Chapitre 13

L'évaluation des systèmes de traduction automatique

13.1. Spécificités de la traduction automatique et de son évaluation

La traduction automatique est l'un des objectifs historiques de l'intelligence artificielle et du traitement automatique des langues (TAL). Le problème a l'avantage de s'énoncer très simplement : produire un texte dans une langue cible qui soit la traduction (au sens courant, humain) d'un texte source donné. La portée applicative de la tâche est également facile à comprendre.

Malheureusement, en l'absence d'une définition opérationnelle ou algorithmique de la traduction, la conception des systèmes de traduction automatique (TA) reste une tâche ardue. Devant les imperfections des traductions produites actuellement par ces systèmes, l'importance d'une évaluation quantitative de leur qualité apparaît clairement. Comme pour d'autres problèmes du TAL, l'existence de mesures de qualité communément admises par les experts est un facteur essentiel de progrès. Or, comme nous allons le voir dans ce chapitre, l'évaluation d'un système de TA, et notamment l'évaluation de la qualité d'une traduction, reste un problème difficile, sujet à de nombreux débats.

Précisons tout d'abord l'objet du chapitre : ce sont les systèmes visant une traduction *totalement automatique* du texte source, c'est-à-dire n'impliquant pas de révision humaine. Outre ces systèmes, il existe de nombreux outils d'*aide à la*

Chapitre rédigé par Anthony HARTLEY et Andrei POPESCU-BELIS.

traduction, dont l'évaluation implique la participation de sujets humains. A ce titre, leur évaluation s'appuie sur des mesures de l'utilisabilité et sur des techniques d'évaluation des interfaces humain-machine qui ne relèvent pas de ce chapitre.

Les dictionnaires électroniques, intégrés ou non à des éditeurs de textes multilingues, constituent déjà un exemple d'outil d'aide à la traduction, mais ce sont surtout les mémoires de traduction qui ont connu récemment un succès notable. Ces logiciels permettent de dériver des paires de phrases, dont l'une est la traduction de l'autre, à partir de textes déjà traduits, grâce à l'alignement par l'humain de la source et de la cible. Afin d'évaluer un tel outil sans faire appel à un utilisateur humain, on peut par exemple s'intéresser à sa capacité à apparier des phrases d'un texte à traduire avec des phrases déjà stockées dans la mémoire de traduction.

Toutefois, l'évaluation véritablement informative d'un tel outil s'intéressera plutôt à l'accroissement de la productivité d'un sujet – traducteur humain – qui l'utilise.

Dans ce chapitre, nous nous intéresserons aux différentes façons de mesurer la qualité d'un système produisant une traduction qui se veut achevée, et qui doit être utilisée telle quelle par différents utilisateurs humains, ou par d'autres systèmes de TAL (par exemple en recherche documentaire multilingue).

Dans la section 13.2, nous dresserons un tableau d'ensemble des différentes méthodes d'évaluation proposées, en choisissant une approche fondée sur des principes normalisés, et qui tient compte du contexte d'utilisation d'un système de TA.

La section 13.3 se concentrera sur la « qualité intrinsèque » des textes produits par un système, en brossant un tableau critique des mesures les plus répandues – tant celles faisant appel à des juges humains, que celles, plus récentes, fondées sur des algorithmes automatisables.

Dans la section 13.4, nous passerons en revue quelques campagnes marquantes d'évaluation de la TA, durant la dernière décennie, avec leurs points forts et leurs faiblesses.

Enfin, la section 13.5 exposera quelques critères formels d'analyse des mesures d'évaluation, avec des exemples d'application dans la communauté de la TA, avant une esquisse des perspectives du domaine, dans la section 13.6.

13.2. Le cadre théorique de l'évaluation de la TA

A défaut d'une mesure qui résolve à elle seule toutes les difficultés de l'évaluation de la TA, on rencontre un ensemble de mesures possibles, structurées selon leur contexte d'emploi. Un cadre théorique a été récemment synthétisé, qui s'inspire des normes de l'ISO (organisation internationale de normalisation), et permet, comme nous allons le voir, de cerner les difficultés du domaine.

13.2.1. *Exploitation des normes ISO/IEC pour l'évaluation des logiciels*

Deux séries de normes de l'ISO concernent l'évaluation des logiciels en général, et s'appliquent donc aussi au TAL et aux logiciels de TA. La série ISO/IEC 9126-1 à 4 s'intéresse aux qualités générales des logiciels, alors que la série ISO/IEC 14598-1 à 6 décrit le processus d'évaluation. Selon ces normes, évaluer un système, c'est mesurer sa *qualité*, où la qualité est l'ensemble des caractéristiques du système qui permettent de répondre aux besoins de ses utilisateurs [ISO 01].

Les caractéristiques de qualité sont regroupées en six catégories générales : fonctionnalité,

- fiabilité ;
- utilisabilité ;
- efficacité ;
- possibilité de maintenance ;
- et portabilité.

Catégories qui elles-mêmes se subdivisent en sous-catégories. Dans cette hiérarchie, les éléments dont on peut mesurer concrètement la qualité sont les subdivisions terminales, appelées *attributs*.

Par conséquent, l'évaluation d'un logiciel, donc la mesure de sa qualité, se ramènent à la mesure d'un ou de plusieurs attributs. Pour chaque attribut, on utilise une *métrique* qui lui assigne un niveau de qualité sur une échelle associée à la métrique. Notons que le terme « métrique » ne présuppose pas, dans ce cadre, toutes les propriétés mathématiques d'une métrique (souvent, les « métriques » utilisées ne satisfont pas l'inégalité triangulaire).

Selon l'ISO, on peut distinguer la qualité interne, la qualité externe et la qualité à l'usage. La première peut être mesurée grâce à des attributs internes du système – sans exécution – lors des évaluations dites « en boîte de verre », puisque les caractéristiques intrinsèques du système sont alors transparentes aux évaluateurs.

Pour la TA, des exemples d'attributs de qualité internes sont la taille du dictionnaire utilisé par un système, le nombre de règles de transfert, ou la taille du modèle de langage.

En principe, la qualité interne influence la qualité externe, qui est mesurée grâce à des attributs externes, en faisant fonctionner le système – lors d'évaluations dites « en boîte noire », car on s'intéresse seulement aux résultats produits par le système. Pour la TA, les différents aspects de la qualité du texte produit (voir la section 13.3) constituent des attributs de qualité externes, comme par exemple la durée nécessaire pour traduire un texte.

Enfin, la qualité à l'usage doit être mesurée en plaçant le système dans son contexte d'utilisation, et en quantifiant son efficacité au sein du processus de production. Pour la TA, cela présuppose l'analyse des performances des utilisateurs du système, et cela s'effectue au cas par cas, selon leurs tâches et leurs profils. Dans le cas particulier où le système de TA est encapsulé dans un autre système, qui joue le rôle d'utilisateur exclusif de ses résultats (par exemple un système de recherche d'information), l'évaluation à l'usage doit se faire en mesurant les performances du système encapsulant, avec des mesures adaptées à sa propre tâche.

13.2.2. *Evaluation contextuelle de la TA : le cadre FEMTI*

L'évaluation d'un système par rapport à une tâche donnée revient à fixer les attributs de qualité pertinents pour la tâche, ainsi que les métriques qui permettent de les quantifier. Si les six catégories définies par l'ISO ont, en principe, chacune leur importance, les chercheurs s'intéressent souvent à la fonctionnalité de leurs systèmes.

