La machine interprète

François Yvon

La machine interprète[1]

François Yvon

Université Paris-Saclay, CNRS, LIMSI

Chacun peut faire l'expérience en quelques clics : les récents progrès des logiciels de traduction automatique (TA) les rendent aujourd'hui utilisables pour une vaste gamme de contextes et d'utilisateurs, comme en attestent les statistiques massives d'utilisation de ces outils. En première analyse, ces progrès résultent d'une conjonction de trois facteurs : la disponibilité de très grands corpus de textes traduits (corpus dits « parallèles »), comme ceux produits par les services de traduction du Parlement Européen ; la conception d'algorithmes d'apprentissage statistiques ou neuronaux (apprentissage dit « profond ou *deep learning*»), capables d'extraire de ces corpus des modèles prédictifs , c'est-à-dire des modèles capables de prendre des décisions de traduction qui reproduisent les décisions des traducteurs humains observées dans les exemples d'apprentissage ; enfin, l'accroissement de la puissance de calcul des ordinateurs qui permet de rendre opérationnelle la mise en œuvre de ces algorithmes à très grande échelle, sur des corpus contenant des milliards de mots.

Les principaux usages de la TA se divisent traditionnellement entre ceux qui permettent l'assimilation superficielle d'un document rédigé en langue étrangère ; ceux qui visent à la publication d'un document et requièrent une intervention humaine dans une phase de correction (ou post-édition) ; et, enfin, ceux qui rendent possible une interaction langagière entre des locuteurs parlant des langues différentes. Les premiers correspondent aux usages massifs des systèmes proposés par les grands opérateurs d'Internet (Google Translate, Microsoft Translator, etc.) ; les seconds sont plus réservés aux utilisateurs professionnels (rédacteurs ou traducteurs) à travers des outils de TA embarqués au sein d'environnements de traduction assistée par ordinateur (TAO) ; enfin, les troisièmes sont typiquement implantés sur des dispositifs mobiles, téléphones portables ou casques de traduction simultanée et sont destinés à faciliter les échanges inter-langues. C'est sur ce dernier type d'applications que nous nous focalisons ici, en exposant sommairement les principes de leur fonctionnement, puis en discutant des principaux défis qui restent à relever pour espérer un jour faire tomber les barrières linguistiques dans les situations de communication interpersonnelle.

Les systèmes de traduction simultanée

[1] Cet article est une version initiale d'un texte paru dans 'Vers le cyber-monde : humain et numérique en interaction', sous la Direction de M. Bouzhegoub, J. Daafouz et C. Jutten, CNRS éditions, Paris, 2021.

Les systèmes de traduction simultanée (TS) constituent la dernière évolution des systèmes de traduction automatique de parole, dont les premiers prototypes datent des années 1980. Initialement limités à l'assistance à la réalisation de quelques tâches finalisées prédéfinies, comme la réservation d'une chambre d'hôtel dans un pays étranger, ces systèmes ont progressivement évolué pour prendre en charge des énoncés de moins en moins contraints : prises de parole dans des institutions multilingues comme le parlement européen, exposés et cours diffusés en ligne, réunions de travail, sans parler des applications militaires. En plus de ces usages domestiques, la traduction simultanée peut être essentielle dans certaines situations critiques : permettre un dialogue entre personnel médical et patients, à l'hôpital comme lors d'interventions humanitaires à l'étranger ; fournir une assistance juridique à des migrants ; faciliter les échanges pacifiés entre militaires et populations civiles lors d'interventions de maintien de la paix, etc. Par ailleurs, la traduction simultanée pourrait également jouer un rôle dans le développement d'agents conversationnels polyglottes : plutôt que de redévelopper des systèmes de dialogue évolués dans toutes les langues, une alternative de plus en plus viable consiste, en effet, à traduire tous les échanges avec la machine depuis et vers l'anglais, langue « maternelle » de Siri, Alexa et autres Cortana.

