Le rôle des métriques d'évaluation dans le processus de recherche en TAL

Andrei Popescu-Belis

IDIAP Research Institute
Centre du Parc
Av des Prés-Beudin 20
Case postale 592
CH-1920 Martigny
Suisse
andrei.popescu-belis@idiap.ch

RÉSUMÉ. *Le traitement automatique des langues (TAL) relève à la fois de la démarche scientifique et de la démarche technologique. Dans les deux cas, l'évaluation des systèmes informatiques implémentés est indispensable pour estimer le succès d'une recherche. S'inspirant du cadre ISO pour l'évaluation des logiciels, et utilisant une typologie des systèmes de TAL fondée sur la place de la langue parmi les données d'entrée ou de sortie, cet article analyse le rôle central des métriques d'évaluation à plusieurs étapes du processus de recherche en TAL. L'accent est mis sur les métriques qui comparent un résultat produit avec des résultats corrects attendus. L'analyse de plusieurs situations d'évaluation, en particulier le cas des systèmes de traduction automatique, illustre l'importance d'un choix cohérent des métriques et de l'utilisation conjointe de plusieurs métriques. L'influence du contexte d'utilisation sur le choix des métriques et le cas des systèmes interactifs sont discutés en conclusion.*

ABSTRACT. *Research in natural language processing (NLP) has both scientific and technological dimensions. In both cases, it is necessary to evaluate the implemented systems in order to assess the success of a study. This article, grounded in the ISO framework for software evaluation, introduces a typology of NLP systems based on the role of language as input or output data, in order to analyze the central role of evaluation metrics at several stages of the NLP research process. The article focuses on the evaluation metrics that compare the response of a system to a set of correct responses. The analysis of several evaluation examples, in particular the case of machine translation systems, shows the importance of a coherent choice of metrics and of the joint use of several metrics. The influence of the context of use on the set of metrics and the case of interactive systems are discussed as a conclusion.*

MOTS-CLÉS : *systèmes de TAL, évaluation, normes ISO, caractéristiques de qualité, métriques d'évaluation.*

KEYWORDS: *NLP systems, evaluation, ISO standards, quality characteristics, evaluation metrics.*

TAL. Volume 48 – n° 1/2007, pages 67 à 91

1. Introduction

La perception de l'évaluation dans le domaine du traitement automatique des langues (TAL) varie considérablement selon les chercheurs, les utilisateurs des technologies ou les bailleurs de fonds. Alors que certains prennent position contre l'évaluation et ses effets sur le monde de la recherche, d'autres déplorent les limites des évaluations actuelles et souhaitent leur généralisation et leur perfectionnement. Il paraît donc légitime de s'interroger sur la nature et l'utilité de l'évaluation en TAL en se fondant sur une conception de l'évaluation suffisamment large pour être communément acceptée, mais aussi suffisamment précise pour énoncer des conclusions à portée applicative. C'est à la définition d'un tel modèle de l'évaluation, à son analyse épistémologique et à ses conséquences sur le choix des métriques d'évaluation que cette étude est consacrée.

Partant de l'observation que les dimensions scientifique et technique du TAL font toutes deux intervenir des systèmes informatiques dotés de fonctionnalités linguistiques (section 2), l'étude adopte le cadre des normes de l'ISO pour l'évaluation des logiciels, dont elle résume les développements les plus récents (section 3). Une analyse de la spécificité des tâches relevant du TAL permet ensuite de mieux cerner les difficultés de l'évaluation dans ce domaine (section 4). Tout en reconnaissant la possibilité que les données traitées par les systèmes de TAL varient considérablement (section 5), une analyse de la place de la langue parmi les données d'entrée ou de sortie conduit à un classement des systèmes en quatre catégories (section 6). Cette conception unifiée de l'évaluation satisfait ainsi aux exigences du TAL en tant que science et technologie à la fois, et suggère que la différence entre ces deux dimensions relève davantage des données que des métriques utilisées pour l'évaluation.

La contribution des métriques d'évaluation à différentes étapes du processus de recherche en TAL est analysée dans la section 7, qui fait référence aux approches empiriques ou guidées par les données, tout en préservant la généralité du modèle proposé. Une analyse des métriques d'évaluation en termes de distances est proposée dans la section 8 pour les différents types de systèmes de TAL à entrée/sortie. Cette analyse débouche sur une série de recommandations pour la bonne définition des métriques, éclairées par des observations tirées de campagnes d'évaluation récentes (section 9). La conclusion de l'étude (section 10) aborde brièvement la place de l'utilisateur, en considérant son influence sur le choix des métriques et son rôle dans l'évaluation des systèmes de TAL interactifs.

2. Objectifs du TAL : entre science et technique

Une discussion du rôle de l'évaluation en TAL est inséparable d'une discussion des objectifs de cette discipline qui, par ses nombreuses ramifications, avoisine la linguistique et les sciences cognitives d'un côté, et l'informatique et l'ingénierie des

systèmes d'information d'un autre. On peut affirmer, en toute généralité, que les recherches en TAL placent sous le signe de la modélisation informatique deux objectifs complémentaires : l'étude des langues et de la faculté de langage d'une part, et l'exécution automatique de certaines tâches linguistiques d'autre part. La première dimension se rattache davantage à une vision scientifique du TAL, alors que la seconde relève du domaine applicatif ou technologique.

La modélisation informatique permet d'opérationnaliser certains modèles cognitifs visant des capacités linguistiques humaines et d'effectuer certains traitements permettant une meilleure compréhension d'un système linguistique, offrant notamment des « instruments » à la linguistique (Habert, 2005). La modélisation informatique peut également viser à reproduire certaines capacités linguistiques humaines indépendamment de la compréhension de leur mécanisme, se rattachant alors à l'ingénierie des langues, dont l'objectif est d'accroître l'efficacité et la productivité des tâches qui font intervenir un contenu linguistique (Cunningham, 1999).

Les rapports sont nombreux entre la dimension scientifique et la dimension technique du TAL. La modélisation informatique d'une capacité linguistique humaine aboutit de fait à un système informatique reproduisant ce traitement. Inversement, la conception de méthodes efficaces pour traiter certaines tâches linguistiques peut inspirer le linguiste ou le cogniticien dans son travail de modélisation, bien que les systèmes les plus efficaces ne soient pas nécessairement ceux qui sont fondés sur les modèles les plus plausibles cognitivement.

Il reste que ces deux dimensions du TAL partagent une composante expérimentale essentielle, qui est l'utilisation de systèmes informatiques dans le processus de recherche et développement, ainsi qu'une interrogation méthodologique commune, portant sur les critères de succès des recherches entreprises. Nous nous interrogerons donc d'abord sur la notion de qualité des systèmes informatiques, et examinerons ensuite le rôle de l'évaluation dans les recherches en TAL[1].

3. Évaluation des systèmes informatiques : normes de l'ISO/IEC

Les travaux dans le domaine du TAL étant essentiellement liés au développement de systèmes informatiques, le jugement de qualité que l'on porte sur ces systèmes constitue de toute évidence un élément essentiel permettant de juger le succès des recherches elles-mêmes. Or, il existe déjà un cadre éprouvé pour décrire

[1] Toutes les recherches en TAL n'exigent pas la construction d'un système : certaines s'intéressent pas exemple aux ressources linguistiques informatisées ou à l'évaluation. La description proposée ici ne concernera donc pas directement ce type de recherches, à moins qu'un système automatique ne soit employé pour aider à développer des ressources linguistiques ou à accomplir une évaluation.

et estimer la « qualité » des logiciels tout au long de leur cycle de développement, et nous allons l'adopter comme point de départ de nos propositions. L'importance réduite du matériel dans les systèmes de TAL justifie en effet l'intérêt exclusif porté à la qualité du logiciel.

La qualité des logiciels a fait l'objet d'une longue série de normes élaborées sous l'égide de l'Organisation internationale pour la normalisation (ISO) depuis le début des années 1990. Certaines normes ont connu plusieurs versions, et l'ensemble a subi une importante réorganisation sous le nom de SQuaRE[2] au début des années 2000 (Azuma, 2001), toutes les normes prévues n'étant pas encore achevées. La série de normes ISO/IEC 9126 est ainsi divisée en cinq groupes :

a. le groupe 9126-1*n* (*n* = 0, 1) offre une perspective d'ensemble sur les notions et les processus liés à la qualité ;
b. 9126-20 normalise la notion de modèle de qualité ;
c. 9126-3*n* (*n* = 0 à 5) s'intéresse aux métriques de qualité des logiciels et à leur documentation ;
d. 9126-40, principale nouveauté de la série, concerne la spécification des qualités requises ;
e. 9126-5*n* (*n* = 0 à 3) décrit le processus d'évaluation de divers points de vue, reprenant ainsi le contenu de la série ISO/IEC 14598-1 à 6 déjà publiée.