Par exemple, pour des systèmes de dictée vocale, on pourra mesurer la proximité de la transcription produite par le système par rapport à la transcription correcte établie par les juges humains. Pour la TA, la fonctionnalité se subdivise en plusieurs sous-catégories et attributs, qui reflètent des aspects de la qualité du texte produit. D'autres attributs sont également importants, tels ceux liés au comportement temporel (vitesse du système), à l'utilisation des ressources, et à la facilité de mise à jour. Dans l'état actuel des systèmes de TA, la fonctionnalité peut même parfois être moins importante, pour l'évaluation, que d'autres caractéristiques de qualité, selon la tâche prévue pour le système [CHU 93].

On voit donc apparaître la nécessité de structurer les différents attributs de qualité selon les besoins de l'évaluation, c'est-à-dire le contexte d'utilisation prévu pour le système de TA, les caractéristiques des utilisateurs, etc. Dans les normes de

l'ISO, l'influence du contexte sur l'évaluation n'apparaît qu'à travers quelques exemples (voir [HOV 03], 3.1 pour une analyse). Le projet EAGLES visait, lui, à appliquer le cadre ISO au TAL [EAG 96].

Or, pour l'évaluation de la TA, l'influence du contexte est centrale, comme le montrent Hovy *et al.* [HOV 99, HOV 03], ainsi que le schéma d'évaluation préconisé par JEIDA [NOM 92a, NOM 92b]. Des synthèses portant sur les attributs de qualité et les métriques ont été également proposées respectivement par Van Slype pour la TA [VAN 79], et par Sparck Jones et Galliers pour le TAL [SPA 96].

Une synthèse récemment réalisée pour la TA met en avant de façon systématique le rôle du contexte d'utilisation dans la sélection des attributs de qualité [HOV 03]. En effet, le cadre FEMTI pour l'évaluation de la TA (*framework for the evaluation of mt in isle*[1]) offre d'abord aux évaluateurs la possibilité de définir les exigences de l'évaluation, en termes de caractéristiques de la tâche de traduction assignée au système à évaluer, de caractéristiques des utilisateurs prévus et des textes à traduire. Organisées de façon hiérarchique, et complétées par des considérations sur les buts et l'objet de l'évaluation, ces instructions constituent la première partie de FEMTI.

La seconde partie de FEMTI développe la hiérarchie des caractéristiques de qualité jusqu'aux attributs et à leurs métriques. Cette hiérarchie est ancrée à la racine dans les six classes définies par l'ISO, puis particularisée pour les systèmes de TA, comme il ressort du tableau ci-après.

La synthèse de FEMTI a été rendue possible par la participation de plusieurs experts du domaine, dans le cadre d'une série d'ateliers qui ont proposé des exercices pratiques d'évaluation de la TA, des communications orales, et des discussions d'experts.

Ces ateliers sont décrits sur le site Internet de FEMTI, et dans [HOV 03] ; un exemple sera développé dans la section 13.5. Dans son état actuel, FEMTI regroupe la plupart des attributs et métriques utilisés par la communauté, avec de nombreuses références aux travaux qui les définissent et/ou les emploient. Plusieurs améliorations sont à apporter à l'avenir, notamment l'automatisation des liens entre la première et la seconde partie, qui devra permettre à terme la spécification automatique d'une évaluation de TA en fonction du contexte d'utilisation souhaité pour les systèmes ; et aussi, l'analyse de chaque métrique en termes de cohérence statistique, de corrélation avec d'autres métriques, et de coût.

---

1. Un des résultats du projet ISLE, consultable à l'adresse : http://www.issco.unige.ch /projects/isle/femti ou bien à http://www.isi.edu/natural-language/mteval.

Exigences de l'évaluation (1)	Caractéristiques et attributs de qualité (2)
1. But de l'évaluation 2. Objet de l'évaluation 3. Caractéristiques de la tâche 3.1. Assimilation 3.2. Dissémination 3.3. Communication 4. Caractéristiques de l'utilisateur 4.1. Utilisateur de la TA brute 4.2. Utilisateur de la TA achevée 4.3. Organisation utilisatrice 5. Caractéristiques du texte à traduire 5.1. Type de document 5.2. Auteur 5.3. Sources d'erreur	1. Caractéristiques internes des systèmes de TA 1.1. Type de l'algorithme de traduction 1.2. Ressources linguistiques : langues, dictionnaires, glossaires, corpus alignés, grammaires 1.3. Caractéristiques du processus : préparation du texte, postédition, interaction avec le système 2. Caractéristiques externes du système 2.1. Fonctionnalité 2.1.1. Adéquation : lisibilité du texte produit, intelligibilité ; cohérence, cohésion, style 2.1.2 Précision : fidélité au texte source, consistance, correction terminologique 2.1.3. Bonne formation : ponctuation, items lexicaux, morphologie, syntaxe 2.1.4. Interopérabilité 2.1.5. Conformité 2.1.6. Sécurité 2.2. Fiabilité 2.3. Utilisabilité (ergonomie) 2.4. Efficacité 2.4.1. Efficacité temporelle : temps de prétraitement, vitesse de traduction brute, temps de post-traitement 2.4.2. Utilisation des ressources : mémoire, lexique, nettoyage, taille du logiciel 2.5. Possibilités de maintenance 2.5.1. Analysabilité 2.5.2. Stabilité 2.5.3. Testabilité 2.5.4. Possibilités de changement : dictionnaires, grammaires, ajout d'une langue 2.6. Portabilité 2.7. Coût

Tableau 13.1. *Vue simplifiée du cadre FEMTI pour l'évaluation de la TA*

Le tableau 13.1 fournit un aperçu des principales caractéristiques de qualité des systèmes de TA, et des paramètres définissant les contextes d'utilisation. Les attributs internes sont naturellement spécifiques à la TA, alors que les attributs externes sont des raffinements des six caractéristiques ISO de base.

Les attributs de qualité les plus typiques en TA sont ceux ayant trait à la qualité du texte produit, regroupés sous la fonctionnalité (2.1). Ce sont en effet ceux qui permettent de répondre à la question : « Est-ce que la traduction produite est convenable ou non ? ».

Dans la mesure où « convenable » doit s'entendre par rapport à une certaine utilisation, plusieurs attributs caractérisent cette qualité, divisés en deux sous-catégories :

– les attributs ayant trait à la qualité du texte produit en lui-même, en tant que texte dans la langue cible (bonne formation grammaticale, lisibilité, etc.) ;

– et les attributs ayant trait à la proximité (sémantique, stylistique, etc.) du texte produit et du texte source.

Outre ces attributs, pour lesquels des métriques et des campagnes d'évaluation seront décrites dans les sections 13.3 et 13.4, d'autres attributs sont également pertinents [CHU 93].

Citons ainsi la vitesse de traduction, qui peut être primordiale dans des applications de recherche d'information multilingue, où l'on doit traduire rapidement de grandes quantités de textes, avec une certaine tolérance sur la qualité. Les possibilités de mise à jour, notamment pour les ressources lexicales, sont une autre caractéristique importante, par exemple pour des systèmes qui doivent traduire une terminologie spécifique à un domaine [SEN 03].

On constate donc que, contrairement à d'autres problèmes du TAL, la traduction automatique présente de multiples facettes à évaluer, chacune ayant son importance.

13.2.3. *Formalisation de l'évaluation par étapes*

Pour conclure cette section, il est important de résumer les principales étapes de l'évaluation d'un système de TA. Cette division s'inspire des normes ISO/IEC [ISO 00], interprétées par EAGLES en vue du TAL [EAG 96], et résumés dans [POP 99]. Les étapes principales d'une évaluation sont :

– la définition des qualités requises des systèmes (ici, on définit un contexte d'utilisation grâce à FEMTI, puis on sélectionne les caractéristiques de qualité pertinentes) ;

– la spécification de l'évaluation par le choix des métriques et du mode d'emploi (procédé d'application, données, etc.) ;

– l'exécution de l'évaluation ;

– la conclusion et le rapport final.

