Longtemps tournés en dérision, les systèmes de traduction automatique ont accompli d'incroyables progrès en quelques années seulement. Leur secret ? Puissance de calcul et statistiques.

La traduction automatique est-elle un domaine de recherche ancien ? Ou son essor est-il récent, lié à celui d'internet ?

Les deux sont vrais. La traduction automatique s'est énormément développée ces dernières années, et je vous expliquerai pourquoi. Mais l'idée d'automatiser le processus de traduction a toujours fasciné. Des philosophes comme Descartes et Leibniz s'interrogeaient déjà sur la possibilité de formaliser le langage pour le dénuer de toute ambiguïté, et donc le rendre facilement traduisible. Pendant la Seconde Guerre mondiale, grâce au déchiffrement automatisé des messages cryptés allemands, les Alliés ont développé des connaissances quantifiées sur les langues, l'organisation ou la fréquence d'emploi de telle ou telle lettre. Après 1945, les pionniers de l'informatique, qui avaient souvent eux-mêmes travaillé pour la défense, ont imaginé appliquer les mêmes techniques à la traduction, en faisant des comptages sur les mots, et non plus sur les lettres. D'autant qu'avec le début de la guerre froide, les Américains se sont penchés sur les grandes masses de textes en russe, concernant par exemple les avancées scientifiques et techniques des Soviétiques, avec l'idée de les indexer et d'y chercher des informations. Les besoins en traduction automatique sont devenus évidents.

Sur quelles difficultés a-t-on buté ?

Le problème, bien connu, est celui de l'ambiguïté du langage naturel : il y a plusieurs sens par mot. Et beaucoup de mots, de l'ordre de 50 000 dans un dictionnaire usuel du français, 400 000 quand on prend toutes les formes possibles des mots, et jusqu'à 1 million ou plus si on y ajoute les noms propres, les mots composés, les termes techniques… L'ambiguïté existe aussi au niveau des phrases. Si vous écrivez : « J'ai vu Clara avec les jumelles », voulez-vous dire que vous avez regardé Clara à l'aide de jumelles ? Ou bien que c'était elle qui portait des jumelles ? Ou bien encore qu'elle était accompagnée de… deux filles ? La somme des possibilités conduit à une explosion combinatoire qu'à l'époque il aurait fallu coder à la main, sans recours à l'électronique : c'était absolument impossible. En réalité, pour arriver à déchiffrer les mots, il faut percevoir l'intention d'emploi derrière, laquelle ne se trouve pas dans le texte, mais dans le contexte. Et le contexte, l'ordinateur ne sait pas ce que c'est. Pour le prédire, il lui serait nécessaire d'avoir, sur le monde, des connaissances de sens commun, ce qu'il n'a pas.

Qu'entendez-vous par « connaissance de sens commun » ?

Ce n'est pas simple de répondre, car les chercheurs en discutent sans toujours le définir. Je vous donne deux exemples : « On peut manger

des fraises » et « S'il y a un feu rouge, je dois arrêter ma voiture ». Beaucoup de connaissances de sens commun sont acquises par l'expérience, et varient selon la culture : ainsi se repérer dans l'espace peut se faire avec une boussole ou bien en regardant le Soleil. Même si votre culture ignore le sens du feu rouge, vous présumez que la présence de cet objet a une signification : voilà une autre connaissance de sens commun. Si vous croisez un fruit dont vous ignorez tout, la première question que vous allez vous poser est : est-ce comestible ? Vous allez avoir une interrogation par rapport à l'objet. Tout cela échappe à l'ordinateur. En revanche, on peut alimenter les bases de connaissances utilisables par les ordinateurs avec les milliards de textes disponibles sur le web. Une bonne partie des connaissances que nous acquérons au contact du monde, l'ordinateur va les trouver dans ce corpus, comme la proximité fréquente de « manger » et « fraises » (ce qu'on appelle des « cooccurrences ») qui lui permettra d'établir qu'une fraise est comestible. Le lien entre feu rouge, voiture et arrêt est plus difficile à déduire, mais pas impossible. L'ordinateur n'aura pas une connaissance formalisée du lien entre ces différents éléments, mais l'apparition simultanée et fréquente de ces mots dans des phrases permet malgré tout d'inférer qu'il y a un lien entre eux.

Comment l'ordinateur regarde-t-il une langue ? Comme un ensemble de symboles ?

On a longtemps pensé que les langues étaient des systèmes symboliques, et que traduire revenait à effectuer des calculs sur des symboles. Au nom de quoi on a tenté de nourrir la machine avec des règles qui correspondent en partie à ce qu'on apprend à l'école : une phrase se compose d'un groupe nominal et d'un groupe verbal, etc. (ce qui revient à modéliser la syntaxe). Mais ce qui convient pour la syntaxe est très difficile à transposer à la sémantique. Comment modéliser des règles permettant de déterminer *a priori* le sens d'un mot dans une phrase ? C'est là que les statistiques deviennent utiles. Le linguiste américain George Zipf avait remarqué, dès les années 1930, qu'en français, par exemple, il y a des mots comme les déterminants qu'on trouve dans toutes les phrases, et d'autres, très rares, susceptibles de n'apparaître qu'une seule fois même dans un texte long. C'est aussi vrai au niveau de l'usage des mots : la plupart ont un sens fréquent (la fraise comme fruit) et d'autres plus rares (pensez à la fraise de veau). Par défaut, il vaut mieux considérer le premier usage, beaucoup plus fréquent.

Que faire de ces informations statistiques sur les mots ?

À condition de lui fournir des millions, voire des milliards d'exemples d'usage, l'ordinateur va déterminer le sens des mots en contexte. Pour reprendre l'exemple précédent, « J'ai vu Clara avec les jumelles », du fait de l'apparition conjointe fréquente de « voir » et de « jumelles », il va inférer que « jumelles » est sans doute rattaché au verbe : ce serait l'interprétation la plus probable, même si l'autre reste possible. Au-delà, les traitements sont fondés sur une hypothèse aussi simple que puissante : les mots employés dans les mêmes contextes ont des sens proches. S'il y a un trou dans une phrase, l'ordinateur saura fournir l'ensemble des possibilités qui conviendraient pour la compléter, en proposant une liste de mots ordonnés, du plus ou moins probable. Il a ainsi une vision complète du lexique, des proximités sémantiques entre mots, et des liens possibles entre les mots (entre « voir » et « jumelles », par exemple).

La traduction automatique doit donc son succès aux statistiques ?

Oui, et plus largement à deux choses : à la masse de textes aujourd'hui disponibles, sur le web essentiellement, et à la puissance de calcul de l'ordinateur. Désormais, il n'y a plus vraiment de règles au sens où on l'entend dans la grammaire traditionnelle, mais des milliards de contraintes qui s'appliquent sur les mots, comme on l'a vu ou, quand on est dans un cadre de traduction automatique, sur les liens entre mots d'une langue à l'autre. Étant donné l'énorme masse de données disponibles pour l'entraînement, c'est-à-dire pour la mise au point du modèle de traduction, l'ordinateur traduit le plus souvent sans trop d'erreurs, même dans le cadre d'applications grand public où on a affaire à une grande variété de couples de langues, et à des textes de différente nature, sur n'importe quel sujet.