La notion de qualité est définie comme l'ensemble des caractéristiques du logiciel qui lui permettent de répondre aux besoins de ses utilisateurs (ISO/IEC, 2001 : p. 11). Selon la norme ISO/IEC 14598-1 (1999 : p. 12, fig. 4) le cycle de vie du logiciel débute en effet par une analyse des besoins des utilisateurs auxquels le logiciel répondra, à savoir les « exigences de qualité à l'usage ». Cette analyse conduit à des spécifications fonctionnelles qui correspondent à des « exigences de qualité externe », qui seront finalement internalisées lors de la phase de conception du logiciel en « exigences de qualité interne ». Lorsque le logiciel est implémenté, il devient possible de mesurer les différents types de qualités afin d'évaluer le logiciel.

L'approche ISO distingue par conséquent trois types de caractéristiques de qualité : internes, externes et à l'usage. Les qualités internes peuvent être mesurées sans exécution du logiciel – évaluations dites « en boîte de verre » – alors que les qualités externes doivent être mesurées en faisant fonctionner le logiciel – évaluations dites « en boîte noire » où l'on s'intéresse aux résultats produits. Enfin, la qualité à l'usage doit être mesurée en plaçant le système dans un contexte d'utilisation, expérimental ou final, et en observant dans quelle mesure le système aide ses utilisateurs à accomplir leurs tâches (nous y reviendrons dans la section 10).

Les caractéristiques de qualité internes influencent naturellement les caractéristiques de qualité externes, mais cette influence est souvent difficile à

[2] SQuaRE : *Software Quality Requirements and Evaluation* (spécification et évaluation de la qualité des logiciels).

prédire pour les systèmes de TAL, comme expliqué à la section suivante. Plus difficile encore est de prédire la qualité à l'usage à partir des qualités internes et externes. D'ailleurs, si la qualité à l'usage est souvent exprimée en termes d'efficacité, efficience (ou rendement), satisfaction et sûreté de l'utilisateur (ISO/IEC, 2004), les qualités internes et externes se décomposent, quant à elles, selon un modèle de qualité bien différent.

Les caractéristiques de qualité internes et externes d'un système sont en effet regroupées en six catégories : fonctionnalité, fiabilité, utilisabilité, efficacité (du système), possibilité de maintenance, et portabilité – catégories qui elles-mêmes se subdivisent en sous-catégories en fonction du type de système évalué et de sa tâche. Dans cette hiérarchie, les éléments dont on peut mesurer directement la qualité sont les subdivisions terminales, appelées *attributs*. Pour chaque attribut on utilise une *métrique* qui assigne à cet attribut, par le biais d'un processus de *mesure*, un niveau de qualité sur une échelle associée à la métrique. Évaluer un système, c'est donc mesurer sa qualité en utilisant une décomposition fondée sur un modèle de qualité accompagné de métriques pour chaque attribut pertinent dans le contexte d'utilisation prévu.

4. L'évaluation comme critère de succès des recherches en TAL

L'aperçu qui précède conduit naturellement à la question suivante : en quoi les propositions de l'ISO concernent-elles le TAL, dans sa visée scientifique ou technique ? Pour ce qui est de la dimension technique, la réponse est sans ambiguïté : lorsque l'on doit déterminer si une recherche à visée applicative a abouti, il faut évaluer le logiciel construit en le comparant aux objectifs initiaux de la recherche afin de déterminer s'ils ont été atteints.

Mais la question de l'évaluation est également déterminante pour les recherches fondamentales, dépourvues de visées applicatives directes. En effet, la modélisation informatique faisant partie intégrante du TAL, le succès de ces recherches est fortement lié aux performances des systèmes utilisés comme modèles. Par conséquent, afin de démontrer le succès d'une recherche à visée théorique, il faudra également rendre compte des performances du système informatique implémenté.

Seule une véritable évaluation permet en effet de démontrer qu'un système possède le comportement prévu par ses concepteurs, dans des conditions expérimentales contrôlées. Ces conditions peuvent d'ailleurs varier considérablement selon la recherche entreprise : seule demeure constante la nécessité de l'évaluation. L'évaluation des performances d'un système implémenté n'est certes pas la seule source d'arguments démontrant la qualité d'une recherche fondamentale, mais s'ajoute à des considérations analytiques sur la plausibilité cognitive ou linguistique du modèle ou sur ses possibilités de généralisation.

Si l'application du modèle ISO à l'évaluation d'un système informatique de TAL permet d'assurer une certaine normalisation de la procédure et de la terminologie (EAGLES EWG, 1996), cela ne résout pas les principales difficultés de l'évaluation. En effet, les nombreuses recherches en évaluation qui continuent d'être présentées aux congrès spécialisés (tels ACL, EACL, LREC ou TALN) montrent que l'évaluation en TAL demeure une question ardue, puisque d'importantes tâches ne disposent pas encore de métriques d'évaluation universellement acceptées, par exemple la traduction automatique de la langue écrite ou parlée (Blanchon *et al.*, 2004), comme nous le verrons plus loin.

La principale difficulté de l'évaluation en TAL est due, selon nous, à l'impossibilité de fournir des spécifications formelles pour les tâches de traitement faisant intervenir les données linguistiques. Il paraît en effet difficile de spécifier une telle tâche sans faire directement référence à la faculté humaine du langage. (Par exemple, on ne peut définir en termes formels la traduction d'une langue à une autre car il n'existe aucune procédure formelle permettant de vérifier qu'un texte est bien la traduction d'un autre.) Or, le modèle ISO exige une spécification formelle pour pouvoir tester si un système fonctionne conformément à ses objectifs. Par conséquent, à défaut d'une telle spécification, ce sont les métriques d'évaluation qui permettent de tester sur des exemples combien le comportement d'un système est proche de celui qui est implicitement souhaité.

De surcroît, puisque la résolution de problèmes en TAL fait souvent usage d'heuristiques plutôt que d'algorithmes déterministes, il est souvent difficile d'établir un lien clair entre les qualités internes d'un système et ses qualités externes. Par exemple, le nombre de règles d'un analyseur syntaxique n'influence pas de manière directe sa couverture. Or, alors que dans la perspective ISO les qualités internes sont supposées figurer en parallèle avec les qualités externes au sein d'un modèle de qualité, dans le cas du TAL ces deux ensembles de qualités et les métriques associées seront souvent très différents, ce qui rend encore plus complexe l'ensemble des métriques à disposition. De nouvelles métriques sont régulièrement proposées en TAL – une dizaine pour la traduction automatique depuis 2002 (Hartley et Popescu-Belis, 2004 ; Callison-Burch *et al.*, 2006) – alors que d'autres travaux visent à organiser les métriques existantes en modèles de qualité spécifiques à certaines tâches (EAGLES EWG, 1996 ; Sparck Jones et Galliers, 1996; Hovy *et al.*, 2002). L'évaluation à l'usage, fondée sur l'utilisation d'un système en vue d'accomplir une tâche possédant ses propres métriques de succès, fait figure à part, et nous y reviendrons dans la section 10 à la fin de l'article.

Les arguments ci-dessus montrent l'importance et la difficulté de l'évaluation dans le processus scientifique du TAL. Il s'agit toutefois à ce stade d'une conception très générale de l'évaluation comme moyen de comparaison entre le comportement souhaité d'un système et son comportement observé. Il ne faudrait pas conclure que tout type d'évaluation est également indispensable, comme par exemple l'évaluation sur de grandes quantités de données. Si l'évaluation est

indispensable à la recherche, le type d'évaluation mis en œuvre dépend en grande partie des objectifs fixés *a priori*, notamment en ce qui concerne les données d'évaluation, sur lesquelles nous allons nous pencher brièvement.

5. Le choix des données d'évaluation

Les méthodes d'évaluation permettent de comparer un modèle informatique à la capacité linguistique que le chercheur vise à modéliser ou à reproduire, mais elles ne contraignent pas *a priori* le choix des données de test, qui peut se faire en fonction des objectifs d'une recherche. Le chercheur est certes tenu d'évaluer son système pour démontrer sa réussite, mais reste maître de délimiter un ensemble de données auxquelles celui-ci est destiné. Les débats méthodologiques qui traversent le TAL concernent ainsi davantage la nature et la taille des données d'évaluation que la présence ou l'absence d'évaluation.

L'adéquation aux données est un critère essentiel de jugement d'une théorie linguistique. Les théories peuvent en effet être évaluées en considérant les phénomènes linguistiques qu'elles permettent de décrire ou d'expliquer. Cette méthodologie ne semble pas sujette à contestation, et le débat entre les différents courants linguistiques porte davantage sur la nature des données à considérer. Chomsky (2002) argumente par exemple contre l'utilisation de grandes quantités de données « brutes », leur préférant une focalisation sur des phénomènes linguistiques plus rares mais plus informatifs, un point de vue qui s'oppose ainsi aux linguistes de corpus. Une métaphore souvent utilisée par Chomsky est celle des expériences de physique qui se concentrent sur des situations en apparence sans rapport avec la réalité afin d'isoler au maximum le phénomène à expliquer. Cela revient ainsi à éliminer ses nombreuses interférences avec d'autres phénomènes, qui se manifestent dans un cadre expérimental insuffisamment contrôlé (Chomsky, 2002 : p. 124-128).