Pour ce qui est de l'application des métriques, on peut distinguer trois étapes [EAG 96, POP 99] :

– la mesure proprement dite de chaque attribut avec la métrique choisie ;

– l'appréciation de chaque valeur obtenue (chiffre ou classe) sur une échelle de scores établie en fonction des nécessités de l'évaluation ;

– et l'intégration des scores en un résultat final, si cela est souhaité, par exemple pour comparer des systèmes.

Dans cette optique, l'évaluation dépendante du contexte qui est préconisée par FEMTI se ramène à la sélection des métriques et la pondération des scores lors de l'intégration, dictées par l'application de TA envisagée.

Pour clore cette section, il faut évoquer d'autres synthèses portant sur l'évaluation de la traduction, cette fois-ci humaine.

En effet, on peut penser que l'appréciation des étudiants dans les écoles de traduction – et plus généralement dans les établissements qui pratiquent les exercices de traduction, version ou thème – présuppose une méthode d'évaluation systématique. On constate que l'expertise des correcteurs regroupe de façon intuitive plusieurs des attributs de qualité contenus dans FEMTI, notamment ceux qui constituent les sous-catégories adéquation, précision et bonne formation. Lors de la notation d'une traduction, ces attributs sont intégrés souvent inconsciemment par les correcteurs produisant une note finale. Des tentatives existent pour introduire plus de précision dans ces corrections, tels les critères de certification de l'Association américaine des traducteurs [ATA 02], ou les normes de qualité pour les documents traduits dans l'industrie automobile [SAE 01, WOY 02] – sur lesquels nous reviendrons plus bas. Il est heureux de constater que les attributs de qualité définis dans ces documents se retrouvent le plus souvent dans FEMTI, notamment en liaison avec la qualité du texte produit, vers laquelle nous nous tournons maintenant.

13.3. Métriques visant la qualité du texte produit

Dans la section précédente, nous avons dressé un tableau de l'ensemble des principales caractéristiques de qualité des systèmes de TA. Ici nous nous concentrons sur les caractéristiques contribuant à la fonctionnalité (partie 2.1) qui ont trait à la qualité du texte produit, c'est-à-dire l'adéquation, la précision et la bonne formation. Selon la terminologie proposée par John White [WHI 03], il s'agit de « l'évaluation déclarative » dont la portée intéresse plusieurs publics – l'utilisateur final (traducteur ou lecteur), le manager, le développeur, l'investisseur et le revendeur – et qui mérite de ce fait une attention particulière.

Notre objectif, quoique limité, n'en est pas simplifié pour autant, ceci pour plusieurs raisons. Pour évaluer un attribut, il faut normalement pouvoir le comparer à un idéal qui soit « correct » ou « le meilleur ». Or, dans le domaine de la traduction il est admis que cet idéal n'existe pas. Etant donné un grand nombre de traductions (humaines) d'un même texte source, il est probable qu'il n'y aura pas d'accord général sur le choix de la meilleure traduction et qu'aucune traduction ne sera jugée parfaite. Bref, « l'étalon-or » (*gold standard*) que l'on peut imaginer plus ou moins facilement pour la correction orthographique ou syntaxique nous fuit ; il n'y a pas une seule bonne réponse. En concevant des métriques, il nous faut donc nous accommoder de la variabilité légitime des traductions comme de la subjectivité des juges appelés à les évaluer.

L'appel aux juges humains entraînant non seulement la subjectivité mais aussi des dépenses considérables en argent et en temps, il n'est pas surprenant que des travaux récents cherchent à se passer d'intervention humaine. Nous allons donc considérer tour à tour l'approche humaine et l'approche automatisée.

Rappelons d'abord ce que nous entendons par *métrique* : il s'agit d'un test particulier qui vise à évaluer un *attribut* particulier du système de TA à l'aide d'une *technique* particulière. Les attributs sont des propriétés souhaitables du système ou des résultats qu'il produit, par exemple, dans le cas présent, *lisibilité* ou *fidélité*. Une technique va associer une méthode de collecte de réponses – par exemple, un questionnaire à choix multiples – avec une échelle dont l'interprétation fournit une mesure de qualité.

13.3.1. *Métriques nécessitant des juges humains*

On recense trois types d'approches qui requièrent la participation de juges humains :

– on peut inviter ceux-ci à accomplir une tâche à l'aide d'un document traduit ;

– ou bien à analyser les erreurs dans la traduction ;

– ou encore à prononcer un jugement intuitif sur la qualité de celle-ci.

Mesurer la capacité d'un sujet à accomplir une tâche à l'aide d'un document traduit (la qualité à l'usage) est une approche qui remonte aux expériences réalisées en 1971 par H. Wallace Sinaiko (rapportées par [VAN 79] et par [FAL 91]), qui consistaient à faire exécuter à des pilotes des tâches extraites d'un manuel d'utilisateur et traduites du vietnamien vers l'anglais.

Le juge observe le sujet et classe sa performance, consigne par consigne, sur une l'échelle suivante : aucune erreur/erreurs mineures/erreurs majeures. La méthode rappelle les pratiques des rédacteurs techniques cherchant à tester l'utilisabilité d'un manuel d'utilisateur au stade d'avant-projet, et connaît d'ailleurs les mêmes limitations. Pour estimer l'efficacité de la traduction, il faut aussi quantifier l'utilité – pas forcément optimale – du texte source, et l'aptitude du sujet à accomplir la tâche sans consignes. Ceci nécessite l'emploi d'un nombre relativement important de sujets, de préférence des professionnels du domaine couvert par le document. La procédure devient alors lente et coûteuse. Enfin, cette approche ne peut s'appliquer qu'à une classe restreinte de documents, à savoir les textes de type « mode d'emploi ».

Dans une expérience plus récente [WHI 00], on a étudié l'acceptabilité de documents traduits pour l'accomplissement de cinq tâches qui faisaient partie du travail habituel des sujets : filtrage, détection, triage, extraction d'informations, et résumé. Les textes, qui constituaient un sous-ensemble du corpus DARPA'94 [WHI 92-94], étaient des articles de journaux traduits du japonais vers l'anglais.

Chaque sujet avait deux missions :

– d'abord, porter un jugement binaire, intuitif et instantané sur l'utilisabilité pour une tâche donnée d'un ensemble de 15 traductions ;

– ensuite, exécuter une des cinq tâches citées, à titre d'exercice.Le filtrage consistait à trier des traductions selon leur pertinence par rapport à un thème donné (oui/non/indécis), la détection à les trier selon cinq centres d'intérêt, et le triage à les classer par ordre de pertinence à l'intérieur de trois domaines d'intérêt.

Un même jeu de 15 traductions a servi à ces trois exercices. Pour l'exercice d'extraction d'informations, les sujets devaient coder les différents types d'entités nommées (personnes, lieux, dates, etc.), alors que pour le résumé il s'agissait d'indiquer à quel degré les informations présentes dans la traduction humaine de l'article source étaient préservées par la traduction automatique. Sept traductions ont servi à ces deux derniers exercices.

Quant aux métriques, on a fait appel à celle qui est normalement adoptée pour l'exercice en question :

– rappel pour le filtrage et la détection ;
– rappel et précision pour l'extraction ;
– fidélité, sur une échelle de 5 à 1, pour le résumé ;
– et classement ordinal relativement à un classement étalon pour le triage.

On peut interpréter les scores pour chaque tâche comme une mesure de la tolérance de celle-ci envers une traduction imparfaite, et arriver par là à un classement des tâches elles-mêmes en termes de tolérance relative. Le classement issu des jugements spontanés et celui issu des exercices se sont avérés identiques, la première méthode étant donc beaucoup plus rentable en termes d'effort et de temps.

