choix a été fait de manière à ce que les scores soient répartis de manière plus uniforme dans l'intervalle [1*,* 5]. Cependant, cette décision implique également que la fiabilité d'un annotateur est directement liée à la variance de son score.

Ce faible accord inter-annotateur et la décision de définir les étiquettes à prédire comme des moyennes pondérées des trois notes nous obligent à relativiser les (faibles) performances obtenues par notre système. La tâche traitée lors de la campagne d'évaluation WMT est, en effet, plus compliquée que la prédiction de la qualité de la traduction, puisqu'il faut également modéliser et prédire le désaccord des annotateurs. Ces observations montrent également qu'il est difficile de définir *a priori* (c.-à-d. sans effectuer les corrections) un jugement sur la difficulté de la post-édition et qu'il est peut-être nécessaire de choisir autrement les étiquettes.

3.2 Difficulté de définir la post-édition minimale

Nous avons montré, dans la section précédente, que l'estimation, a priori, de la difficulté de post-édition était difficile et ne permettait pas d'obtenir des étiquettes fiables. Ceci peut être dû soit à l'impossibilité d'évaluer la difficulté des corrections minimales sans réaliser celles-ci, soit au fait que la notion même de « correction minimale » varie grandement d'un correcteur à l'autre. Dans cette section, nous souhaitons déterminer parmi ces hypothèses laquelle est la plus réaliste.

Dans le cadre du projet TRACE[5], un grand corpus de post-éditions a été collecté : celui-ci contient 6 923 traduction automatiques de l'anglais vers le français et 6 692 traduction automatique automatiques du français vers l'anglais, toutes ces traductions étant post-éditées par des traducteurs professionnels. Afin de s'approcher le plus possible de conditions réalistes et comme cela est généralement demandé dans les campagnes d'évaluation de traduction automatique, il a été demandé aux correcteurs d'effectuer la correction *minimale* (au sens du nombre de mots modifiés) permettant d'obtenir une traduction correcte, aussi bien par rapport à la conservation du sens qu'à la fluidité et la grammaticalité de la traduction.

Pour chaque direction de traduction 1 000 traductions ont été, indépendamment, corrigées deux fois. À notre connaissance, c'est la première fois que deux annotateurs différents corrigent les mêmes phrases, permettant une comparaison des post-éditions. Pour des raisons de place, nous décrirons, dans le reste de cette section, uniquement les résultats obtenus en sur le corpus de traductions de l'anglais vers le français. Les résultats pour la direction français vers anglais sont similaires.

(5) http://anrtrace.limsi.fr, TRACE un projet du programme Contenu et Interaction de l'Agence Nationale pour la Recherche.