Encadré

L'activité du traducteur, qui travaille sur des documents écrits, s'oppose par bien des aspects à celle de l'interprète, qui prend en charge des contenus oraux. Ces deux activités sont bien distinctes et correspondent à des cursus de formation professionnelle différents. Compte-tenu de la multiplicité réelle des contextes et usages de traduction envisagées, les praticiens de la TA ne distinguent en général pas aussi clairement ces deux usages et utilisent le terme de « traduction de parole » – qui ne correspond à aucune forme d'activité humaine. Au terme d'« interprétation », qui est ambigu en français, les chercheurs lui préfèrent celui de « traduction simultanée », qui met en avant un trait propre à cette activité, à savoir l'exigence de traduire un discours fugace, avec de fortes contraintes temporelles liées à la situation d'interaction.

Un système de traduction simultanée s'interpose entre plusieurs locuteurs parlant des langues différentes et prend en charge les opérations de traduction qui permettent que chacun des intervenants puisse s'exprimer dans sa propre langue. Pour ce faire, trois grandes étapes de traitement doivent être réalisées séquentiellement : la transcription automatique de la parole en langue source sous une forme textuelle normalisée, la traduction en langue cible, puis la synthèse vocale du texte ainsi produit (voir Figure 1). Chacune de ces étapes de traitement pose des problèmes difficiles et l'amélioration des systèmes de TS implique des progrès de chacun de ces composants. Il existe également des défis spécifiques, relatifs aux principaux contextes d'usage de ces systèmes, qui sont

au centre des questions étudiées par les chercheurs en traduction de parole et sur lesquels nous nous attardons plus longuement dans la suite.

(Insérer Figure 1)

Comme la traduction automatique, et pour des raisons assez similaires, la transcription de parole a fait des progrès remarquables, en vitesse comme en performances. À titre d'illustration, le système de transcription automatique de parole spontanée du LIMSI était en 2016 environ 100 fois plus rapide que celui de 1996, pour un taux d'erreurs divisé par 7. Les outils de transcription automatique sont aujourd'hui intégrés dans de nombreux dispositifs d'interactions familiers, au premier rang desquels les agents conversationnels vocaux (comme Alexa ou Siri). Dans le cadre de la TS, comme pour toutes les applications qui impliquent des conversations spontanées, cette étape de traitement continue de buter sur les questions relatives à l'extrême variabilité de la parole, au besoin de traiter des énoncés arbitraires, pouvant contenir des mots rares et des noms propres. Ainsi, la traduction simultanée de « mais le MOSE[2] pourra-t-il vraiment protéger Venise de l'acqua alta » donne « mais le mot pourra-t-il vraiment protéger Venise de l'aquarelle ? », car le traducteur ne connait pas le projet MOSE et ne comprend pas l'intégration du mot italien « aqua alta » dans une phrase en français). La prise en compte des accents régionaux ou étrangers, et de la parole émotive sont également d'autres problématiques auxquelles se heurte la TS. Enfin, reconnaître la parole dans des environnements bruités reste un problème difficile et constitue une autre source importante d'erreurs de transcription, erreurs qui ne pourront être corrigées par les étapes de traitement ultérieures. À l'autre extrémité de la chaîne de traitement, les systèmes de synthèse vocale à partir du texte constituent des briques de traitement relativement éprouvées, capables de lire des textes arbitraires avec une voix très proche d'une voix naturelle, avec une expressivité encore très limitée, ce qui peut s'avérer insuffisant pour restituer fidèlement l'intonation de l'énoncé source.

Les difficultés de la traduction automatique

L'étape de traduction automatique pose des problèmes encore plus délicats : en premier lieu, parce que la TA statistique ou neuronale n'est jamais aussi performante que lorsqu'elle peut être entraînée sur de très gros volumes de données parallèles, que l'on souhaite aussi représentatives que possible de l'application visée. Si ces données existent, au moins pour quelques domaines bien définis, pour la traduction de textes, elles sont bien plus difficiles à trouver pour les applications de traduction simultanée – puisqu'il est très rare que l'activité des interprètes soit enregistrée et rendue accessible aux outils d'apprentissage automatique (exception faite des corpus issus du travail des interprètes du

[2] Le MOSE est le projet de digues flottantes destiné à protéger Venise des crues.