Si l'approche précédente vise à estimer directement la qualité d'un texte traduit en vue d'une utilisation particulière en aval de la traduction, l'analyse des erreurs linguistiques dans les textes traduits par un système de TA prétend à une application plus générale, en ce sens qu'elle vise la bonne formation linguistique aux niveaux morphologique, lexical et syntaxique [LEH 88]. Le premier problème est de s'entendre sur une typologie des erreurs qui ne soit ni floue ni subjective [FLA 94].

La solution de Loffler-Laurian consiste à établir les catégories d'erreurs sur la base des corrections apportées à des traductions brutes par plusieurs post-éditeurs différents [LOF 96], ce qui conduit au tableau suivant :

– vocabulaire et terminologie ;
– sigles et noms propres ;
– prépositions ;
– déterminants ;
– temps verbaux ;
– voix verbales ;
– modalités ;
– négations ;
– ordre des mots ;
– problèmes généraux d'agencement.

Il faut ensuite statuer sur la gravité des erreurs, et prendre la décision éventuelle de les pondérer [MIN 93]. Cette décision dépendra de la finalité de l'évaluation et va éventuellement attribuer une pondération différente à une même erreur selon son impact sur la compréhension, ou bien sûr le temps d'édition, ou encore sur la

difficulté de correction des algorithmes. Ce dernier cas suppose une évaluation dite « en boîte de verre », où le développeur a accès aux représentations intermédiaires des modules de traitement du système de TA. De ce fait, certaines catégories d'erreurs peuvent dépendre de l'architecture du système [COR 03], contrairement aux catégories génériques qui relèvent des évaluations « en boîte noire ».

Une autre approche de la bonne formation linguistique, applicable aussi au taux de couverture du système, consiste à construire des batteries de phrases tests (*test suites*) qui mettent en jeu de façon systématique et exhaustive les structures syntaxiques de la langue source, et qui visent souvent les points de contraste par rapport à celles de la langue cible. Les phrases tests permettent au développeur d'évaluer de manière contrôlée la performance du système, et permettent même d'automatiser la détection des erreurs[2].

Il est difficile d'atteindre le même degré d'objectivité lorsqu'il s'agit de faire évaluer par des juges humains les attributs d'adéquation et de précision (voir le tableau 13.1 et le site FEMTI). De façon générale, il faut employer un nombre suffisant de juges pour pouvoir pallier à leur subjectivité et à la variabilité des jugements qui en découle. Par adéquation (*suitability*) nous entendons des caractéristiques du texte cible considéré indépendamment du texte source, notamment ici la lisibilité et l'intelligibilité.

La lisibilité, dite aussi fluidité (*fluency*), caractérise une phrase qui se laisse lire facilement et naturellement. Le juge lit le texte traduit phrase par phrase, sans savoir quelles informations sont censées y être présentes, et accorde à chaque phrase une note sur une échelle allant, par exemple, de 1 à 5. Tous les points sur l'échelle peuvent être ancrés dans une définition, ou bien on peut se borner à définir seulement les deux extrémités de l'échelle, et supposer que les points intermédiaires délimitent des intervalles de qualité constants. Pour l'intelligibilité, on peut procéder de manière identique, sauf que l'échelle ira cette fois-ci du « complètement inintelligible » au « parfaitement intelligible ». On peut calculer le score pour le texte dans son ensemble en faisant la moyenne des scores pour les phrases individuelles, et traiter la variabilité entre juges en ayant recours aux techniques statistiques habituelles.

L'évaluation de la précision (*accuracy*) d'une traduction s'intéresse à la préservation du contenu du texte source dans le texte cible. On peut procéder en demandant aux juges de répondre, après lecture du texte, à des questions à choix multiples ; plus les réponses sont justes, plus la précision, ou l'informativité (*informativeness*) de la traduction est considérée grande. La compilation de tels

2. TSNLP : http://tsnlp.dfki.uni-sb.de/tsnlp/ ; DIET : http://diet.dfki.de/c_as.html.

questionnaires exige, cependant, des compétences particulières et du temps, ce qui rend cette approche relativement coûteuse, même si elle est assez objective. Il est donc plus courant de faire appel au principe de l'échelle et au texte source pour évaluer la fidélité de la traduction.

Plusieurs procédures sont possibles. Si l'on dispose de juges bilingues (encore relativement coûteux), on peut aligner les textes source et cible et inviter les juges à indiquer, segment par segment, dans quelle mesure les informations contenues dans le texte source sont préservées dans la traduction. John B. Carroll a introduit une variante intéressante sur ce thème [PIE 66] : les juges ont d'abord lu le segment traduit pour ensuite noter l'informativité du texte source sur une échelle allant de « contient moins d'informations que la traduction » à « fait toute la différence du monde ». Avec des juges monolingues (moins coûteux), c'est une traduction humaine qui remplace le texte source comme texte de référence, mais l'opération de traduction risque elle-même d'introduire des distorsions, comme nous avons déjà constaté.

On peut procéder à la manière de Carroll, mais le plus souvent les sujets lisent la traduction humaine de référence avant la traduction automatique. Cette traduction de référence peut être rédigée en reproduisant le contenu propositionnel, mais non les effets stylistiques de l'original, afin de ne pas distraire les juges. Cela suppose qu'il est possible de simplifier sans perte d'informations.

13.3.2. *Métriques automatisées*

La motivation pour l'élaboration de métriques automatisées, en termes de coûts et de temps, apparaîtra avec encore plus force quand nous aurons apprécié plus loin les ressources mobilisées lors des grandes campagnes d'évaluation. Pour l'instant nous nous bornons à une appréciation des principes et des limites de quelques expériences portant sur l'automatisation. La démarche commune consiste, dans un premier temps, à calculer un score qui soit en corrélation étroite avec les jugements humains, en général des jugements d'intelligibilité, de fluidité ou de précision. Dans un deuxième temps, quand la fiabilité de la métrique aura été suffisamment démontrée, elle pourra être employée de façon autonome.

Si les tentatives d'automatisation remontent au moins à [BRE 94], c'est la métrique BLEU [PAP 01] qui a eu le plus d'impact. Le paramètre fondamental est la proximité de la traduction automatique par rapport à une ou plusieurs traductions professionnelles. La mesure de la proximité est basée sur le taux d'erreurs au niveau des mots (*word error rate*), métrique adoptée avec succès pour la reconnaissance de

la parole. On calcule les *n*-grams pour des valeurs de *n* allant de 1 à 4 ; les 1-grams correspondraient à la fidélité et les valeurs plus élevées de *n* rendraient compte de la fluidité. On peut jouer sur différentes pondérations de ces valeurs et de la pénalité dite de brièveté qui sanctionne les traductions plus courtes que la phrase de référence (voir, par exemple, [DOD 02, NIE 00, VOG 00]).

Lors de la première mise à l'essai de cette méthode, deux groupes de juges (unilingues et bilingues) ont évalué la qualité de cinq traductions vers l'anglais à partir du chinois, sur 250 paires de phrases. L'échelle allait de 1 (très mauvais) à 5 (très bon). La corrélation des jugements de ces évaluateurs unilingues avec les scores BLEU a été très forte.

Des critiques ont été adressées à BLEU à plusieurs titres. D'abord, la métrique exigerait plusieurs traductions de référence (au mieux quatre), dont la production coûte cher.

Cependant, [COU 03] a trouvé que l'on obtient de fortes corrélations avec les juges humains, même en utilisant une seule traduction de référence, à condition traiter 500 phrases au lieu de 250.