Au sein du TAL proprement dit, il n'est pas rare de voir surgir la question de l'évaluation dans l'opposition entre approches symboliques et approches statistiques, alors que le facteur discriminant nous paraît être davantage la variété des données. Par exemple, Cunningham *et al.* (1995 : p. 4) évoquent l'exemple d'une grammaire de l'anglais conçue vers 1985 pour « couvrir un ensemble de phrases test justifié linguistiquement » mais qui ne permettait d'analyser syntaxiquement presque aucune phrase extraite au hasard d'un journal. La conclusion des auteurs est que « de tels conflits entre les résultats et l'intuition illustrent l'absence totale d'évaluation dans les projets d'analyse syntaxique de l'époque » (nous traduisons).

En réalité, le fonctionnement adéquat d'un système sur un jeu de phrases test constitue une évaluation tout à fait acceptable. Ce qui pose problème dans l'exemple ci-dessus, ce n'est pas l'absence d'évaluation, mais plutôt le passage d'un ensemble de données de test sur lesquelles le système présente un comportement satisfaisant à un autre ensemble, beaucoup plus vaste, sur lequel le comportement n'est plus du

tout satisfaisant. Le cas cité, exemplaire lorsque l'on évoque l'application des systèmes de TAL à des données peu contraintes, illustre plutôt la nécessité de tester un système sur des données typiques de celles qu'il aura plus tard à traiter – à l'exception du cas où l'on souhaite plutôt mesurer l'adaptabilité d'un système à un nouveau domaine ou comparer des systèmes conçus pour des domaines différents. Évaluer des systèmes de TAL sur des domaines très restreints peut être totalement légitime si ces domaines sont conformes aux objectifs d'une recherche.

Les métriques d'évaluation figurent donc nécessairement dans un modèle unifié des systèmes de TAL, par-delà les oppositions entre approches symboliques ou statistiques, modèles de compétence ou de performance (en termes chomskyens), approches rationalistes ou empiriques, ou systèmes de traitement à grande ou à faible échelle (Church et Mercer, 1993 ; Cunningham, 1999 : section 4). Les données apparaissent comme le compagnon indispensable des métriques d'évaluation externes, et nous allons maintenant focaliser la discussion sur la place de la langue parmi ces données.

6. Classification des systèmes de TAL selon les données d'entrée et de sortie

Afin de pouvoir préciser notre conception de l'évaluation et développer un modèle plus spécifique des métriques d'évaluation, nous allons proposer un classement des systèmes de TAL selon la place des données linguistiques dans leur fonctionnement. Nous adoptons ici une approche inspirée de l'image de la « boîte noire » pour nous intéresser seulement aux entrées et aux sorties des systèmes de TAL. Le contenu linguistique – parlé, écrit ou même signé – peut ainsi figurer en entrée d'un système de TAL, en sortie, ou des deux côtés ; le système peut ou non nécessiter une interaction avec un utilisateur humain.

Lorsque la langue figure seulement en entrée d'un système de TAL, on parlera de système d'analyse ou d'annotation (type A) : ce sont le plus souvent des tâches de classification du contenu linguistique en un nombre limité de catégories. Par exemple, selon Habert (2005 : section 1), « l'annotation consiste à ajouter de l'information (une interprétation stabilisée) aux données langagières : sons, caractères et gestes. Elle associe deux ou trois volets : (i) segmentation pour délimiter des fragments de données et/ou ajout de points singuliers ; (ii) regroupement de segments ou de points pour leur affecter une catégorie ; (iii) (éventuellement) mise en relation de fragments ou de points »[3]. La compréhension automatique de textes peut être conçue comme une série de tâches d'annotation, comme dans l'approche de l'évaluation définie par les *Message Understanding Conferences* (Hirschman, 1998).

[3] En toute généralité, les relations entre fragments pourraient aussi faire l'objet d'une catégorisation, et ainsi de suite.

Réciproquement, lorsque la langue figure seulement en sortie d'un système de TAL, on parlera de système de génération ou de synthèse (type G). Le contenu linguistique peut être généré à partir d'informations non linguistiques, comme dans le cas de la génération de bulletins météorologiques, par exemple, ou bien il peut être généré par la transformation des données linguistiques d'entrée : nous parlerons alors plutôt de systèmes de type AG qui combinent l'analyse et la génération.

Certains systèmes enfin n'accomplissent leur fonction qu'au terme d'un échange avec un utilisateur humain, à savoir au terme d'une série d'entrées/sorties conditionnées par des réactions de l'utilisateur, sans que l'on puisse considérer uniquement la dernière sortie du système comme le résultat de l'ensemble des échanges. Cette catégorie correspond donc aux systèmes interactifs, et en particulier aux systèmes de dialogue humain-machine. On parlera alors du type AGI même si la langue n'intervient pas toujours dans leurs entrées et sorties à la fois.

La distinction de ces quatre types – A, G, AG ou AGI – est importante pour la modélisation qui va suivre, mais est-elle exhaustive ? Pour répondre à cette question, nous avons examiné les champs de recherche du TAL tels qu'ils sont classifiés par deux ouvrages récents à visée encyclopédique (Dale *et al.*, 2000 ; Mitkov, 2003). Après avoir traduit les titres des chapitres selon la terminologie française usuelle, nous avons fusionné les deux listes obtenues, tout en éliminant les champs correspondant à des questions méthodologiques ou d'infrastructure (par exemple les automates à états finis ou les ontologies). Le classement selon les quatre types figure dans le tableau 1 ci-dessous.

Cette analyse montre que les principaux champs de recherche du TAL se laissent aisément classer dans l'une des catégories A, G, AG ou AGI, et qu'une majorité de ces champs relèvent de l'analyse ou annotation automatique (type A). La liste des champs diffère selon les deux ouvrages utilisés – les champs qui figurent dans les deux ouvrages à la fois sont marqués avec l'exposant « 2 ».

Plusieurs des champs de recherche indiqués dans le tableau 1 ont fait l'objet de campagnes d'évaluation, et nous les avons répertoriées au niveau francophone. Les tâches du tableau 1 marquées avec l'exposant « J » ont fait l'objet d'évaluations présentées aux journées du réseau FRANCIL de l'Aupelf-Uref en 1997 (Mariani, 1997 ; Adda *et al.*, 1998) alors que celles marquées avec « T » ont été évaluées au sein de la récente plateforme EVALDA de l'action Technolangue (Chaudiron et Mariani, 2006). On constate ainsi que des systèmes des quatre types – A, G, AG ou AGI – ont fait l'objet d'évaluations quantitatives, avec toutefois une prédominance des systèmes d'analyse (A). Nos propositions de modélisation se concentrent également sur le type A, tout en étant applicables aux types G et AG, alors qu'une analyse du cas des systèmes AGI sera proposée dans la section finale.

Type A	**Type G**
Reconnaissance de la parole[J, T]	Génération automatique de textes[2]
Reconnaissance de l'écriture	Génération de rapports à partir de bases de données
Alignement de textes[J, T]	Génération de présentations multimédias
Segmentation en phrases[2]	Synthèse vocale[J, T]
Segmentation en mots	
Étiquetage morphosyntaxique[2, J]	**Type AG**
Désambiguïsation lexicale[2]	Traduction automatique[2, T]
Acquisition d'informations lexicales[2]	Systèmes de question/réponse [T]
Extraction de collocations	Résumé automatique
Extraction terminologique[J, T]	
Analyse syntaxique[2, T]	**Type AGI**
Inférence grammaticale	Systèmes de dialogue/interaction en langage naturel[2, J, T]
Analyse sémantique[J]	Systèmes de dialogue multimodal
Résolution des anaphores et des co-références	Enseignement des langues assisté par ordinateur
Structuration du discours	Systèmes d'aide à la rédaction
Reconnaissance des intentions	Interfaces aux bases de données
Détection de l'auteur	
Recherche documentaire[J]	
Indexation automatique	
Extraction d'informations[2]	
Fouille de textes[2]	

Tableau 1. *Classification des principaux champs de recherche du TAL selon la place des données linguistiques en entrée et/ou en sortie (Dale* et al.*, 2000; Mitkov, 2003). Abréviations : A : analyse, G : génération, I : interaction, 2 : présence dans les deux ouvrages cités, J : évalué aux JST '97, T : évalué par Technolangue.*

7. Présence de l'évaluation dans le processus de recherche en TAL

Le processus de recherche en TAL passe par la conception d'un modèle informatique d'une certaine fonction ou capacité linguistique. Lorsque ce modèle prend la forme d'un système à entrée/sortie, tels les systèmes de type A, G ou AG, la recherche suit généralement les étapes représentées schématiquement dans la figure 1 ci-après. Cette description ne préjuge pas de l'approche choisie pour la modélisation, et englobe aussi bien les systèmes symboliques, à base de règles, que les méthodes fondées sur l'apprentissage automatique, où la phase d'entraînement doit être clairement séparée de la phase de test.

