La recherche en traduction automatique depuis un laboratoire de sciences humaines et sociales

Thierry Poibeau

La recherche en traduction automatique depuis un laboratoire de sciences humaines et sociales

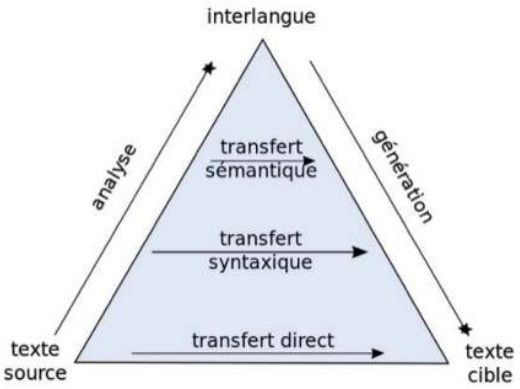
*Directeur de recherche CNRS au sein du laboratoire Langues, Textes, Traitements informatiques, Cognition (Lattice, UMR8094, CNRS / ENS-PSL / Université Sorbonne Nouvelle), Thierry Poibeau est actuellement titulaire d'une chaire Prairie (*Paris Artificial Intelligence Research Institute*) sur le thème du traitement automatique des langues (TAL) et des applications dans le domaine des humanités numériques.*

La première expérience publique de traduction automatique, en 1954 (menée conjointement par des chercheurs d'IBM et de l'université Georgetown). Elle devait permettre la traduction du russe vers l'anglais en quelques années © Wikimedia Commons

La traduction automatique (TA) est un domaine de recherche ancien, dont les débuts ont coïncidé avec ceux de l'informatique, juste après la Seconde Guerre mondiale. La TA a considérablement évolué ces dernières années : jusqu'aux années 2010, les systèmes étaient très défectueux, et ne pouvaient rendre des services qu'à la marge, dans des contextes particuliers (pour avoir une vague idée du contenu d'un texte écrit dans une langue inconnue, par exemple). Depuis quelques années, la donne a changé : les systèmes produisent aujourd'hui des traductions généralement considérées comme de bonne qualité et sont employés (directement ou avec une étape de correction / post-édition) dans un nombre croissant de cadres professionnels.

Les recherches dans le domaine ont pris un tour paradoxal. Alors qu'on a longtemps cru que ce serait l'injonction de « connaissances linguistiques » qui améliorerait la TA, ce sont en fait des systèmes entièrement automatisés, traitant d'énormes quantités de données (mais sans apport humain), qui se sont imposés. Mais cela doit justement nous interroger. Pourquoi de tels systèmes sont-ils plus performants que les systèmes des générations précédentes ? Ces systèmes nous apprennent-ils quelque chose sur la langue, et sur la traduction elle-même ? Plus prosaïquement, est-ce que la TA va mettre les traductrices et traducteurs au chômage ? Enfin, la TA a-t-elle encore un intérêt en tant que domaine de recherche du point de vue des sciences humaines et sociales (SHS) ?

Le laboratoire Lattice s'intéresse fortement à ces questions, et plus généralement à l'évolution du Traitement automatique des langues (TAL), dont la TA a toujours été l'application phare. Les recherches sur ce thème au Lattice portent sur plusieurs axes, mais (peut-être paradoxalement) pas directement sur la mise au point de nouveaux systèmes (ce qui pourrait pourtant sembler être le cœur du domaine). Il y a évidemment encore des améliorations à



Le triangle de Vauquois. À la fin des années 1960, Bernard Vauquois représente les différentes approches de traduction automatique sous la forme d'un triangle, allant du plus simple (traduction mot à mot) jusqu'au plus abstrait, nécessitant d'encoder le sens, l'idée étant qu'il faut atteindre le haut du triangle pour pouvoir traduire avec finesse et exactitude.

apporter aux systèmes de TA, mais ceux-ci sont avant tout fondés sur l'analyse automatique de grandes masses de données (corpus parallèles, c'est-à-dire en situation de traduction et « alignés » au niveau de phrase, d'une part ; corpus monolingues permettant d'avoir des connaissances plus générales sur chaque langue particulière, d'autre part). Ce type de recherche nécessite des infrastructures importantes et a un coût non négligeable, pour des améliorations aujourd'hui somme toute marginales. Surtout, eu égard à la qualité des systèmes aujourd'hui disponibles, les enjeux semblent principalement ailleurs.