Ensuite, BLEU semblerait privilégier les systèmes statistiques en leur attribuant des scores plus élevés que ne le font les juges humains [COU 03, BAB 03, BAB 04]. Plus encore, [TUR 03] montre que BLEU produit des corrélations moins bonnes sur des documents longs et prétend qu'une métrique qui ne considère que les 1-grams serait plus fiable[3], tout en reconnaissant que les métriques automatisées sont loin de pouvoir remplacer les jugements humains. Cette mise en garde est énoncée aussi par [AKI 03], qui propose la métrique RED [AKI 01] basée sur la distance d'édition des mots (*word edit distance*), c'est-à-dire les opérations d'édition nécessaires pour transformer une chaîne de mots en une autre. RED serait moins tolérante envers les remplacements et les déplacements de mots, mais moins sensible que BLEU au choix des traductions de référence et plus robuste envers les co-occurrences à distance.

Même si [COU 03] a utilisé BLEU pour évaluer des langues cibles autres que l'anglais, apparemment avec succès, [AKI 03] attire l'attention sur la tendance de BLEU à sous-estimer la qualité de la traduction là où des particules sont omises ou mal traduites. Cette observation, qui vaut pour le japonais, pourrait s'appliquer aussi aux langues morphologiquement plus riches que l'anglais, comme le français. Dans le même sens, [OCH 01] fait remarquer que le taux d'erreur au niveau des mots ne

3. http://nlp.cs.nyu.edu/GTM/.

distingue pas les mots importants des mots peu importants. Cette remarque vise non seulement les particules, mais aussi la variation légitime au niveau lexical, ce qui requiert plusieurs traduction de référence pour BLEU. Inspirée des techniques d'extraction d'information, [BAB 04] génère automatiquement des pondérations pour les *n*-gram, compte tenu des différences entre les fréquences de ceux-ci dans le texte présent et leurs fréquences dans le reste du corpus à traduire[4].

L'approche proposée par [RAJ 01, RAJ 02] est motivée par le désir de classer des systèmes de TA selon leur performance sans faire appel à des traductions de référence. Elle s'appuie sur la définition de deux scores, l'un syntaxique et l'autre sémantique. Le premier est calculé sur la base du profil quantitatif des dépendances syntaxiques identifiées par un analyseur syntaxique.

Le second part de l'hypothèse que si le contenu sémantique d'un document est bien préservé en traduction, la position du document source dans l'espace de vecteurs sémantiques extrait d'un corpus de référence en langue source sera comparable à la position du document cible dans l'espace de vecteurs sémantiques extrait du corpus de référence en langue cible. Bien que prometteuses par rapport à BLEU, ces deux métriques exigent la mise en œuvre de ressources et d'outils linguistiques importants, en l'occurrence le parseur XELDA et – comme corpus de référence – le corpus JOC composé de 6229 documents du Journal Officiel de la Communauté européenne.

Nous pouvons donc conclure que beaucoup de travaux restent à faire dans le domaine des métriques automatisées et qu'il est sans doute vain d'espérer trouver une seule métrique qui réponde aux nécessités de tous les intéressés.

13.4. Analyse des campagnes d'évaluation en TA

Le passage en revue des grandes campagnes d'évaluation en TA témoigne d'une évolution intéressante des motivations : mesurer la rentabilité de la TA par rapport la traduction humaine (ALPAC 1966) ; évaluer le rendement des subventions à la recherche (DARPA 1992, 1994) ; stimuler les recherches à l'aide d'un protocole simple permettant de comparer les progrès réalisés (NIST 2000) ; et développer une métrique automatisée fiable, adaptée au français (CESTA 2002).

4. http://www.comp.leeds.ac.uk/bogdan/ltv-mt-eval.html.

13.4.1. *Campagne ALPAC, 1966*

Nous commençons cet historique par le fameux rapport ALPAC [PIE 66[5]] qui a donné un coup d'arrêt aux recherches en TA en détournant les subventions du gouvernement américain vers l'intelligence artificielle et le TALN. Cette recommandation a résulté d'une expérience d'évaluation de systèmes de TA anglais-russe, qui avait conclu que la TA était plus lente, moins précise et plus chère que la traduction humaine.

Quatre textes scientifiques ont été traduits par trois traducteurs humains et trois systèmes de TA. Ensuite 36 phrases ont été extraites au hasard de chacun des textes et présentées aux évaluateurs, chacun ne jugeant qu'une seule traduction d'une phrase source donnée, au total 144 phrases chacun. Pour l'attribut de l'intelligibilité, 18 étudiants ne connaissant pas le russe ont eu recours à une échelle à 9 points, dont nous avons défini les extrémités plus haut. La mesure de la fidélité a été faite avec les méthodes bilingue et unilingue, et l'échelle à 10 points déjà décrite. On a observé une forte corrélation non seulement entre les jugements d'intelligibilité portés par les évaluateurs unilingues et les bilingues, mais aussi entre l'intelligibilité et la fidélité, ce qui amènerait à conclure que pour comparer des « systèmes » de traduction humains ou automatiques, il suffirait de mener les expériences moins onéreuses d'évaluation de l'intelligibilité pour en déduire la fidélité.

13.4.2. *Campagnes initiées par la DARPA, années 1990*

Les campagnes de la DARPA dans les années 1990 avaient pour but de mesurer et de comparer les performances de prototypes issus de trois projets de recherche qui instanciaient des principes de traitement différents (statistiques, linguistiques, hybrides) et qui traduisaient à partir de trois langues sources différentes (espagnol, français, japonais), d'où une nécessité absolue de métriques « boîte noire » [WHI 92-94].

La précision a été caractérisée par deux tests différents – le questionnaire à choix multiples (informativité), et la mesure sur une échelle de 1 à 5 de la fidélité des segments traduits par les systèmes par rapport aux segments correspondants dans une traduction de référence humaine, lue en premier. L'adéquation a été caractérisée par un test de fluidité basé sur l'échelle à cinq points décrite plus haut.

L'envergure des campagnes est impressionnante : en tout 14 systèmes ont participé et, pour chaque couple de langues, 100 textes sources de quelque 400 mots

5. Disponible également à l'adresse : http://www.nap.edu/books/ARC000005/html/.

chacun ont été traduits par les systèmes et par deux traducteurs humains. Pour chaque métrique, chaque traduction a reçu entre 6 et 25 jugements ; le score attribué au texte intégral est la moyenne des jugements individuels, comme le score attribué au système est la moyenne des scores de ses traductions.

Les conclusions sont venues conforter celles de l'ALPAC : la qualité des traductions humaines était supérieure ; les deux mesures de précision étaient fortement corrélées, celles-ci étant aussi bien corrélées avec la fluidité. Si la validité des métriques semble indépendante du genre du texte traduit (scientifique, journalistique), la performance relative de deux systèmes sur un genre textuel ne sera pas forcément maintenue sur un autre.

13.4.3. *Campagnes initiées par le NIST, 2002-2003*

Les deux campagnes organisées par le NIST en 2002 et 2003 ont repris sous une forme légèrement modifiée les métriques humaines DARPA pour la fluidité et la fidélité, et la métrique automatisée de [PAP 01]. Elles ont toutes les deux visé le chinois et l'arabe comme langues sources, auxquelles la campagne 2003 a ajouté une « langue surprise », en l'occurrence le hindi. Et elles ont pris la forme d'un concours ouvert à tous, les ressources linguistiques – textes sources et corpus d'entraînement – étant mis à disposition sur le site Internet du NIST[6]. Les données sources consistaient en une centaine de bulletins d'information (dépêches d'agences) diffusés par les médias ou par l'Internet.

L'objectif principal était de stimuler les travaux en TA en comparant les progrès réalisés sur un laps de temps court, comme dans les concours organisés dans d'autres domaines du TALN. Il est intéressant de noter en 2002 comme en 2003 l'idée d'un « éventuel recours à la seule évaluation automatisée dans les campagnes à venir si cela s'avère suffisant ». L'édition 2004 nous dira si les organisateurs estiment leur métrique automatisée désormais assez fiable pour pouvoir se passer de métriques nécessitant des juges humains.