7.1. *Étapes du processus de recherche en TAL*

Lorsqu'une capacité linguistique est choisie au commencement d'une recherche, son étude emprunte deux chemins parallèles (figure 1). Le premier doit cerner les fondements et les limites de cette capacité chez les sujets humains, en produisant ainsi au passage des données de référence, alors que le second vise la modélisation informatique à proprement parler. L'évaluation joue un rôle clé à quatre étapes de ce cheminement, correspondant aux boîtes en trait plein sur la figure 1, la dernière boîte regroupant deux étapes.

Premièrement, les métriques d'évaluation permettent de mesurer la stabilité de la capacité linguistique étudiée à travers les individus, et de déterminer les limites de fiabilité des jugements humains, en comparant plusieurs traitements des mêmes données (voir section 7.2). Si cette fiabilité est insuffisante, les instructions d'annotations, voire l'ensemble de l'étude, doivent être repensés.

Ensuite, ce sont également des métriques d'évaluation qui permettent l'ajustement du système ou l'apprentissage automatique (lorsqu'une telle méthode est utilisée), en indiquant au développeur ou à l'algorithme d'apprentissage une distance par rapport à la capacité linguistique idéale qui est visée.

Lorsque le système est finalisé, l'évaluation proprement dite indique si son comportement est conforme à l'objectif initial, à savoir si sa fonctionnalité principale est satisfaite (voir section 7.3). Notons que pour des systèmes à visée applicative, l'évaluation doit aussi considérer d'autres caractéristiques de qualité du modèle ISO. L'extrapolation des performances futures du système à partir de son score sur les données de test, un point essentiel souvent considéré comme allant de soi, demeure une question difficile à étudier objectivement, comme le montre l'exemple de l'analyseur syntaxique cité à la section 5 ci-dessus.

L'évaluation finale n'exclut pas d'autres types d'analyses, dans la perspective de la capacité linguistique modélisée par un système. Le succès peut être mesuré également par la cohérence interne du modèle utilisé, ses analogies de structure ou de fonctionnement avec la cognition humaine, ou les indications qu'il fournit à propos des connaissances requises pour maîtriser la capacité linguistique étudiée.

Enfin, si les métriques d'évaluation sont peu coûteuses à appliquer (par exemple si les scores sont calculés automatiquement), il est possible de les utiliser en vue d'analyser la contribution des divers composants du système à ses performances globales. On peut ainsi estimer l'utilité des composants individuels d'un système à base de règles (Popescu-Belis, 2003a) ou la pertinence des traits utilisés pour l'apprentissage automatique (Zufferey et Popescu-Belis, 2004). Par exemple, on pourra mesurer pour chaque trait la performance du système en utilisant seulement ce trait : plus la différence avec le score minimal est grande, et plus ce trait est important. On peut également mesurer la performance du système en enlevant tour à tour chaque trait : plus les performances diminuent en enlevant un trait, et plus ce trait est important pour la capacité étudiée.

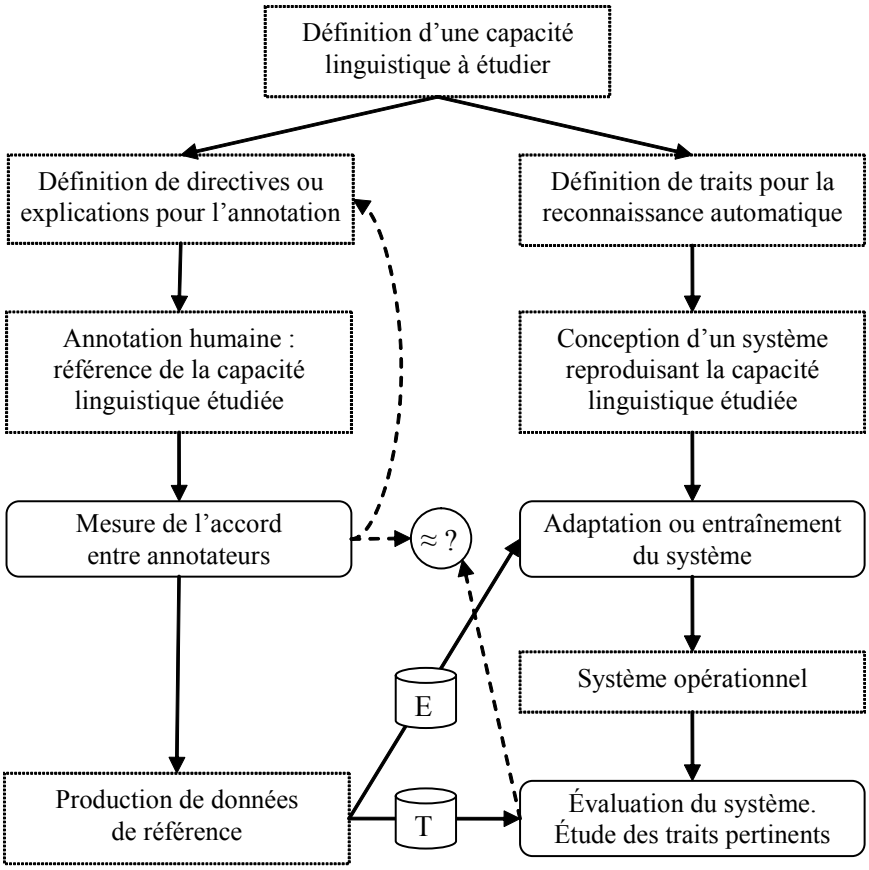

Figure 1. *Le processus de modélisation des capacités linguistiques en TAL. Les étapes encadrées par un trait plein font intervenir des métriques d'évaluation. Les données d'entraînement et de test sont indiquées respectivement par « E » et « T ». Le point d'interrogation signale la comparaison entre les performances humaines et celles du système.*

7.2. *Évaluation de la fiabilité de l'annotation de référence*

Que l'on se place dans la perspective des systèmes de type A, G ou AG, il faut constater que ce sont les humains qui sont considérés comme la référence de la capacité étudiée. Or, leurs performances varient à cause des erreurs d'attention et de l'ambiguïté inhérente aux données linguistiques. C'est pourquoi il est essentiel de quantifier l'accord entre les humains accomplissant une certaine tâche linguistique, ainsi que la stabilité des jugements successifs d'un même sujet, qui serviront de limite supérieure aux performances attendues des systèmes informatiques.

La mesure de cet accord fait classiquement intervenir un ensemble déterminé de données d'entrée. Dans le cas des systèmes de type A, ces données sont annotées (ou catégorisées) par plusieurs sujets humains auxquels la tâche a été expliquée. Les métriques d'évaluation permettent alors d'estimer la différence ou la distance entre deux annotations, ce qui, en fonction de leur nature, peut être une tâche complexe. En outre, la définition d'une échelle de valeurs afin de déterminer un niveau d'accord acceptable constitue également un problème difficile.

L'une des métriques les plus connues mesurant l'accord sur des tâches de classification linguistique est le coefficient *kappa* (Cohen, 1960), mais la signification de son échelle de valeurs est encore sujette à discussion (Carletta, 1996; Craggs et McGee Wood, 2005; Di Eugenio et Glass, 2004). Le type d'argument apparaissant dans ces discussions est bien illustré par l'exemple cité par Habert (2005 : 3.2) : « le paradigme actuel d'évaluation en termes de rappel/précision […] peut conduire à sous-estimer les hésitations inter et intra-annotateurs lors de la construction de données de référence (Sparck Jones, 2001). Les zones de flou semblent globalement minorées. » En d'autres termes, la dernière proposition signale que les scores calculés pour l'accord entre annotateurs ne reflètent pas convenablement la véritable importance des points de désaccord.

La plupart des tâches de type A du tableau 1 peuvent être définies par une seule sortie de référence – en anglais, *ground truth* ou *gold standard*. Les données sont alors accompagnées d'une estimation de la confiance accordée à la référence, autrement dit une mesure de sa variabilité admise, dérivée de l'accord entre annotateurs. Par exemple, en recherche documentaire, on admet la possibilité que les juges puissent manquer des documents justes, ou au contraire en proposer des faux, *a fortiori* dans le cas du *pooling* automatique utilisé parfois pour construire les données de référence (Tague-Sutcliffe, 1996 ; Chaudiron, 2004, section 12.2.1) ; l'estimation de la marge d'erreur peut alors se révéler ardue ou impossible.

Dans le cas des systèmes de type G ou AG, il serait illusoire de chercher à accorder les réponses des évaluateurs humains, étant donné la variabilité des résultats linguistiques considérés comme acceptables. Bien que l'on ne puisse pas exclure les erreurs de performance, il est courant de considérer chaque réponse humaine comme une référence valable de la capacité linguistique étudiée. Par conséquent, les données de référence contiennent dans ce cas un faible échantillon de réponses correctes, car l'espace de toutes les réponses correctes – phrases ou textes acceptables – est très vaste.