Parlement européen, mais qui concernent ces énoncés très particuliers que sont les interventions lors des débats parlementaires). À défaut, il faudra s'accommoder de corpus parallèles construits pour d'autres applications, d'autres domaines ou d'autres registres et normaliser la parole transcrite afin qu'elle puisse être traitée par des outils qui ont été entraînés à traduire des textes (plus précisément : des phrases isolées) respectant les normes spécifiques de l'écrit. Ceci impliquera des transformations très substantielles, tant la parole spontanée retranscrite diffère des énoncés écrits : en premier lieu, identifier dans le flot de parole des unités s'apparentant à des phrases ; puis les débarrasser des marques de leur caractère oral en supprimant les pauses remplies (euh, hum, etc), les répétitions, les hésitations, et autres disfluences ; enfin ajouter des marques typographiques et des ponctuations. Tout en sachant que ces ajustements essentiellement formels ne combleront qu'une partie de l'écart entre oral et écrit, qui s'exprime aussi et surtout par des choix lexicaux, des structures syntaxiques, et plus généralement une organisation du discours différentes, conduisant parfois à dérouter les meilleurs systèmes de TA. À titre d'exemple, comparons ainsi l'énoncé très spontané reproduit ci-dessous, extrait d'une intervention télévisuelle de Dominique Voynet en 1995, à sa retranscription normalisée sous forme écrite :

« est-ce que c'est sérieux de proposer de proposer aux gens, et notamment, aux femmes hein, de travailler à temps partiel en étant payées à temps partiel ? ça, c'est de la précarisation, c'est de la c'est c'est de la flexibilité mal comprise. (…) vous savez, je crois la façon de dans laquelle on a euh avec laquelle on a reconnu de façon précipitée la Slovénie et la Croatie et pas la Macédoine et pas euh euh et pas les autres états de l'ex-Yougoslavie euh sous la pression de l'Allemagne, je crois que c'est quelque chose qui nous fait pas honneur aujourd'hui. »

« Est-il sérieux de proposer aux gens, et notamment aux femmes, de travailler à temps partiel en étant payées à temps partiel ? C'est de la précarisation, de la flexibilité mal comprise. (…) Je crois que la façon précipitée dont on a reconnu la Slovénie et la Croatie mais pas la Macédoine ni les autres états de l'ex-Yougoslavie sous la pression de l'Allemagne ne nous fait pas honneur aujourd'hui.»

À cette première difficulté, s'ajoutent les limitations systèmes de TA traduisant des écrits, comme les problèmes liés à la prise en compte de phénomènes linguistiques dont la portée s'étend sur plusieurs phrases. C'est, par exemple, le cas de la traduction de pronoms, comme l'illustre l'échange (imaginaire) suivant :

- (A) Mary bought her daughter a new car -> Mary a acheté une nouvelle voiture à sa fille.
- (B) Combien l'a-t-elle payée ? -> How much did she pay for it?
- (C) I don't know, I don't think it was very costly / Je sais pas, je crois pas qu'elle était très chère.

Cet échange pose en fait plusieurs problèmes de référence : ainsi, dans (B), la traduction vers l'anglais doit produire « for it » alors qu'en l'absence de tout contexte, on pourrait tout aussi bien traduire

« how much did she pay her? » ; dans (C) la traduction en français de « it » doit enfin utiliser « elle », qui réfère à « la voiture » dont il est fait mention deux phrases plus haut. Pour progresser dans la résolution de ces limitations, ici comme pour de nombreux autres cas, doter les systèmes de TA de connaissances factuelles sur le monde et d'une certaine forme de capacité de raisonnement semble aujourd'hui la piste la plus prometteuse – et paradoxalement peut-être une des moins explorée – dans les travaux récents qui se concentrent principalement sur l'amélioration des méthodes neuronales.