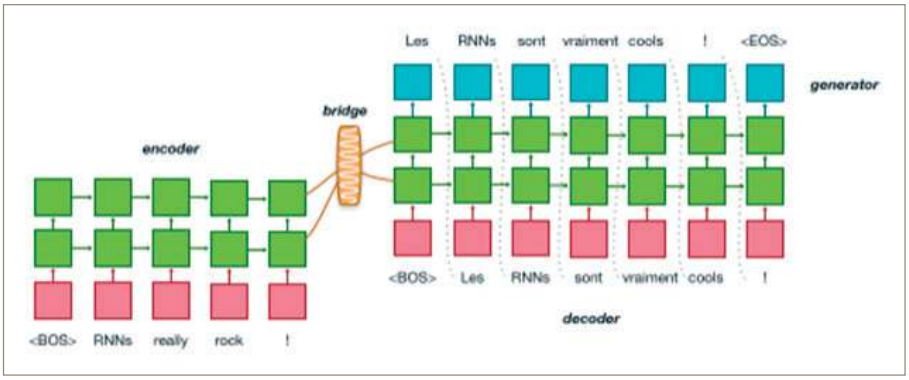
Un premier axe de recherche concerne la réflexion critique sur l'évaluation des performances rapportées. Plusieurs groupes de recherche (essentiellement industriels, mais pas seulement), ont proclamé à la fin des années 2010, pour la TA, des performances égales ou supérieures à celles de traducteurs humains (*supra-human performance*, en anglais). Ce type de déclaration est facilement repris par la presse et laisse penser que la traduction automatique est une question résolue, où la machine fait aujourd'hui mieux que l'humain. En fait, l'évaluation des performances est un domaine de recherche en soi, et il est clair qu'il n'y a pas de mesure parfaite : ce qui fait une « bonne traduction » est quelque chose de complexe, peu formalisable et éminemment subjectif, dépendant en partie du contexte d'utilisation (c'est pourquoi on préfère en général classer différentes traductions en les comparant entre elles : il a été démontré que l'évaluation est ainsi beaucoup plus fiable. Autrement dit, on sait assez bien comparer des traductions, mais on ne sait pas vraiment dire ce qu'est une bonne traduction). Pour en revenir à la question des performances soi-disant « supra humaines » récentes, un examen attentif montre que ces résultats ont été obtenus sur des couples de langue particuliers (avec en général, l'anglais comme langue source ou langue cible), et avec des textes aussi très particuliers (en général, des news, c'est-à-dire des dépêches d'agence et autres textes courts et factuels, avec une langue très littérale et donc relativement facile à traduire). On est encore loin d'avoir des systèmes parfaits, « tout terrain », quels que soient la langue et le domaine considéré.

Il est malgré tout indéniable que les performances des systèmes de TA sont aujourd'hui impressionnantes et doivent être regardées de manière attentive, en particulier pour ce qui concerne les conséquences pour le marché de la traduction. Les performances sont généralement bonnes pour les langues bien représentées sur Internet (en gros, une quinzaine de langues indo-européennes, ainsi que le chinois, le japonais, le coréen, l'arabe, etc.). Les performances reflètent assez directement les investissements dans le domaine, qui eux-mêmes reflètent le marché potentiel (en termes d'audience, de pouvoir d'achat, etc.) Il n'en reste pas moins que l'immense majorité des langues n'est pas couverte : les ressources disponibles pour mettre au point des systèmes pour la plupart des langues sont insuffisantes (même si les modèles multilingues permettent en partie de combler le manque de données parallèles). Les langues restant à couvrir intéressent peu les industriels, faute de rentabilité (même si des entreprises comme Meta ou Google mettent au point des systèmes couvrant plusieurs dizaines de langues, il faut garder en mémoire qu'il y a entre 6 000 et 7 000 langues dans le monde ; même parmi les dizaines de langues actuellement outillées, les performances varient largement).