L'étude de la capacité linguistique humaine aboutit donc à la production d'un jeu de données de référence, composé de données d'entrée brutes et d'une ou plusieurs sorties de référence (annotations ou texte générés). Les données peuvent à ce stade être divisées en deux parties : certaines seront utilisées pour développer, mettre au point, ou entraîner un système (données signalées par un « E » dans la figure 1), alors que d'autres pourront être cachées aux développeurs en vue d'évaluer plus tard le système sur des données auxquelles celui-ci n'était pas

explicitement préparé (données « T »). La nature et la partition de ces données font partie des conventions *a priori* de la recherche entreprise, comme indiqué dans la section 5, et sont en partie conditionnées par le coût de production des données et par l'utilisation à laquelle on destine le système de TAL.

7.3. *Évaluation des qualités externes relevant de la fonctionnalité*

L'évaluation d'un système de TAL obtenu comme résultat d'une recherche peut se faire selon un grand nombre de caractéristiques de qualité, suivant en cela l'approche ISO, selon laquelle l'estimation de la qualité d'ensemble dépend notamment de l'utilisation prévue. Toutefois, c'est la *fonctionnalité linguistique* des systèmes qui fait la spécificité et la difficulté des recherches en TAL. Alors que beaucoup de paramètres contribuent à la qualité d'un système (par exemple sa vitesse, sa convivialité, sa facilité de mise à jour), nous allons nous concentrer dans ce qui suit sur les caractéristiques de qualité relevant de la fonctionnalité, à savoir la capacité à effectuer un traitement linguistique défini par une description initiale et par des données de référence. La plupart des campagnes d'évaluation initiées par les chercheurs ou les bailleurs de fonds visent ainsi l'évaluation d'une seule caractéristique de fonctionnalité, formulée de façon indépendante d'un contexte particulier d'utilisation[4].

Le principe d'évaluation de la fonctionnalité, pour les systèmes A, G et AG, est la comparaison des résultats produits par le système avec ceux élaborés par les annotateurs humains pour les mêmes données d'entrée. Le critère essentiel de cette comparaison, signalé par un point d'interrogation dans la figure 1, est le niveau d'accord mesuré entre les juges humains eux-mêmes, car celui-ci fournit une véritable unité de mesure de la tolérance sur les résultats produits par le système. En d'autres mots, lorsque l'on compare les résultats du système aux résultats de référence, une distance inférieure à la distance moyenne entre les annotations humaines est synonyme de réponse correcte. De même, on ne peut départager les réponses de deux systèmes si leur distance est inférieure à la variabilité des annotations humaines. Le principe d'évaluation par comparaison des résultats du système avec les résultats de référence s'étend aux systèmes de type G ou AG, sachant toutefois que l'ensemble des réponses « correctes » n'est connu que par un échantillon de réponses. Nous allons y revenir dans les sections suivantes.

[4] Nous avons déjà mentionné les campagnes présentées aux JST Francil en 1997 ou la récente action Technolangue, ainsi que les évaluations MUC. On peut leur ajouter à titre d'exemple le programme ACE de la DARPA – *Automatic Content Extraction* (Doddington *et al.*, 2004) – et les campagnes de la DARPA évaluant la traduction automatique, qui se sont concentrées sur la qualité du texte produit (White et O'Connell, 1994; Doddington, 2002).

8. Métriques fondées sur des distances : importance et utilisation

L'évaluation fondée sur des données de référence et sur des mesures de similarité conçues comme des distances à un ensemble de réponses correctes est un moyen incontournable pour estimer le succès d'une recherche en TAL. Il est par conséquent utile de modéliser le fonctionnement de telles métriques, en s'inspirant notamment d'études récentes, pour en déduire ensuite quelques principes de bonne définition et d'utilisation, qui seront exposés dans la section 9.3.

Nous utilisons ici le terme « métrique » en référence aux normes de l'ISO, bien que ces métriques ou distances ne satisfassent souvent pas les propriétés mathématiques d'une métrique (identité, réflexivité et inégalité triangulaire). Nous ne discuterons pas le cas des métriques internes de la fonctionnalité, qui examinent (sans exécuter le système) les caractéristiques de qualité telles que la nature des algorithmes employés ou la quantité de ressources linguistiques disponibles. Comme indiqué à la section 4, il est difficile de déterminer l'influence exacte de ces caractéristiques internes sur les performances externes des systèmes de TAL, et c'est pourquoi la grande majorité des évaluations en TAL utilisent plutôt des métriques externes (évaluations en boîte noire).

Les métriques externes de la fonctionnalité peuvent être modélisées comme des distances entre la réponse d'un système (sur un certain jeu de données de test) et la réponse ou les réponses correctes attendues. L'évaluation externe de la fonctionnalité détermine si l'exécution du système résultant d'une recherche en TAL est proche du comportement satisfaisant, celui-ci étant défini comme l'ensemble de résultats acceptables possibles, obtenus par traitement des données de référence. Cet ensemble peut ou non être explicité ; ses frontières peuvent être plus ou moins nettes ; sa taille peut être très faible (une seule réponse correcte ou presque) ou bien très grande (notamment si la réponse prend la forme d'un texte entier) ; toutefois, son existence formelle paraît difficile à contester, au risque de ne pouvoir indiquer ce qu'est une réponse satisfaisante.

Réciproquement, on peut formuler l'hypothèse qu'il est nécessairement difficile d'évaluer la fonctionnalité propre au TAL sans disposer des ressources externes que sont les données de référence ou les juges humains. En effet, si une telle métrique « intrinsèque » était trouvée, elle serait sans doute incorporée par les développeurs dans leurs systèmes afin de guider ceux-ci dans la recherche des réponses optimales. Une telle métrique serait donc au fond équivalente à une heuristique spécifique au problème traité, et par conséquent rendrait l'évaluation superflue, du moment qu'elle guiderait nécessairement le système vers une réponse optimale. Tant que les développeurs ne pourront trouver une telle métrique (que la nature même du langage semble exclure), l'évaluation reste indispensable, en utilisant des données de test accompagnées d'annotations ou de réponses de référence, que les développeurs ne connaissent pas à l'avance.

La question de savoir si l'ensemble des réponses satisfaisantes peut être déterminé avec une précision convenable relève de l'étude de la capacité linguistique humaine (partie gauche de la figure 1). Il semble plus facile d'aboutir à un accord des juges humains sur des annotations (type A) que sur du texte généré (types G ou AG). Comme nous l'avons indiqué dans la section 7.2, dans le premier cas, l'annotation attendue pour un jeu de données d'entrée pourra être unique ou presque unique, alors que dans le second cas l'ensemble des réponses satisfaisantes sera vaste et en pratique impossible à énumérer. C'est pourquoi les métriques externes, dans le second cas, font souvent appel à des juges humains, qui pourront déterminer grâce à leurs compétences linguistiques si une réponse du système appartient ou non à l'ensemble des réponses satisfaisantes. Ainsi, alors que les annotateurs humains construisent directement l'ensemble des réponses satisfaisantes pour les tâches de type A, les évaluateurs humains vérifient mentalement si une réponse produite par un système de type G ou AG figure dans cet ensemble.

Selon la nature du système, la conception des métriques comme distances doit être précisée comme suit. Dans le cas des tâches d'analyse (A), la réponse correcte étant connue pour les données de test (unique ou avec une faible variation), la difficulté de l'évaluation réside dans la quantification de la distance entre un résultat produit par un système et la réponse correcte. S'il est en général facile de vérifier par une procédure automatique (et *a fortiori* par des juges humains) qu'une réponse est identique à la réponse correcte, il est bien plus difficile d'estimer une distance lorsque ces réponses sont différentes.

La difficulté à quantifier les distances est illustrée par les systèmes de résolution de la référence, pour lesquels toutes les distances entre annotations qui ont été proposées sont complexes à définir (Vilain *et al.*, 1995; Popescu-Belis, 2000). De manière analogue, l'insuffisance des mesures simplistes de la distance entre deux étiquetages des actes de dialogue a été récemment signalée (Lesch *et al.*, 2005).

Les métriques d'évaluation fondées sur les distances s'étendent également aux systèmes de type G ou AG[5]. L'ensemble des réponses « correctes » n'est alors plus connu que, au mieux, par un échantillon de réponses, de taille en général extrêmement réduite par rapport à l'ensemble tout entier. La qualité d'un résultat, à savoir sa distance à l'ensemble des réponses correctes, doit être estimée à partir de la distance à chacun des éléments de l'échantillon – une estimation *a priori* loin d'être déterministe, mais qui passe aussi par le calcul d'une distance entre la réponse et chaque élément.