À ces difficultés propres au matériau linguistique, s'ajoutent, enfin, toutes les questions relatives à la prise en compte des phénomènes extra-linguistiques (l'intonation, les contacts visuels, les expressions faciales, les postures, etc.) qui sont essentiels au bon déroulement des interactions langagières entre humains. En particulier, leur rôle dans la régulation / négociation des tours de parole, ou encore dans l'établissement et le maintien d'un accord sur le contenu, les termes et modalités de l'interaction verbale sont bien documentés ; a contrario, lorsque ces indices sont moins directement disponibles, par exemple lors de conversations par téléphone, les risques de malentendus et de confusions augmentent fortement. Fortement enracinées dans l'expérience intime des locuteurs, ces formes de communication non-verbales varient fortement d'une culture à l'autre et ne pourront être restituées telles quelles. Faute de les prendre en charge, les outils de traduction simultanée doivent intégrer dans leur conception d'autres dispositifs pour assurer que l'interaction verbale, même médiatisé par la machine, pourra se dérouler avec le moins d'accrocs possible. On pourra, par exemple, proposer à chaque locuteur la retranscription textuelle dans sa langue de ce que la machine a reconnu, afin que les erreurs de la reconnaissance vocale soient repérées au plus tôt ; ou bien encore, faute de parvenir à restituer en langue cible la prosodie et l'émotivité de la langue source, faire entendre la voix originale, tout en donnant à lire la version traduite. Il reste en la matière un énorme champ d'investigation à explorer pour concevoir des dispositifs d'interaction vocale qui inspireront la confiance dans la fidélité des traductions produites, et permettront de détecter – et de réparer – les situations de malentendus.

Si chaque étape de traitement (transcription – traduction – synthèse) pose donc des difficultés qui lui sont propres, et qui peuvent faire l'objet d'améliorations relativement indépendantes, la manière optimale de les enchaîner en séquence fait également l'objet de recherches poussées. Deux aspects ont été particulièrement étudiés. Le premier concerne la spécification des meilleures interfaces entre ces différentes étapes. Il est ainsi possible de limiter l'effet des incertitudes de la transcription vocale en proposant à la traduction plusieurs séquences de mots également probables, en laissant aux étapes ultérieures de traitement le soin de lever les ambiguïtés restantes. Le second concerne le cadencement des différentes opérations de traitement. La traduction d'un énoncé n'est en théorie possible qu'une fois que le locuteur aura achevé son tour de parole, retranscrit par une ou plusieurs phrases. Ceci implique, pour le récepteur, un temps d'attente qu'il s'agira de réduire au minimum pour préserver la spontanéité de l'interaction : diminuer la latence du système implique de commencer à traduire (et de

synthétiser) des fragments incomplets, tâche souvent risquée, voire parfois impossible lorsque la construction du début de la phrase cible requiert des informations qui ne viendront qu'à la toute fin de la phrase source. Les situations de traduction depuis l'allemand vers le français fournissent de nombreux exemples de cette difficulté, puisque le verbe principal allemand n'est souvent pas connu avant les derniers mots de la phrase allemande, alors qu'il sera souvent nécessaire pour commencer à former la phrase en français. C'est le cas dans l'exemple suivant où le verbe principal (möchte / aimerait) n'est disponible à l'interprète qu'à la toute fin de la phrase allemande, alors qu'il est nécessaire dès les premiers mots de la phrase traduite :

- Weiss du, dass Paul den nächsten Weihnachtsurlaub im bretonischen Haus bei uns verbringen möchte?
- Sais-tu que Paul aimerait passer les prochaines vacances de Noël dans la maison de Bretagne avec nous ?

Comment éviter les erreurs ?

Ces questions d'architecture logiciel continuent de faire l'objet d'une recherche soutenue, avec, en particulier, l'espoir de parvenir à une traduction directe de parole à parole qui s'affranchirait de l'enchaînement des trois étapes décrites plus haut pour produire des énoncés sonores en langue cible à partir des enregistrements en langue source. De nombreux travaux s'intéressent déjà à exploiter les techniques neuronales d'apprentissage de bout-en-bout pour condenser les deux premières étapes (transcription et traduction) en une seule, afin de générer directement des transcriptions en langue cible. Cette perspective éviterait en particulier que les erreurs des étapes initiales (en particulier de la transcription automatique) se répercutent sur les traitements ultérieurs.

Reste enfin, pour le concepteur de systèmes de traduction simultanée, un dernier écueil à surmonter : celui de l'évaluation de la qualité du système. Il est bien sûr possible d'évaluer séparément chaque composant. L'absence de traduction de référence, produite par un humain, est un obstacle à la mise en œuvre des évaluations automatiques qui ont été si bénéfiques pour le développement de la TA. Et quand bien même de telles références seraient disponibles, il serait légitime de questionner leur utilité, tant le succès d'une interaction réussie ne peut se mesurer par comparaison avec une référence linguistique, mais demande d'évaluer le succès des buts de l'interaction (l'accomplissement d'une tâche, dans les cas les plus simples ; la fluidité de l'échange, pour des conversations non-finalisées). On peut penser que la diffusion à très large échelle des outils de TS permettra à leurs opérateurs de progressivement collecter les grands corpus de conversation naturelle qui font aujourd'hui défaut et aideront à répondre de manière plus satisfaisante à ces questions.