Au-delà, les systèmes de TA nous interrogent sur la notion de compréhension. Pendant longtemps, on a supposé que pour traduire, il fallait représenter le sens de façon aussi précise que possible (et la notion de contexte, car le sens dépend du contexte). Mais qu'est-ce que le sens ? Comment représenter la notion de contexte ? La TA (et depuis, le TAL plus généralement) a permis de répondre à ces questions en proposant une solution à la fois simple et directe : le contexte, c'est tout simplement les mots qui cooccurrent (apparaissent ensemble) à un moment donné. Les systèmes automatiques sont très efficaces pour analyser des millions, voire des milliards d'exemples et (schématiquement) associer chaque mot à des contextes précis (autrement dit, la bonne traduction en fonction du contexte). D'où les milliards de paramètres souvent évoqués : de par les capacités de calcul quasi infinies actuellement disponibles, il est ainsi possible d'élaborer des modèles ultra précis, au niveau du mot, mais aussi du syntagme et de la phrase, d'où la qualité des traductions obtenues. Ces modèles, très concrets et directement utilisables sur un plan pratique, interrogent donc aussi les conceptions que l'on peut avoir sur la langue. Les grandes théories linguistiques sont bousculées au profit d'approches en apparence plus terre à terre, simplement fondées sur l'analyse distributionnelle (la « distribution » des mots en contexte). La sémantique a longtemps été un « monde parallèle » qu'on cherchait en vain, mais peut-être est-ce juste une question d'usage. Ces questions sont débattues et les systèmes actuels ont relancé les discussions avec les collègues de philosophie, de psychologie et de sciences cognitives.

Quid de l'impact de la TA sur le marché de la traduction ? Il est très difficile de répondre à cette question, faute de chiffres fiables quant au marché de la traduction. Le monde de la traduction est un domaine très éclaté (avec la plupart des traducteurs travaillant en tant qu'indépendants, mais dépendant en fait des plateformes de traduction qui recueillent l'essentiel de la demande). On peut toutefois constater plusieurs tendances :

- De nombreuses études convergent pour montrer que la demande de traduction augmente, soutenue par la mondialisation de l'économie (augmentation des échanges entre les différentes parties du monde).
- L'explosion des budgets de traduction est suivie de près dans toutes les organisations et la situation au sein des institutions de l'Union européenne (UE) est intéressante à cet égard. Les institutions restent pour la plupart multilingues, même si des voix se font parfois entendre pour ne garder qu'une poignée de langues, voire passer intégralement à l'anglais. La TA est aujourd'hui largement utilisée, alors que le nombre de traductions augmente et que, suivant les institutions, le nombre de traducteurs humains



L'architecture d'un système de traduction neuronale

se maintient ou, plus généralement, est en baisse. Pour l'UE, la TA est le seul moyen de maintenir aujourd'hui le volume de traduction en respectant les contraintes en termes de ressources humaines et de budget (ainsi, l'irlandais a récemment été introduit comme langue officielle, sans augmentation du nombre de traducteurs dans la plupart des institutions, mais en réorganisant les services de traduction et en utilisant plus fortement la TA).