Les solutions proposées consistent souvent en une distance qui peut être calculée automatiquement, puis en une suite d'estimations sur un ensemble de données de

[5] En toute logique, on pourrait être tenté d'évaluer séparément les composantes A et G des systèmes AG, mais cela est rarement fait, dans la mesure où la forme de représentation interne utilisée pour combiner les deux traitements n'est pas imposée par la tâche. Une telle évaluation nécessiterait ainsi des données spécifiques à chaque système, dont le coût d'élaboration paraît élevé par rapport au bénéfice retiré.

test, en espérant une convergence des distances estimées vers une valeur stable grâce à la loi des grands nombres. Ainsi, Soricut et Brill (2004) proposent un cadre général pour des métriques automatiques fondées sur des *n*-grammes pouvant être appliquées à diverses tâches de type AG et pouvant reproduire divers types de jugements humains, guidés soit par le rappel soit par la précision.

L'illustration typique des systèmes de type AG est fournie par l'évaluation des systèmes de traduction automatique (Hartley et Popescu-Belis, 2004). Si l'on se restreint, pour simplifier, à l'évaluation phrase par phrase, on constate que l'ensemble des traductions acceptables d'une phrase donnée est potentiellement vaste (selon le domaine et la longueur) mais borné. Par conséquent, il a été proposé d'estimer la distance d'une phrase candidate à l'ensemble des traductions acceptables en calculant la distance à un échantillon de 1 à 4 phrases seulement. Ainsi, la métrique BLEU (Papineni *et al.*, 2001) calcule la distance entre deux phrases grâce au nombre de *n*-grammes commun ($n = 1, 2, \ldots$).

Testée empiriquement, BLEU fait montre d'une certaine corrélation avec les jugements humains, mais de nombreuses exceptions ont été signalées (Culy et Riehemann, 2003 ; Callison-Burch *et al.*, 2006). Par exemple, l'évaluation d'une bonne traduction humaine ne faisant pas partie de l'échantillon de référence est souvent loin d'obtenir des scores BLEU élevés (Popescu-Belis, 2003b ; Hamon *et al.*, 2006). Les juges humains, quant à eux, utilisent parfois également des échantillons de référence pour estimer la qualité d'une réponse – ils y sont même obligés s'ils ne connaissent pas la langue source – et calculent mentalement des distances fondées en général sur le contenu sémantique et l'acceptabilité linguistique.

9. Le choix des métriques et leurs limites

9.1. *Motivations des développeurs et des évaluateurs*

Le défi des évaluateurs, selon notre modèle, est de définir des distances entre des « réponses » (annotations ou textes générés) qui puissent mesurer véritablement les capacités linguistiques désirées. Ces métriques seront le plus souvent utilisées également par les développeurs de systèmes qui chercheront à maximiser les performances quantitatives obtenues, y compris dans le cas où les développeurs jouent aussi le rôle d'évaluateurs. Or, s'il paraît légitime d'améliorer un système de TAL en se guidant sur l'évaluation – comme nous l'avons fait pour la résolution des co-références dans (Popescu-Belis, 2003a) – le risque est que l'amélioration des scores ne traduise pas une véritable amélioration de la « qualité », si les métriques utilisées n'approximent qu'imparfaitement la qualité visée.

Du point de vue des évaluateurs, l'objectif de l'évaluation est de dégager la « qualité » d'un système dans un contexte d'utilisation bien délimité, mais qui ne peut pas être entièrement connu à l'avance notamment à cause de la variabilité et de

l'innovation constante du contenu linguistique à traiter. Par conséquent, l'un des plus grands biais d'une évaluation est d'utiliser des données de test qui ont déjà été à disposition des développeurs du système. Dans ce cas, le résultat du système sera en principe très proche de la réponse correcte, et donc illustre très mal sa capacité future à traiter des données non vues auparavant.

Les développeurs chercheront quant à eux à approximer au mieux le contexte futur d'utilisation du système, et en particulier à optimiser (légitimement) le système pour le type particulier de données le caractérisant. Les développeurs pourront également tenter d'optimiser le système par rapport à la métrique d'évaluation utilisée dans ce processus. Que peuvent alors faire les évaluateurs pour mesurer la véritable qualité des systèmes ? La réponse passe encore une fois par un usage raisonné des métriques.

Idéalement, la distance entre le résultat d'un système et l'ensemble des réponses correctes devrait refléter la qualité du résultat telle qu'elle serait estimée par des juges humains en contexte. Optimiser un système selon une telle métrique reviendrait alors certainement à augmenter sa qualité. Mais il est souvent difficile ou impossible de trouver une telle distance. Une solution prudente consiste alors à utiliser plusieurs métriques « imparfaites » et de ne considérer comme valides que les comparaisons confirmées par toutes les métriques, ou par une majorité d'entre elles[6].

L'exemple classique de distances imparfaites mais complémentaires sont les métriques de rappel et de précision (Salton et McGill, 1983 : chapitre 5), définies à l'origine pour la recherche documentaire, mais applicables à toute tâche visant à identifier des éléments pertinents parmi un ensemble d'éléments candidats. Aucune des deux métriques ne mesure à elle seule la distance entre l'ensemble des éléments identifiés par le système et l'ensemble correct ; c'est leur moyenne harmonique, ou *f-mesure*, qui est en général utilisée. L'utilisation de métriques complémentaires inspirées du rappel et de la précision a aussi été proposée pour la reconnaissance automatique de la parole, où le taux d'erreur de mots semblait pourtant être une métrique unique suffisante (McCowan *et al.*, 2004). De la même façon, les performances des systèmes de résolution des coréférences sont souvent énoncées selon plusieurs métriques (Popescu-Belis, 2000). Enfin, la traduction automatique est l'un des domaines où la multiplication des métriques et le risque d'optimisation sur la plus utilisée sont très plus visibles, comme nous allons le montrer.

---

[6] Une solution plus radicale consisterait à changer de métrique entre l'entraînement et l'évaluation, ou tout au moins éviter d'indiquer laquelle des métriques fournies avec les données d'entraînement sera utilisée pour juger les résultats sur des données de test. Toutefois, une telle approche ne serait acceptable par les développeurs que si le problème à résoudre pouvait être spécifié indépendamment des métriques d'évaluation.

9.2. *Exemple de la traduction automatique*

Dans le cas de la traduction automatique, il a été observé que l'utilisation d'un modèle de langage pour contraindre la sortie d'un système, comme le font tous les systèmes statistiques, permet d'améliorer le score de la métrique BLEU. Le risque que les systèmes de traduction soient optimisés pour BLEU et ne visent plus directement une qualité perceptible par les humains est ainsi souligné par l'un des responsables des campagnes DARPA GALE, J. Olive : « Une bonne partie de l'amélioration des technologies de traduction automatique durant les deux ou trois années précédentes est due aux approches statistiques couplées à des procédures d'optimisation utilisant BLEU. Je crains toutefois que ce paradigme ne soit en train d'atteindre ses limites, si ce n'est déjà fait » (Accipio Consulting, 2006 : p. 33, nous traduisons).

De nouvelles métriques ont donc récemment été étudiées pour les campagnes DARPA GALE, notamment la métrique HTER, qui requiert des juges humains mesurant le temps de post-édition nécessaire pour corriger des traductions automatiques (Przybocki *et al.*, 2006). On constate ainsi un retour aux évaluateurs humains déjà utilisés pour estimer la fidélité et l'intelligibilité des phrases traduites (White et O'Connell, 1994), mais leur coût reste beaucoup plus élevé que celui des métriques automatiques. Ce retour peut être le signe d'une amélioration notable des systèmes de traduction automatique, puisque les distances fondées sur les *n-grammes* deviennent peu fiables lorsque les traductions se rapprochent par leur qualité et par leur variabilité des traductions humaines.

Une solution possible pour préserver l'usage des métriques « automatiques » repose sur l'utilisation de plusieurs métriques différentes, comme dans la campagne française CESTA (Hamon *et al.*, 2006). Les campagnes organisées par le NIST aux États-Unis, dans le cadre du programme TIDES, utilisent également plusieurs métriques ; même si dans le rapport final officiel ne figurent que les scores BLEU. En réalité, et à usage interne, ces campagnes utilisent en plus trois autres métriques automatiques, ainsi que des jugements humains sur une partie des données (NIST, 2006). La variation corrélée de plusieurs métriques fournit alors un indice plus fiable d'une différence de qualité que la variation d'une seule métrique.

9.3. *Critères de cohérence des métriques*

La responsabilité des évaluateurs est d'utiliser des métriques d'évaluation qui reflètent le plus précisément possible la qualité recherchée, ou, à défaut, des combinaisons de métriques complémentaires. En principe, les métriques calculées par des juges humains fournissent une référence pour les jugements de qualité, mais elles sont coûteuses à appliquer et par conséquent difficilement reproductibles en dehors des campagnes d'évaluation. Les distances qui peuvent être calculées automatiquement sont quant à elles beaucoup plus facilement utilisables par les développeurs. Outre la nécessité, énoncée plus haut, que ces métriques capturent la

fonctionnalité linguistique désirée, un certain nombre d'autres critères de cohérence doivent être satisfaits (Popescu-Belis, 1999).