Les progrès des outils de traitement de la parole et de traduction automatique neuronale semblent nous rapprocher de cet idéal babélien si souvent mis en scène dans la littérature de fiction, et dans lequel les barrières entre les langues seront entièrement abolies, la technologie permettant à chacun de s'exprimer dans la langue de son choix en toute circonstance. Nous avons discuté dans cet article des principales étapes de traitement automatique qui sont nécessaires pour mettre en œuvre ces systèmes de traduction simultanée, en essayant de montrer qu'en dépit d'avancées tangibles (au moins pour quelques contextes et couples de langues), un long chemin reste à accomplir pour atteindre cet horizon. Progresser dans cette voie permettra, dans le plus court terme, d'améliorer les communications inter-linguistiques dans un certain nombre de situations stéréo-typiques (voyages à l'étranger, interventions dans des réunions internationales, etc). Des retombées sont également attendues dans le domaine de l'aide à l'apprentissage des langues.

Ajoutons pour terminer qu'il faut rester sans illusion sur la nature de cet accomplissement : quand bien même les barrières linguistiques seraient abolies, les sources d'ambiguïté et de mécompréhension continueront à proliférer dans les interactions verbales. N'est-ce pas ce que nous enseignent déjà nos interactions quotidiennes dans notre langue maternelle ?

Eléments de bibliographie

Claire Blanche-Benveniste (2010) : Approches de la langue parlée en français - Nouvelle édition. Collection « L'essentiel Français », Editions Ofrys.

Frédéric Landragin (2013). Dialogue homme-machine : Conception et enjeux. Hermès-Lavoisier, Paris.

Jean-Paul Haton (2016) : La parole numérique, analyse, reconnaissance et synthèse du signal vocal. Collection « l'Académie en poche », Presses de l'Académie Royale de Belgique.

Thierry Poibeau (2019) : Babel 2.0 : Ou va la traduction automatique ? Editions Odile Jacob, Paris.

Glossaire

Architecture logicielle. L'architecture logicielle décrit d'une manière schématique les différents éléments d'un systèmes informatique, leurs interrelations et leurs interactions. Dans le cas de l'interprète, l'architecture logicielle décrit comment les trois traitement principaux (transcription, traduction, synthèse) interagissent et se cadencent.

Agent conversationnel. Un dialogueur, chatbot ou encore agent conversationnel est un programme informatique capable de soutenir des interactions langagières avec un utilisateur humain, par exemple pour répondre à des questions ou pour recevoir des demandes d'actions à exécuter.

Apprentissage automatique. L'apprentissage automatique désigne un ensemble d'algorithmes destinés à extraire des régularités à partir d'un ensemble d'exemples, à des fins d'analyse

exploratoire ou d'aide à la décision. L'algorithme d'apprentissage s'entraine pour améliorer ses performances dans l'exécutions de tâches sans être explicitement programmé pour une tâche particulière.

Corpus parallèle. Un corpus parallèle est un ensemble de documents associés à leur traduction dans au moins une autre langue. L'alignement des corpus met en correspondance les fragments en langues source avec les fragments équivalents en langue cible, pour des fragments de taille variable : des textes, des paragraphes, des phrases, voire des mots.

Latence. Pour un système vocal, la latence est la durée qui s'écoule entre l'énoncé du message et la fin de son traitement par la machine. Dans le cas de l'interprète, on s'intéresse en particulier à la durée entre le début d'un énoncé et le commencement de sa traduction par la machine.

Interaction langagière. Une interaction langagière est un échange entre deux acteurs (humains ou non humains) qui exploite des canaux linguistiques : la parole, l'écriture, la langue des signes.

Langue cible. La langue cible est la langue vers laquelle on doit traduire un énoncé.

Langue source. La langue source est la langue depuis laquelle on doit traduire un énoncé.