13.4.4. *Autres spécifications d'évaluations : JEIDA, ATA, SAE, CESTA*

Les critères élaborés en 1992 au Japon par la JEIDA [NOM 92a, NOM 92b, ISA 95, TOM 92] sont différenciés en fonction de deux publics :

– les utilisateurs ;

– et les développeurs.

6. http://www.nist.gov/speech/tests/mt/.

Un premier questionnaire de portée économique permet aux utilisateurs d'identifier le type de système qui est le mieux adapté à leur situation actuelle et celui qui est susceptible de répondre le mieux à leurs besoins futurs.

Les questions et la quantification des réponses sont associées à 14 paramètres, dont :

- le type de document (facile, difficile) ;
- la qualité de la traduction (importante, peu importante) ;
- le domaine d'application (limité, non spécifié) ;
- le temps (urgent, pas urgent).

Pour chaque paramètre on calcule un score dans l'intervalle [0 ; 100] et les scores sont visualisés sur une « charte radar » (*radar chart*). Ces 14 paramètres correspondent également à 7 types de systèmes de TA (traduction avec post-édition, traduction de haute qualité, outils interactifs d'aide à la traduction, etc.) de sorte que leurs propriétés sont susceptibles elles aussi de visualisation sous forme de charte radar.

Une simple comparaison visuelle fait ressortir le type de système le plus approprié. Un questionnaire supplémentaire de portée technique permet à ceux qui ont déjà décidé d'installer un système d'évaluer leur degré de satisfaction avec sa performance. Les valeurs des paramètres (le système, son exploitation, les dictionnaires, la qualité de la traduction avant et après enrichissement lexical, etc.) sont visualisées sous la même forme graphique et la correspondance avec le profil souhaité est facilement repérable. L'évaluation technique par les développeurs s'appuie sur les mêmes principes, visant des paramètres tels que les représentations intermédiaires utilisées, l'analyse et la synthèse pour mettre en évidence tout écart entre l'état actuel et l'état cible de développement.

La métrique de qualité SAE J2450 a été élaborée par la Société des ingénieurs automobiles aux Etats-Unis pour permettre l'évaluation objective des traductions de la documentation de maintenance, quelles que soient les langues source et cible, et que la traduction soit automatique ou humaine [SAE 01, WOY 02]. Cette volonté de généralisation rend le schéma peu intéressant pour l'évaluation de la TA en ce sens que les catégories d'erreur sont trop générales pour éclairer utilement le développeur.

En effet, on distingue sept catégories :

- terme erroné ;
- erreur syntaxique ;

– omission ;
– erreur morphologique ou d'accord ;
– faute d'orthographe ;
– ponctuation ;
– erreurs diverses.

Les erreurs sont pondérées et peuvent en plus être classées comme graves ou mineures. Pour arriver à la note finale pour le texte cible, on calcule la somme des valeurs numériques de la totalité des erreurs pour la diviser par le nombre de mots dans le texte source. Bien que ce ne soit pas précisé, on peut imaginer de calibrer la pondération pour privilégier soit la bonne formation grammaticale soit la correction terminologique. Une autre préoccupation serait le temps de révision requis avant livraison de la traduction au client.

Le cadre pour l'annotation d'erreurs proposé par l'Association des traducteurs américains [ATA 02] est beaucoup plus large, identifiant en plus les erreurs de registre et de style, les ajouts, la traduction trop littérale (cas de figure fréquent pour la TA), la traduction trop libre, l'incohérence terminologique (cas moins fréquent), l'ambiguïté, l'indécision et le manque de compréhension. Ce schéma a été conçu pour l'encadrement et la formation de traducteurs humains et semble peu adapté aux besoins de la TA.

CESTA (campagne d'évaluation des systèmes de TA) fait partie des initiatives EVALDA en France, qui ont pour objectif la constitution d'une infrastructure d'évaluation des systèmes d'ingénierie linguistique du français[7]. Ce projet cherche à adapter le protocole de [RAJ 01, RAJ 02] et celui de [BAB 03, BAB 04], entre autres, pour créer une « boîte à outils » destinés aux utilisateurs comme aux développeurs. Parmi les participants à la campagne on compte un système statistique et deux systèmes linguistiques, et en fonction des couples de langues, certains systèmes représentent un stade précoce de développement, alors que d'autres sont déjà avancés, cela permettant aussi de bien mettre les métriques à l'épreuve.

Dans la même optique, les ressources générées doivent inclure des traductions produites par des étudiants en plus de celles faites par des professionnels, et on prévoit un éventail de genres textuels plutôt que de se limiter aux seuls bulletins d'information.

7. http://www.technolangue.net.

13.5. Métaévaluation des métriques

La question de la validité des métriques employées semble être, au vu des critiques exprimées dans les deux sections précédentes, éminemment expérimentale. Comment en effet savoir si certaines métriques mesurent bien la qualité de l'attribut concerné autrement qu'en les comparant à des jugements humains ? Qui plus est, pour certaines métriques automatiques, il n'est pas clair quel est l'attribut mesuré, si bien que l'on doit les comparer avec divers jugements humains. Les critères théoriques de métaévaluation des mesures que nous allons d'abord exposer exigent eux aussi une application expérimentale, en vue de comparer des résultats réels. Nous allons résumer quelques exemples de critères (section 13.5.1), puis illustrer ce type d'étude par une expérience collective récente (section 13.5.2).

13.5.1. *Critères d'évaluation des métriques*

Parmi les critères définis pour évaluer la cohérence des métriques [POP 99] on peut retenir les suivants :

– une métrique doit atteindre sa valeur maximale pour des traductions « parfaites » (selon l'attribut respectif), et seulement pour ces traductions ;

– une métrique doit atteindre sa valeur minimale pour les traductions « les plus mauvaises » (selon l'attribut respectif), et réciproquement les traductions « les plus mauvaises » doivent recevoir des scores minimaux. Ce critère étant difficile à étudier tel quel, on peut en étudier quelques cas particuliers grâce à des contre-exemples, en vérifiant si des traductions particulières de faible qualité reçoivent bien un score faible (par exemple des traductions produites par des systèmes simplistes), et inversement si des traductions qui reçoivent un score faible sont bien de faible qualité (on peut imaginer ici des traductions construites manuellement de façon à obtenir un score faible, sans qu'elles soient réellement déficientes) ;

– une métrique doit être monotone, à savoir elle doit classer les traductions (selon un attribut donné) de la même façon que le feraient des juges humains. Ce critère est testé nécessairement de façon expérimentale.

La comparaison théorique et empirique des métriques, tout particulièrement en termes de fiabilité, de corrélation et de coût, est plus que jamais nécessaire. On peut parler d'un véritable effort de métaévaluation, qui vise à déterminer les métriques les moins coûteuses à appliquer *et* qui sont le mieux corrélées avec les aspects de la qualité qui intéressent les évaluateurs. Cet effort récent peut être mis en relation avec le développement de systèmes de TA de plus en plus performants, qui doivent être évalués souvent pour déterminer si les changements logiciels qui sont constamment effectués permettent d'augmenter les qualités attendues. Ainsi, pour les systèmes

fondés sur un apprentissage statistique, les modifications peuvent être quotidiennes, en fonction des algorithmes d'apprentissage ou des corpus préparés.

13.5.2. *Comparaison des métriques sur des traductions humaines et automatiques*

Dans le contexte des travaux ayant abouti au cadre FEMTI, la consultation des experts et des utilisateurs de la TA était une priorité, afin d'obtenir une image aussi précise que possible des qualités requises et des métriques les plus couramment utilisées. Ces consultations comportaient souvent des applications pratiques, qui permettaient aux organisateurs de tester l'applicabilité du cadre, mais aussi le comportement des métriques qui y figurent en seconde partie.