Une métrique définie comme distance à un ensemble de réponses correctes doit d'abord satisfaire les critères dits de « borne supérieure » et « borne inférieure ». La distance devrait atteindre sa valeur minimale (correspondant au meilleur score) pour les résultats corrects et seulement pour eux. Si cela est souvent le cas pour les distances qui comparent des annotations (dans la limite de l'accord entre annotateurs), le critère est moins souvent satisfait pour la comparaison de textes générés, du fait des grandes différences entre plusieurs textes acceptables.

Inversement, une distance devrait atteindre sa valeur maximale (correspondant au score minimal) sur les réponses des systèmes dits « minimaux » (en anglais *baseline*), en d'autres mots pour les plus mauvaises réponses et seulement pour celles-ci. Ce critère est difficile à vérifier car il est difficile d'examiner toutes les réponses minimales possibles, qui peuvent être les réponses d'un système aléatoire, celles d'un système facile à construire, ou celles d'un système qui fait toujours les mauvais choix. Par ailleurs, il est souvent difficile de calculer la valeur minimale théorique d'une métrique, et encore plus difficile de trouver des exemples concrets de réponses qui obtiennent effectivement ce score. Ces incertitudes peuvent être vues comme autant de failles d'une métrique, que les développeurs pourront exploiter pour augmenter leurs scores avec un minimum d'effort.

Un critère plus général de cohérence est celui de la monotonie : la distance par rapport aux réponses correctes d'un résultat *X* meilleur qu'un autre résultat *Y* devrait être plus faible que celle de *Y*. La vérification de ce critère passe souvent par une étude de corrélation avec les jugements de qualité humains[7]. En outre, une distance peut être plus indulgente (respectivement plus sévère) qu'une autre si les scores qu'elles fournit sont systématiquement plus élevés (respectivement plus faibles) que ceux de l'autre. Ces notions permettent d'expliquer pourquoi la mesure *kappa* est préférable à d'autres pour mesurer l'accord entre annotateurs : *kappa* est plus sévère que la f-mesure lorsque les deux annotations sont proches, donc elle permet d'estimer plus précisément de faibles différences d'annotation.

10. Présence de l'utilisateur humain : conclusion et perspectives

Notre analyse a mis en évidence le rôle central des métriques d'évaluation en TAL et les défis que leur définition pose aux évaluateurs. Quelle que soit la visée d'une recherche en TAL, la mesure de son succès passe par l'évaluation du système

[7] Cette corrélation est un critère d'analyse essentiel pour les métriques d'évaluation de la traduction automatique : alors que l'étude proposant la métrique BLEU insiste sur sa corrélation significative avec l'adéquation et la fidélité assignée par des humains (Papineni *et al.*, 2001), d'autres études indiquent au contraire ses limites (Callison-Burch *et al.*, 2006) et proposent des solutions permettant d'améliorer cette corrélation (Babych et Hartley, 2004).

développé, dans des conditions conformes aux objectifs de la recherche. L'évaluation permet aussi de définir une marge d'erreur sur les scores attendus, grâce à l'étude de l'accord entre annotateurs humains. Quant à la mesure des performances, les métriques externes fondées sur la distance entre un résultat et un ensemble de réponses correctes offrent aux évaluateurs flexibilité et reproductibilité. C'est pourquoi elles sont fréquemment utilisées, parfois accompagnées de métriques impliquant des juges humains lorsque l'évaluation requiert davantage de fiabilité.

Étant donné le rôle central de l'évaluation, il ne serait pas absurde de conclure que le problème épistémologique central de toute question étudiée en TAL est la définition de métriques d'évaluation ainsi que l'élaboration de données de référence. La définition de chaque question apparaît en effet inséparable de la définition d'une métrique d'évaluation.

Il est toutefois une classe de systèmes auxquels le modèle des métriques comme distances entre des résultats ne semble pas s'appliquer : ce sont les systèmes interactifs, notés AGI dans notre classification. Ces systèmes, appelés aussi « symbiotiques » (King et Underwood, 2006), ne produisent de résultats qu'après une série d'interactions avec un utilisateur humain, et c'est pourquoi le modèle à entrée/sortie s'applique difficilement à eux, bien que leurs composants soient parfois des modules du type A ou G. Les méthodes d'évaluation de ces systèmes (Dybkjær *et al.*, 2004), à défaut de s'appuyer sur des mesures externes de la fonctionnalité, font appel aux méthodes d'évaluation fondées sur la tâche ou à l'usage.

Une analyse de ces méthodes d'évaluation qui soit compatible avec l'approche généralisatrice proposée ici est également possible, mais dépasse les objectifs de notre article. Une première distinction séparerait l'évaluation fondée sur une tâche, au cours de laquelle le système évalué est utilisé par des sujets humains ou par d'autres modules informatiques dans un contexte idéalisé (mais en principe représentatif d'un contexte réaliste à venir), de l'évaluation à l'usage, qui place le système dans un contexte d'utilisation réel, avec des utilisateurs finaux[8].

L'évaluation par la tâche a été discutée pour la traduction automatique par White *et al.* (2000). Cette approche a inspiré également une métrique récente, HTER, qui mesure le taux d'erreur de traduction en observant l'effort de post-édition nécessaire à un éditeur humain afin d'obtenir une traduction de qualité à partir d'une traduction automatique (Snover *et al.*, 2006). L'évaluation à l'usage, en termes d'efficacité, rendement, satisfaction et sûreté (Bevan, 2001) présuppose une définition encore plus précise de la tâche et du contexte. Dans les deux cas, les principaux problèmes semblent être le coût de l'évaluation et la possibilité de généraliser les résultats obtenus à des tâches ou à des contextes d'utilisation différents.

[8] Ces méthodes s'appliquent en réalité à tout type de système, mais elles deviennent prépondérantes lorsqu'il est difficile d'appliquer des métriques sur des entrées/sorties, dans le cas des systèmes AGI.

Le rôle de l'utilisateur humain est en réalité important pour tous les types de systèmes. En effet, comme l'ont montré les travaux des projets EAGLES et ISLE (EAGLES EWG, 1996; Hovy *et al.*, 2002), la définition d'un modèle de qualité est inséparable de l'analyse préalable du contexte d'utilisation prévu pour un système de TAL. L'importance des caractéristiques de qualité requises d'un système dépend considérablement de la façon dont il sera utilisé, qu'il s'agisse par exemple des systèmes de recherche d'information (Sparck Jones, 2001; Chaudiron, 2004) ou de traduction automatique (Hovy *et al.*, 2002; Estrella *et al.*, 2005).

L'analyse que nous venons de présenter se situe toutefois en amont de la théorie des modèles de qualité contextuels, puisqu'elle vise spécifiquement les métriques associées aux caractéristiques de qualité portant sur la fonctionnalité externe des systèmes de TAL, et non la pondération de ces métriques au sein d'un modèle de qualité complexe. Les développeurs et les évaluateurs institutionnels se concentrent souvent sur les caractéristiques de la fonctionnalité, car celles-ci mesurent les progrès accomplis pour répondre aux défis contemporains du TAL. Si un contexte de qualité précis est spécifié lors d'une évaluation, et *a fortiori* s'il s'agit d'une évaluation à l'usage, un modèle de qualité riche devra être utilisé. En revanche, si l'évaluation concerne une technologie générique, fondamentale, les métriques du type décrit ici seront alors de première importance.

Remerciements

L'auteur remercie pour son soutien le Fonds national suisse de la recherche scientifique (projets n^{os} 200021-103318 et 200020-113604 et Pôle de recherche national IM2), ainsi que les relecteurs anonymes de la revue TAL pour leurs commentaires et suggestions.

11. Bibliographie

Accipio Consulting, Human Language Technologies for Europe, Rapport ITC-IRST et Projet TC-STAR, IST-2002-FP6-506738, 2006.

Adda G., Mariani J., Lecomte J., Paroubek P., Rajman M., « The GRACE French Part-of-Speech Tagging Evaluation Task », *Actes de LREC 1998 (1st International Conference on Language Resources and Evaluation)*, 1998, Grenade, p. 433-442.

Azuma M., « SQuaRE: The Next Generation of the ISO/IEC 9126 and 14598 International Standards Series on Software Product Quality », *Actes de Escom 2001 (12th European Software Control and Metrics Conference)*, 2001, Londres, p. 337-346.

Babych B., Hartley T., « Extending the BLEU MT Evaluation Method with Frequency Weightings », *Actes de ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*, 2004, Barcelone, p. 621-628.

Bevan N., « International Standards for HCI and Usability », *International Journal of Human-Computer Studies*, vol. 55, n° 2001, p. 533-552.

Blanchon H., Boitet C., Besacier L., « Spoken Dialogue Translation System Evaluation: Results, New Trends, Problems and Proposals », *Actes de IWSLT 2004 (1st International Workshop on Spoken Language Translation)*, 2004, Kyoto, p. 95-102.