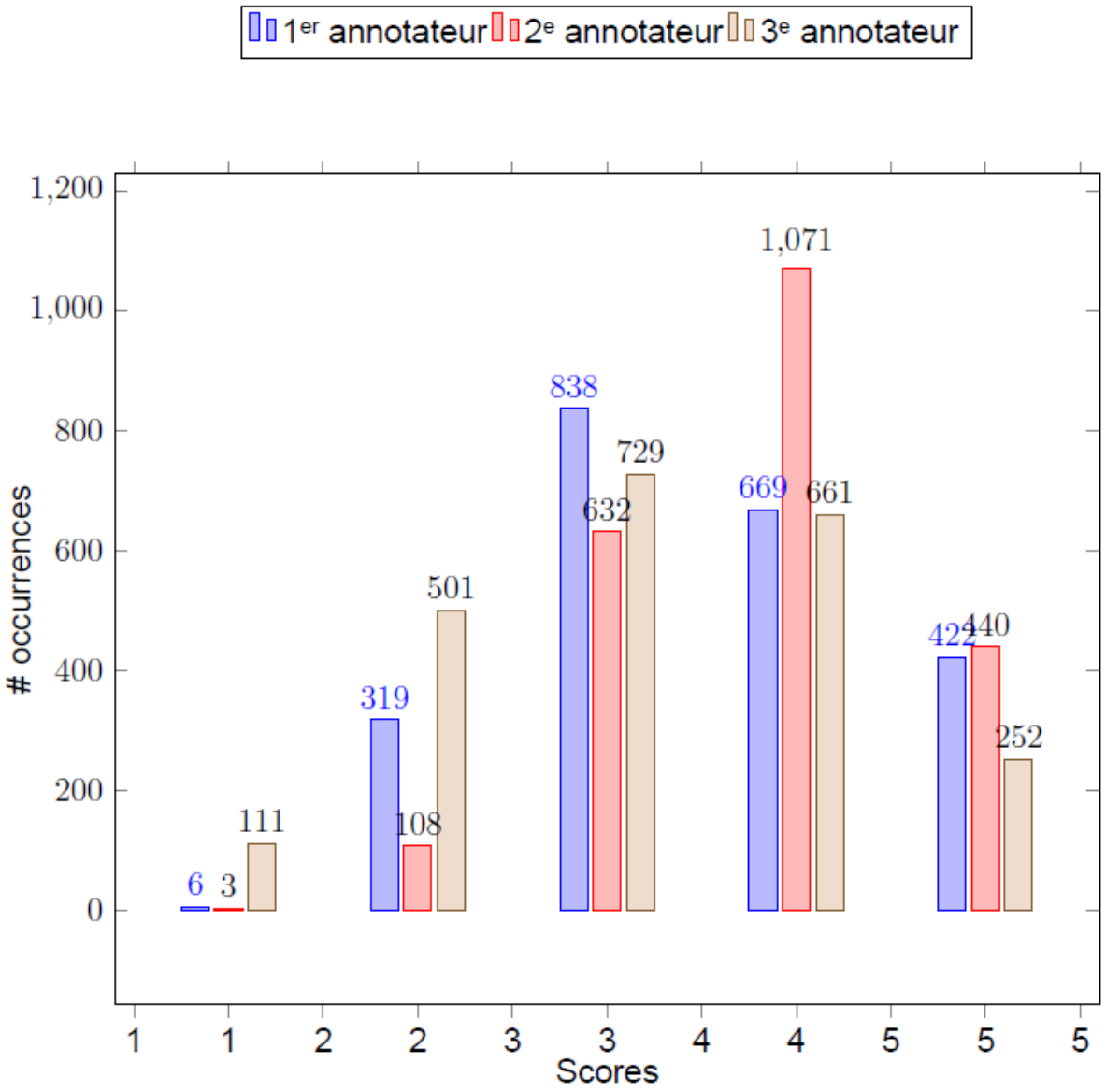

Figure 2 – Distribution des scores pour les trois annotateurs

De manière quantitative, il est possible de quantifier la similarité entre les post-éditions effectuées par les différents correcteurs en mesurant la corrélation entre les scores hTER obtenus en prenant ces corrections comme référence. Cette corrélation est faible : le coefficient de Pearson entre les deux notes n'est que de 0,642 et le τ de Kendall de 0,476 indiquant, intuitivement, que si les traductions étaient ordonnées suivant leur score hTER, deux traductions quelconques ne seraient dans le même ordre, qu'une fois sur deux. De manière globale, les post-éditions produites ne sont identiques que dans 12 % des cas[6]. La distance d'édition moyenne entre les deux post-éditions est de 24 %, ce qui revient à dire qu'il faut, pour passer d'une post-édition à l'autre changer en moyenne un mot sur cinq. Les opérations les plus fréquemment mises en jeu dans cette transformation sont les substitutions de mots (57 % des modifications) suivi des suppressions et des insertions de mots (16 % dans les deux cas) ; les déplacements de mots n'interviennent que dans 11 % des cas.

(6) La comparaison entre les deux corrections ne tient compte ni de la ponctuation, ni de la casse.

Table 3 - Exemple de différences de post-éditions. Les différences les plus marquantes sont indiquées en gras.

1	**source**	'False confession'
	trad. autom.	'Confession fausse'
	correction no 1	**"Confession fausse"**
	correction no 2	**« Faux témoignage »**
2	**source**	Each year, the Member States shall send the Commission a report on the evaluation of the execution and effectiveness of this regulation.
	trad. autom.	Chaque année, les États membres transmettent à la Commission un rapport sur l'évaluation de l'exécution et l'efficacité de cette réglementation.
	correction no 1	Chaque année, les États membres **transmettent** à la Commission un rapport **sur l'évaluation de** l'exécution et l'efficacité **de cette réglementation.**
	correction no 2	Chaque année, les États membres **communiquent** à la Commission un rapport **d'évaluation concernant** l'exécution et l'efficacité **du présent règlement.**
3	**source**	I'm thinking this must be an ancient print date, right.
	trad. autom.	Je retiens ce doit être une date imprimée antique.
	correction no 1	Je pense qu'il s'agit une **ancienne édition, c'est évident.**
	correction no 2	Je pense que ça doit être une **ancienne date d'impression, n'est-ce pas.**
4	**source**	So let's take a tour of this state-of-the-art clean coal facility.
	trad. autom.	Donc prenons un tour de cet état de l'art nettoient la facilité de charbon.
	correction no 1	Alors allons voir **ces installations ultramodernes** de charbon propre.
	correction no 2	Donc faisons une visite de **cette installation** de charbon propre **à la pointe de la technologie.**
5	**source**	Dear Valued Customer, For your problem, please follow the steps below to have a troubleshooting.
	trad. autom.	Cher valorisées à la clientèle, pour votre problème, veuillez suivre les étapes ci-dessous pour avoir un dépannage.
	correction no 1	Cher client estimé, pour votre problème, veuillez suivre les étapes ci-dessous **pour avoir un dépannage**.
	correction no 2	Très cher client, pour votre problème, veuillez suivre les étapes ci-dessous **pour être dépanné**.