Méthode d'apprentissage neuronale. Parmi les méthodes d'apprentissage automatique, les méthodes neuronales s'inspirent (à gros grains) de certains principes de fonctionnement du cerveau humain, en particulier du fait que les représentations et les calculs sont distribués sur un ensemble de composants élémentaires effectuant des opérations simples et qui sont de lointains analogues des neurones biologiques.

Synthèse vocale. La synthèse vocale consiste à produire un énoncé sous forme orale, le plus souvent à partir de la forme écrite, mais également parfois à partir d'une représentation conceptuelle.

Transcription automatique (ou reconnaissance vocale). Un logiciel de transcription automatique accomplit l'opération inverse de la synthèse vocale, qui consiste à retranscrire sous une forme écrite normalisée un énoncé oral.

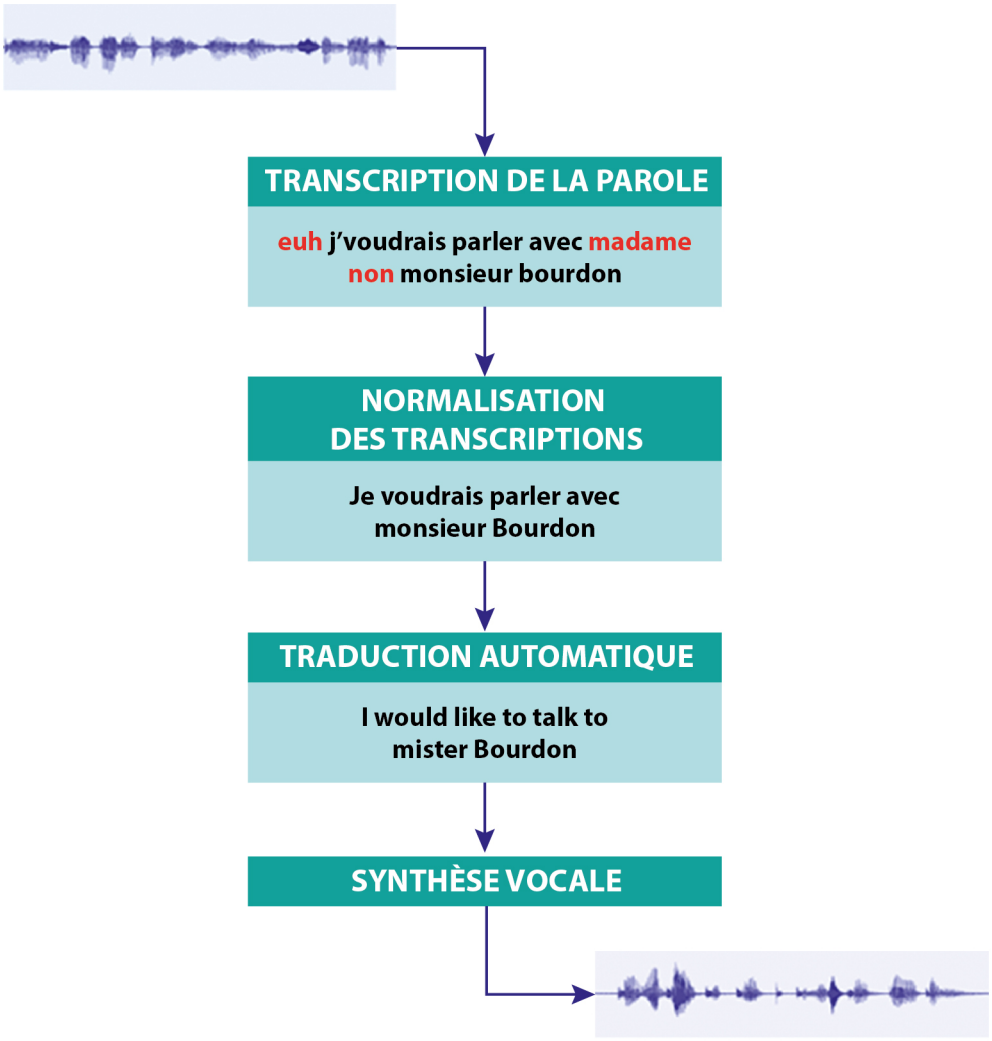

Figure 1 : les étapes d'un système de traduction de la parole