Callison-Burch C., Osborne M., Koehn, P., « Re-evaluating the Role of BLEU in Machine Translation Research », *Actes de EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, 2006, Trente, p. 249-256.

Carletta J., « Assessing Agreement on Classification Tasks: The Kappa Statistic », *Computational Linguistics*, vol. 22, n° 2, 1996, p. 249-254.

Chaudiron S., « La place de l'usager dans l'évaluation des systèmes de recherche d'informations », S. Chaudiron (éd.), *Évaluation des systèmes de traitement de l'information*, Paris, Hermès, 2004, p. 287-310.

Chaudiron S., Mariani J., « Technolangue : The French National Initiative for Human Language Technologies (HLT) », *Actes de LREC 2006 (5th International Conference on Language Resources and Evaluation)*, 2006, Gênes, p. 767-772.

Chomsky N., *On Nature and Language*, Cambridge, UK, Cambridge University Press, 2002.

Church K.W., Mercer R.L., « Introduction to the Special Issue on Computational Linguistics Using Large Corpora », *Computational Linguistics*, vol. 19, n° 1, 1993, p. 1-24.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and Psychological Measurement*, vol. 20, 1960, p. 37-46.

Craggs R., McGee Wood M., « Evaluating Discourse and Dialogue Coding Schemes », *Computational Linguistics*, vol. 31, n° 3, 2005, p. 289-295.

Culy C., Riehemann S.Z., « The Limits of N-Gram Translation Evaluation Metrics », *Actes de Machine Translation Summit IX*, 2003, New Orleans, LA, p. 71-78.

Cunningham H., « A Definition and Short History of Language Engineering », *Natural Language Engineering*, vol. 5, n° 1, 1999, p. 1-16.

Cunningham H., Gaizauskas R., Wilks Y., A General Architecture for Text Engineering (GATE) - A New Approach to Language Engineering, Technical Report Department of Computer Science, University of Sheffield, CS-95-21, 1995.

Dale R., Moisl H., Somers H. (éd.), *Handbook of Natural Language Processing*, New York, NY, Marcel Dekker, 2000.

Di Eugenio B., Glass M., « The Kappa Statistic: A Second Look », *Computational Linguistics*, vol. 30, n° 1, 2004, p. 95-101.

Doddington G., « Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics », *Actes de HLT 2002 (2nd Conference on Human Language Technology)*, 2002, San Diego, CA, p. 128-132.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., Weischedel R., « The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation », *Actes de LREC 2004 (4th International Conference on Language Resources and Evaluation)*, 2004, Lisbonne, p. 837-840.

Dybkjær L., Bernsen N.O., Minker W., « Evaluation and Usability of Multimodal Spoken Language Dialogue Systems », *Speech Communication*, vol. 43, n° 1-2, 2004, p. 33-54.

EAGLES Evaluation Working Group, EAGLES Evaluation of Natural Language Processing Systems, Final Report Center for Sprogteknologi, EAG-EWG-PR.2 (ISBN 87-90708-00-8), 1996.

Estrella P., Popescu-Belis A., Underwood N., « Finding the System that Suits you Best: Towards the Normalization of MT Evaluation », *Actes de ASLIB 2005 (27th ASLIB International Conference on Translating and the Computer)*, 2005, Londres, p. 23-34.

Habert B., « Portrait de linguiste(s) à l'instrument », *Texto!* vol. X, n° 4, 2005.

Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M., Timimi I., « CESTA: First Conclusions of the Technolangue MT Evaluation Campaign », *Actes de LREC 2006 (5th International Conference on Language Resources and Evaluation)*, 2006, Gênes, p. 179-184.

Hartley A., Popescu-Belis A., « Évaluation des systèmes de traduction automatique », S. Chaudiron (éd), *Évaluation des systèmes de traitement de l'information*, Paris, Hermès, 2004, p. 311-335.

Hirschman L., « Language Understanding Evaluations: Lessons Learned from MUC and ATIS », *Actes de LREC 1998 (1st International Conference on Language Resources and Evaluation)*, 1998, Grenade, p. 117-122.

Hovy E.H., King M., Popescu-Belis A., « Principles of Context-Based Machine Translation Evaluation », *Machine Translation*, vol. 17, n° 1, 2002, p. 1-33.

ISO/IEC, *ISO/IEC 14598-1:1999 (E) -- Information Technology -- Software Product Evaluation -- Part 1: General Overview*, Geneva, International Organization for Standardization / International Electrotechnical Commission, 1999.

ISO/IEC, *ISO/IEC 9126-1:2001 (E) -- Software Engineering -- Product Quality -- Part 1:Quality Model*, Geneva, International Organization for Standardization / International Electrotechnical Commission, 2001.

ISO/IEC, *ISO/IEC TR 9126-4:2004 (E) -- Software Engineering -- Product Quality -- Part 3: Quality in Use Metrics*, Geneva, International Organization for Standardization / International Electrotechnical Commission, 2004.

King M., Underwood N., « Evaluating Symbiotic Systems: the Challenge », *Actes de LREC 2006 (5th International Conference on Language Resources and Evaluation)*, 2006, Gênes, p. 2482-2485.

Lesch S., Kleinbauer T., Alexandersson J., « A New Metric for the Evaluation of Dialog Act Classification », *Actes de Dialor 2005 (9th Workshop on the Semantics and Pragmatics of Dialogue)*, 2005, Nancy.

Mariani J., « Compte-rendu des Premières Journées Scientifiques et Techniques (JST'97) du réseau Francophone de l'Ingénierie de la Langue (FRANCIL) de l'AUPELF-UREF », *http://www.limsi.fr/Recherche/FRANCIL/CR-JST.htm*, page consultée le 18 mai 2007.

McCowan I., Moore D., Dines J., Gatica-Perez D., Flynn M., Wellner P., Bourlard H., On the Use of Information Retrieval Measures for Speech Recognition Evaluation, Research Report IDIAP, RR-04-73, 2004.

Mitkov R. (éd.), *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press, 2003.

NIST (National Institute of Standards and Technology), « NIST 2006 MT Evaluation Official Results », *http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html*, version du 1er nov. 2006, page consultée le 18 mai 2007.

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.

Passonneau R. J., Litman D. J., « Discourse Segmentation by Human and Automated Means », *Computational Linguistics*, vol. 23, n° 1, 1997, p. 103-140.

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, 1999, p. 151-162.

Popescu-Belis A., « Évaluation numérique de la résolution de la référence : critiques et propositions », *T.A.L. (Traitement automatique des langues)*, vol. 40, n° 2, 2000, p. 117-146.

Popescu-Belis A., « Evaluation-Driven Design of a Robust Reference Resolution System », *Natural Language Engineering*, vol. 9, n° 3, 2003a, p. 281-306.

Popescu-Belis A., « An experiment in comparative evaluation: humans vs. computers », *Actes de Machine Translation Summit IX*, 2003b, Nouvelle-Orléans, p. 307-314.

Przybocki M., Sanders G., Le A., « Edit Distance: A Metric for Machine Translation Evaluation », *Actes de LREC 2006 (5th International Conference on Language Resources and Evaluation)*, 2006, Gênes, p. 2038-2043.

Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J., « A Study of Translation Edit Rate with Targeted Human Annotation », *Actes de AMTA 2006 (7th Conference of the Association for Machine Translation in the Americas)*, 2006, Cambridge, MA.

Soricut R., Brill E., « A Unified Framework For Automatic Evaluation Using 4-Gram Co-occurrence Statistics », *Actes de ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*, 2004, Barcelone, p. 613-620.

Sparck Jones K., « Automatic language and information processing: rethinking evaluation », *Natural Language Engineering*, vol. 7, n° 1, 2001, p. 29-46.

Sparck Jones K., Galliers J. R., *Evaluating Natural Language Processing Systems: An Analysis and Review*, Berlin / New York, Springer-Verlag, 1996.

Tague-Sutcliffe J.M., « Special Issue: Evaluation of Information Retrieval Systems », *Journal of the American Society for Information Science*, vol. 47, n° 1, 1996.

Vilain M., Burger J., Aberdeen J., Connolly D., Hirschman L., « A Model-Theoretic Coreference Scoring Scheme », *Actes de MUC-6 (6th Message Understanding Conference)*, 1995, Columbia, MD, p. 45-52.

White J.S., O'Connell T.A., « The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches », *Actes de AMTA 1994 (1st Conference of the Association for Machine Translation in the Americas)*, 1994, Columbia, MD.

White J.S., Doyon J.B., Talbott S.W., « Determining the Tolerance of Text-Handling Tasks for MT Output », *Actes de LREC 2000 (2nd International Conference on Language Resources and Evaluation)*, Athènes, p. 29-32.

Zufferey S., Popescu-Belis A., « Towards Automatic Identification of Discourse Markers in Dialogs: the Case of 'Like' », *Actes de SIGDIAL'04 (5th SIGdial Workshop on Discourse and Dialogue)*, 2004, Cambridge, MA, p. 63-71.