Plus qualitativement, la Table 3 reprend des exemples des corrections les plus différentes, ainsi que des phrases sources et des traductions automatiques. Ces exemples illustrent différents types de différences entre les post-éditions :

- sensibilité différente aux traductions littérales : dans de nombreux cas, un correcteur accepte une traduction parfaitement compréhensible et juste d'un point de vue grammatical, même si elle n'aurait jamais été « produite » par un locuteur natif, alors que le second préfère la reformuler (1er et 5e exemples) ;
- reformulation non nécessaire de la traduction automatique (le second correcteur qui corrige « cette réglementation » en « le présent règlement » dans le 2e exemple) ;
- utilisation de paraphrases ou de synonymes sans raisons apparentes (« ultramodernes » *versus* « à la pointe de la technologie » dans le 4e exemple) ;
- ambiguïté liée au manque de contexte en source (« cette installation » *versus* « ces installations » dans le 4e exemple).

Il est intéressant de remarquer que les corrections sont différentes aussi bien quand la traduction automatique est *plutôt* bonne (2e exemple) que quand elle est complètement fausse (3e et 4e exemple).

Toutes ces observations montrent la difficulté inhérente à la tâche de prédiction de qualité en traduction : dans la mesure où la post-édition semble aussi subjective que la traduction elle-même, toutes les étiquettes que l'on pourrait chercher à prédire seront fortement bruitées et leur prédiction compliquée.

Conclusion

Dans ce travail, nous avons décrit la tâche de prédiction de qualité pour la traduction automatique et nous avons évoqué différentes manières dont ces mesures de confiance pourraient contribuer à améliorer l'utilisation de traductions automatiques, et en particulier à améliorer la productivité de la post-édition. Les expériences que nous avons résumées montrent, cependant, que les résultats des systèmes actuels sont loin d'être satisfaisants pour leur utilisation dans des applications réelles, et que l'amélioration de ces systèmes nécessitera une meilleure définition de la tâche et de la manière dont les annotations sont définies et les prédictions sont utilisées.

Remerciements

Ce travail a été partiellement financé par l'ANR-CONTINT au travers du projet TRACE.

Bibliographie

Allauzen, Alexandre et François Yvon (2011). « Méthodes statistiques pour la traduction automatique ». Dans : *Modèles statistiques pour l'accès à l'information textuelle*. Sous la dir. d'Eric Gaussier et François Yvon. Hermès, Paris. Chap. 7, p. 271–356.

Amini, Massih-Réza et al (2011). « Modèles d'ordonnancement pour le résumé automatique et la recherche d'information ». Dans : *Modèles statistiques pour l'accès à l'information textuelle*. Sous la dir. d'Eric Gaussier et François Yvon. Hermès, Paris. Chap. 2, p. 65–93.

Bach, Nguyen, Fei Huang et Yaser Al-Onaizan (juin 2011). « Goodness: A Method for Measuring Machine Translation Confidence ». Dans: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, p. 211–219.

Banerjee, Satanjeev et Alon Lavie (juin 2005). « METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments ». Dans: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, p. 65–72. Url: http://www.aclweb.org/anthology/W/W05/W05-0909.

Blatz, John et al. (2003). *Confidence Estimation for Machine Translation*. Final report, John Hopkins University / CLSP Summer Workshop. — (2004). « Confidence Estimation for Machine Translation ». Dans : *Proceedings of Coling 2004*. Geneva, Switzerland, p. 315–321.

Burstein, Jill et al. (août 1998). « Automated Scoring Using A Hybrid Feature Identification Technique ». Dans: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. Montreal, Quebec, Canada: Association for Computational Linguistics, p. 206–210. doi: 10.3115/980845.980879.

Callison-Burch, Chris et al. (juin 2012). « Findings of the 2012 Workshop on Statistical Machine Translation ». Dans: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada : Association for Computational Linguistics, p. 10–51.

Cornuéjols, Antoine et Laurent Miclet (2002). *Apprentissage artificiel, concepts et algorithmes*. Eyrolle, Paris.

Denkowski, Michael et Alon Lavie (2012). « Challenges in predicting machine translation utility for human post-editors ». Dans: *Proc. AMTA-2012: the Tenth Biennial Conference of the Association for Machine Translation in the Americas*. San Diego, CA.

Felice, Mariano et Lucia Specia (juin 2012). « Linguistic Features for Quality Estimation ». Dans: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, p. 96–103.

Garcia, Ignacio (2011). « Translating by post-editing: is it the way forward? » Dans: *Machine Translation* 25, p. 217–237.