Par exemple, un atelier organisé à Genève en 2001 proposait aux participants de spécifier des évaluations simples liées aux problèmes de TA auxquels ils étaient confrontés, et de les exécuter dans la mesure du possible. Ce type d'expérience a mis en lumière la façon dont les experts et les utilisateurs spécifient une évaluation – le modèle de l'utilisateur et de la tâche étant parfois insuffisamment précisé – ainsi que leurs préférences pour certaines métriques d'évaluation, et les difficultés d'application – notamment le temps élevé requis par les mesures fondées sur des juges humains (voir 13.3.1). Un effort a été consacré aussi à la définition de métriques plus simples, ou à la simplification de métriques existantes, notamment par l'étude des corrélations entre métriques. Certains des résultats ont été publiés peu après l'atelier [RAJ 01, WHI 01].

Une expérience plus récente visait plus explicitement la comparaison de différentes métriques sur un problème réel, à savoir l'évaluation comparative d'un ensemble de traductions d'un même texte [POP 03][8].

Deux séries de dix traductions étaient proposées aux participants, qui avaient pour objectif de « mesurer leur qualité » selon une ou plusieurs métriques résumées dans le manuel accompagnant l'atelier. Chacun des deux textes source, d'une longueur de 400 mots environ, avait été traduit du français vers l'anglais par divers systèmes de TA disponibles sur Internet (non précisés aux participants), mais aussi par des étudiants en traduction. Les participants à l'atelier étant surtout anglophones, une « traduction de référence » en anglais était également fournie, ce qui leur permettait de ne pas faire appel au texte source. En revanche, les participants ignoraient l'origine des traductions, et en particulier le fait que certaines étaient rédigées par des humains. Une analyse attentive aurait certes pu permettre de repérer

8. Ces documents sont disponibles à http://www.issco.unige.ch/projects/isle/mteval-may02/. L'atelier était organisé en marge de la conférence LREC 2002.

ces traductions, mais l'objectif était de tester si différentes métriques appliquées à ces traductions généraient des scores cohérents, notamment en ce qui concerne les traductions automatiques.

Les participants ont appliqué plusieurs métriques automatiques, notamment fondées sur l'algorithme BLEU [PAP 01] ou sa variante élaborée par le NIST [DOD 02], mais avec différentes traductions de référence. Certains participants ont choisi d'utiliser l'unique traduction de référence fournie, d'autres ont produit des traductions de référence supplémentaires (un procédé relativement coûteux), et d'autres ont évalué chacune des traductions candidates par rapport à toutes les autres traductions considérées comme références – une façon peu canonique, mais intéressante, d'appliquer BLEU. Les métriques humaines choisies étaient la fidélité, l'intelligibilité, le temps de lecture (lié à la lisibilité) et le temps de correction – appliquées certes avec un faible nombre de juges, dans un tel exercice.

Pour résumer les résultats obtenus, le classement des traductions humaines obtenu grâce aux métriques n'est pas le même que le classement préalable établi par leur correcteur académique. Les métriques automatiques, utilisant une traduction de référence construite à partir de la meilleure traduction humaine, attribuent naturellement un score élevé au modèle lui-même, et des scores très bas aux autres traductions – des scores inférieurs même à certaines traductions automatiques. Les métriques appliquées par des juges humains ne parviennent pas non plus à restituer le classement académique. Ces résultats montrent que les méthodes spécifiques employées pour l'évaluation de la TA ne s'appliquent pas convenablement à l'évaluation des traductions humaines. Cela soulève la question de l'évaluation future des traductions automatiques, lorsque leur niveau et le type d'erreurs commises seront comparables à ceux des humains, si cette situation se produit un jour.

L'évaluation des traductions produites par les systèmes apparaît plus cohérente, dans cette expérience. La plupart des métriques permettent de déterminer qu'en réalité les sept traductions automatiques sont issues de seulement quatre systèmes, avec des configurations différentes. Les scores obtenus distinguent de façon cohérente deux paires, l'une toujours meilleure que l'autre. Sur ce point, les scores obtenus automatiquement sont en accord avec ceux des juges humains. Les scores ne distinguent pas de façon cohérente à l'intérieur des groupes : l'ordre est ($a > b$) > ($c > d$) pour la première série, et ($b > a$) > ($d > c$) pour la seconde (« $a > b$ » signifie que le système a est meilleur que le système b). L'expérience montre donc une bonne cohérence des métriques sur les textes issus de la TA et cela malgré le faible volume de données utilisées et l'application variée des métriques basées sur BLEU.

13.6. Perspectives

L'évaluation de la traduction automatique demeure un domaine de recherche très actif – on note même un regain récent d'intérêt, à la mesure des enjeux applicatifs croissants que le domaine suscite. L'objectif principal semble être la réduction des coûts de l'évaluation par le développement de mesures automatiques ou des techniques de classement rapide qui reproduisent, avec un niveau d'approximation raisonnable, les résultats de mesures plus fines, plus fiables, mais plus coûteuses. Cette évolution participe donc d'un changement plus global de la nature de l'évaluation : une évaluation de qualité n'est plus l'apanage de campagnes officielles financées par les décideurs, mais se met à la portée des développeurs de systèmes et les guide dans leurs travaux. On peut également mettre en relation ce changement avec l'apparition de systèmes statistiques de TA, dont les erreurs de traduction diffèrent des erreurs des systèmes symboliques. L'évaluation des premiers systèmes est, de par leur nature, plus proche de la boîte noire que celle des seconds.

Naturellement, la réalisation des objectifs d'ensemble de l'évaluation de la TA passe par une série de travaux focalisés, pouvant être intégrés dans un cadre du type FEMTI décrit plus haut. Il est ainsi peu probable qu'une *seule* mesure de qualité puisse répondre à tous les besoins de l'évaluation. Au contraire, il est probable que plus la qualité des systèmes augmentera, et plus des distinctions fines seront nécessaires. De même, plus les utilisations de la TA se diversifieront, et plus le besoin de mesures de qualité spécifiques à chaque utilisation se fera sentir. Ainsi, pour ceux qui visent une utilisation autonome des textes issus de la TA, la fluidité sera un paramètre déterminant, alors que si la TA est utilisée en complément des traducteurs humains, on préférera une mesure de l'utilité des traductions automatiques pour une tâche donnée. On peut estimer que la recherche en TA adopte plutôt la première perspective, alors que les développeurs de systèmes commerciaux, tout en puisant leur inspiration dans les travaux des chercheurs, adopteront plutôt la dernière, pour des évaluations dépendantes d'un contexte d'utilisation.

13.7. Bibliographie

[AKI 01] AKIBA Y., IMAMURA K., SUMITA E., « Using Multiple Edit Distances to Automatically Rank Machine Translation Output », *MT Summit VIII*, Santiago de Compostela, p. 15-20, 2001.

[AKI 03] AKIBA Y., SUMITA E., NAKAIWA H., YAMAMOTO S., OKUNO H.G., « Experimental Comparison of MT Evaluation Methods: RED *versus* BLEU », *MT Summit IX*, Louisiane, Etats-Unis, p. 1-8, 2003.

[ATA 02] AMERICAN TRANSLATORS ASSOCIATION, Framework for Standard Error Marking, ATA Accreditation Program, http://www.atanet.org/bin/view.fpl/12438.html, 2002.

[BAB 03] BABYCH B., HARTLEY A., ATWELL E., « Statistical Modelling of MT output corpora for Information Extraction », *CL2003: International Conference on Corpus Linguistics*, Lancaster, p. 62-70, 2003.