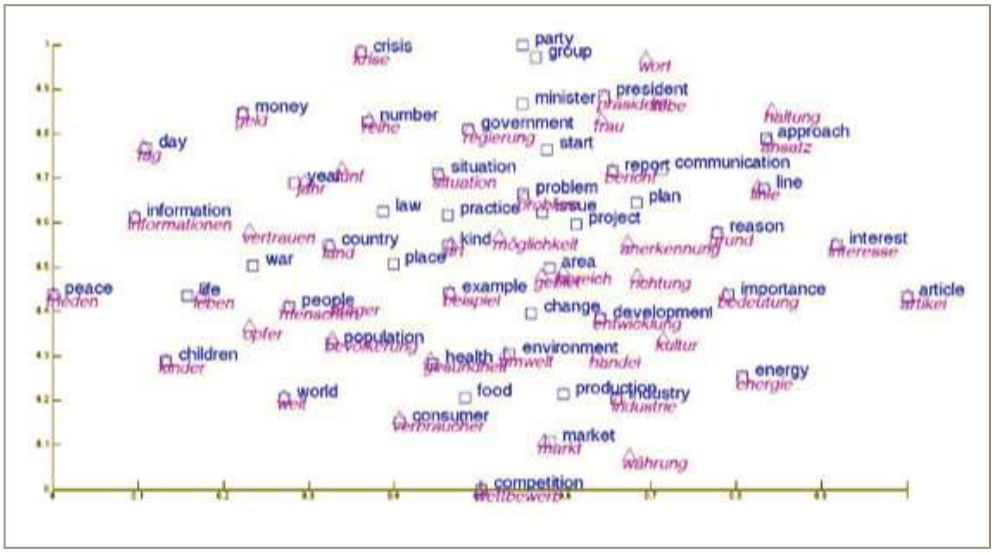
▶ Il faut enfin faire une distinction entre les domaines peu ou pas automatisés (la traduction littéraire) et ceux qui, à l'inverse, ont déjà largement intégré la TA (le sous-titrage multilingue de vidéo par exemple).

Il n'y a donc pas eu jusqu'ici d'impact fort de la TA sur le nombre de traducteurs en exercice, dans la mesure où le marché de la traduction est en expansion, mais la TA est déjà très présente (la TA domine par exemple le monde du sous-titrage de vidéos) et impacte fortement les méthodes de travail et le sens même du métier de traducteur. Les traducteurs travaillent depuis longtemps avec toute une gamme d'outils, mais corriger une traduction faite par un outil est beaucoup plus déstabilisant.

Au niveau de la recherche, notamment au sein des SHS, on notera les multiples réflexions pour essayer d'introduire la TA afin d'augmenter le nombre de textes accessibles en anglais. Le ministère de l'Enseignement supérieur et de la Recherche (MESR) et d'autres acteurs ont multiplié les enquêtes et groupes de travail sur la question, sans réellement aboutir jusqu'ici. Les raisons sont connues : les SHS sont des domaines techniques, et les systèmes de TA génériques n'ont pas le vocabulaire adapté, ce qui risque d'introduire des erreurs et du non-sens dans les traductions produites. Les acteurs ont globalement une image négative de la TA, pour les raisons déjà entrevues (à la fois le manque de précision des systèmes dans les domaines techniques, et les implications sur le métier de traducteur). Ajoutons aussi le fait que la TA fonctionne à partir d'énormes corpus, souvent assemblés de façon obscure, sans l'accord explicite des auteurs / traducteurs. Les traducteurs ont alors une impression de double peine : ils sont directement en concurrence avec la TA, et celle-ci est mise au point à partir de l'exploitation parfois peu éthique de leur propre travail. Un effort pour améliorer la traçabilité des corpus d'entraînement utilisés en TA (et plus généralement, en IA) est donc hautement nécessaire aujourd'hui.

On le voit, l'évolution de la TA ces dernières années a ouvert des voies de recherche prometteuses, mais l'usage des techniques pose aussi des questions redoutables sur le plan social et éthique.

contact&info
▶ Thierry Poibeau, Lattice
thierry.poibeau@cnrs.fr



Représentation graphique d'un espace sémantique bilingue. On voit que chaque mot dans une langue est très proche d'un équivalent dans l'autre langue