Gispert, Adrià de et al. (2012). « N-gram posterior probability confidence measures for statistical machine translation: an empirical study ». Dans: *Machine Translation*, p. 1–30.

González Rubio, Jesús, Daniel Ortiz Martínez et Francisco Casacuberta (juil. 2010). « Balancing User Effort and Translation Error in Interactive Machine Translation via Confidence Measures ». Dans: *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, p. 173–177. url: http://www.aclweb.org/anthology/P10-2032.

He, Yifan et al. (juil. 2010). « Bridging SMT and TM with Translation Recommendation ». Dans: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, p. 622–630. url: http://www.aclweb.org/anthology/P10-1064.

Hovy, Eduard, Maghi King et Andrei Popescu-Belis (2002). « An introduction to MT evaluation ». Dans: LREC-2002: Third International Conference on Language Resources and Evaluation. Workshop: Machine translation evaluation: human evaluators meet automated metrics. Las Palmas Canary Islands, pp. 1–7.

Kanungo, Tapas et David Orr (2009). « Predicting the readability of short web summaries ». Dans: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. Barcelona, Spain: ACM, p. 202–211.

Koehn, Philipp (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, Philipp, Hieu Hoang et al. (juin 2007). « Moses: Open Source Toolkit for Statistical Machine Translation ». Dans: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, p. 177–180.

Koehn, Philipp et Christof Monz (juin 2006). « Manual and Automatic Evaluation of Machine Translation between European Languages ». Dans: *Proceedings on the Workshop on Statistical Machine Translation*. New York City: Association for Computational Linguistics, p. 102–121.

Lavie, Alon et Mark Przybocki (2009). « Introduction to the special issue on "Automated Metrics for Machine Translation Evaluation ». Dans: *Machine Translation* 23 (2). 10.1007/s10590- 010-9071-8, p. 69–70. issn: 0922-6567. url: http://dx.doi.org/10.1007/s10590-010-9071-8.

Lopez, Adam (août 2008). « Statistical machine translation ». Dans : ACM Comput. Surv. 40.3, 8 :1–8 :49. issn : 0360-0300. doi : 10.1145/1380584.1380586. url : http://doi.acm.org/10.1145/1380584.1380586.

Mirko Plitt, François Masselot (2010). « A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Contex ». Dans: *The Prague Bulletin of Mathematical Linguistics* 93, p. 7–16.

Papineni, Kishore et al. (2002). « BLEU : a method for automatic evaluation of machine translation ». Dans : *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, p. 311–318. doi: 10.3115/1073083.1073135. url: http://dx.doi.org/10.3115/1073083.1073135.

Raybaud, Sylvain, David Langlois et Kamel Smaïli (2011). « "This sentence is wrong." Detecting errors in machine-translated sentences ». Dans : *Machine Translation* 25.1, p. 1–34.

Roturier, Johan (2009). « Deploying novel MT technology to raise the bar for quality : a review of key advantages and challenges ». Dans : Ottawa, Canada : International Association for Machine Translation.

Rubino, Raphael et al. (juin 2012). « DCU-Symantec Submission for the WMT 2012 Quality Estimation Task ». Dans: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, p. 138–144.

Snover, Matthew et al. (août 2006). « A Study of Translation Edit Rate with Targeted Human Annotation ». Dans: *Proceedings AMTA*, p. 223–231. url: http://www.mt-archive.info/A MTA-2006-Snover.pdf.

Soricut, Radu, Nguyen Bach et Ziyuan Wang (juin 2012). « The SDL Language Weaver Systems in the WMT12 Quality Estimation Shared Task ». Dans : *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada : Association for Computational Linguistics, p. 145–151.

Soricut, Radu et Abdessamad Echihabi (juil. 2010). « TrustRank : Inducing Trust in Automatic Translations via Ranking ». Dans : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden : Association for Computational Linguistics, p. 612–621.

Specia, Lucia (2011). « Exploiting objective annotations for measuring translation post-editing effort ». Dans : *Proceedings of the 15th conference of EAMT*. Leuven, Belgium, p. 73–80.

Specia, Lucia et al. (2009). « Improving the confidence of machine translation quality estimates ». Dans : *MT Summit XII : proceedings of the twelfth Machine Translation Summi*. Ottawa, Ontario, Canada, p. 136–143.

Ueffing, Nicola et Hermann Ney (oct. 2005). « Word-Level Confidence Estimation for Machine Translation using Phrase-Based Translation Models ». Dans: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language*

Processing. Vancouver, British Columbia, Canada, p. 763–770. url : http://www.aclweb.org/anthology/H/H05/H05-1096.

Wagner, Robert A. et Michael J. Fischer (1974). « The String-to-String Correction Problem ». Dans : *Journal of the ACM (JACM)* 21.1, p. 168–173. issn : 0004-5411. doi : http://doi.acm.org/10.1145/321796.321811.