[BAB 04] BABYCH B., « Weighted N-gram model for evaluating Machine Translation output », *CLUK 2004*, Birmingham, 2004.

[BRE 94] BREW C., THOMPSON H., « Automatic Evaluation of Computer Generated Text », *ARPA/ISTO Workshop on Human Language Technology*, p. 104-109, 1994.

[CHU 93] CHURCH K.W., HOVY E.H., « Good Applications for Crummy MT », *Machine Translation*, vol. 8, n° 1-2, p. 239-258, 1993.

[COR 03] CORREA N., « A Fine-grained Evaluation Framework for Machine Translation System Development », *MT Summit IX*, Louisiane, Etats-Unis, p. 47-54, 2003.

[COU 03] COUGHLIN D., « Correlating Automated and Human Assessments of Machine Translation Quality », *MT Summit IX*, Louisiane, Etats-Unis, 2003.

[DOD 02] DODDINGTON G., « Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics », *HLT 2002 (Human Language Technology Conference)*, San Diego, Californie, 2002.

[EAG 96] EAGLES MT EVALUATION WORKING GROUP, EAGLES Evaluation of Natural Language Processing Systems, Final Report Center for Sprogteknologi, EAG-EWG-PR.2, 1996.

[FAL 91] FALKEDAL K. (DIR.), *Proceedings of the Evaluators' Forum, Les Rasses*, Genève, ISSCO, 1991.

[FLA 94] FLANAGAN M., « Error Classification for MT Evaluation », *AMTA Conference*, Columbia, Etats-Unis, 1994.

[HOV 99] HOVY E.H., « Toward Finely Differentiated Evaluation Metrics for Machine Translation », *EAGLES Workshop on Standards and Evaluation*, Pise, Italie, 1999.

[HOV 03] HOVY E.H., KING M., POPESCU-BELIS A., « Principles of Context-Based Machine Translation Evaluation », *Machine Translation*, vol. 17, n° 1, p. 43-75, 2003.

[ISA 95] ISAHARA H., « JEIDA's Test-sets for Quality Evaluation of MT Systems – Technical Evaluation from the Developer's Point of View », *MT Summit V*, Luxembourg, 1995.

[ISO 00] ISO/IEC, *ISO/IEC 14598-1: Information Technology-Software Product Evaluation-Part 1: General Overview*, International Organization for Standardization/International Electrotechnical Commission, 2000.

[ISO 01] ISO/IEC, *ISO/IEC 9126-1: Software Engineering-Product Quality-Part 1: Quality Model*International Organization for Standardization/International Electrotechnical Commission, 2001.

[LEH 88] LEHRBERGER J., BOURBEAU L., *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*, Amsterdam, John Benjamins, 1988.

[LOF 96] LOFFLER-LAURIAN A.M., *La traduction automatique*, Lille, Presses Universitaires du Septentrion, 1996.

[MIN 93] MINNIS S., « Constructive Machine Translation Evaluation », *Machine Translation (Special Issue on Evaluation of MT Systems)*, vol. 8, n° 1-2, p. 67-76, 1993.

[NIE 00] NIESSEN S., OCH F.J., LEUSCH G., NEY H., « An Evaluation Tool for Machine Translation : Fast Evaluation for MT Research », *LREC 2000 (2nd International Conference on Language Resources and Evaluation)*, Grèce, p. 39-45, 2000.

[NOM 92a] NOMURA H., ISAHARA H., « The JEIDA Report on Machine Translation », *Workshop on MT Evaluation: Basis for Future Directions*, San Diego, Californie, 1992.

[NOM 92b] NOMURA H., ISAHARA H., « JEIDA's Criteria on Machine Translation Evaluation », *IPSJ SIGNotes Natural Language*, Tokyo, Japon, Information Processing Society of Japan, p. 107-114, 1992.

[OCH 01] OCH F.J., NEY H., « What Can Machine Translation Learn from Speech Recognition? » *Workshop on « MT 2010 - Towards a Road Map for MT » at MT Summit VIII*, Espagne, 2001.

[PAP 01] PAPINENI K., ROUKOS S., WARD T., ZHU W.-J., BLEU: a Method for Automatic Evaluation of Machine Translation, Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022), 2001.

[PIE 66] PIERCE J.R., CARROLL J.B., HAMP E.P., HAYS D.G., HOCKETT C.F., OETTINGER A.G., PERLIS A., Computers in Translation and Linguistics (ALPAC Report), report National Academy of Sciences/National Research Council, 1416, 1966.

[POP 99] POPESCU-BELIS A., « L'évaluation en génie linguistique : un modèle pour vérifier la cohérence des mesures », *Langues (Cahiers d'études et de recherches francophones)*, vol. 2, n° 2, p. 151-162, 1999.

[POP 03] POPESCU-BELIS A., « An experiment in comparative evaluation : humans *versus* computers », *MT Summit IX*, Louisiane, Etats-Unis, p. 307-314, 2003.

[RAJ 01] RAJMAN M., HARTLEY A., « Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores », *Workshop on MT Evaluation « Who did what to whom? » at MT Summit VIII*, Espagne, p. 29-34, 2001.

[RAJ 02] RAJMAN M., HARTLEY A., « Automatic Ranking of MT Systems », *Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, vol. 4, p. 1247-1253, 2002.

[SAE 01] SAE INTERNATIONAL, *SAE J2450: Translation Quality Metric*, Warrendale, Etats-Unis, Society of Automotive Engineers, 2001.

[SEN 03] SENELLART J., YANG J., REBOLLO A., « SYSTRAN Intuitive Coding Technology », *MT Summit IX*, Louisiane, Etats-Unis, p. 346-353, 2003.

[SPA 96] SPARCK JONES K., GALLIERS J.R., *Evaluating Natural Language Processing Systems: An Analysis and Review*, Berlin/New York, Springer-Verlag, 1996.

[TOM 92] TOMITA M., « Application of the TOEFL Test to the Evaluation of Japanese-English MT », *Proceedings of AMTA Workshop `MT Evaluation : Basis for Future Directions'*, San Diego, Californie, Etats-Unis, 1992.

[TUR 03] TURIAN J.P., SHEN L., MELAMED I.D., « Evaluation of Machine Translation and its Evaluation », *MT Summit IX*, Louisiane, Etats-Unis, p. 386-393, 2003.

[VAN 79] VAN SLYPE G., Critical Study of Methods for Evaluating the Quality of Machine Translation, European Commission/Directorate for General Scientific and Technical Information Management (DG XIII), BR 19142, 1979.

[VOG 00] VOGEL S., NIESSEN S., NEY H., « Automatic Extrapolation of Human Assessment of Translation Quality », *LREC 2000 (2nd International Conference on Language Resources and Evaluation)*, Grèce, p. 35-39, 2000.

[WHI 92-94] WHITE J.S. *et al.*, *ARPA Workshops on Machine Translation (Series of four workshops on comparative evaluation)*, McLean, 1992-1994.

[WHI 00] WHITE J.S., DOYON J., TALBOTT S., « Determining the Tolerance of Text-Handling Tasks for MT Output », *Second International Conference on Language Resources and Evaluation (LREC'2000)*, Grèce, vol. 1, p. 29-32, 2000.

[WHI 01] WHITE J.S., « Predicting Intelligibility from Fidelity in MT Evaluation », *Workshop on MT Evaluation « Who did what to whom? » at Mt Summit VIII*, Espagne, 2001.

[WHI 03] WHITE J.S., « How to Evaluate Machine Translation », *Computers and Translation: a translator's guide*, Amsterdam, John Benjamins, p. 211-244, 2003.

[WOY 02] WOYDE R., « Translation Needs in Auto Manufacturing », *Multilingual Computing and Technology*, vol. 13, n° 2, p. 39-42, 2002.