Parler n'est pas suffisant

La communication naturelle ne se limite pas aux échanges oraux. De ce fait, une interaction homme-machine n'est idéale que si la communication est «multimodale», c'est-à-dire si elle intègre la parole, mais aussi les gestes de désignation et la perception visuelle. Lorsque nous parlons, nous faisons souvent référence à des objets de notre environnement, que nous voyons et désignons, par des déterminants ou pronoms particuliers, tels «une», «la», «cette», «celle-ci»…

Naturelle chez les êtres humains, cette multimodalité pose à la machine des problèmes d'interprétation et d'influence réciproque entre les modes d'interaction. Par exemple, avec une interaction associée à deux désignations, telle «place cette figure au-dessus de ce texte», la parole, le geste et la perception ne peuvent pas s'interpréter indépendamment.

Depuis dix ans, notre équipe, composée d'informaticiens et de linguistes, étudie la mise en place d'une telle communication : à partir d'études linguistiques, cognitives et informatiques, nous élaborons tout d'abord des modèles informatiques de représentation des connaissances, puis des algorithmes pour traiter un problème particulier, dont nous testons ensuite les capacités de reconnaissance, de compréhension et d'interprétation sur des applications concrètes. Quand l'algorithme est performant, nous l'intégrons enfin à un logiciel ou à une interface.

Ainsi, un de nos modèles linguistiques intègre et comprend les hésitations, les reprises et les corrections, nombreuses à l'oral et qui, bien que diverses, respectent une structure prévisible. Un autre modèle est spécialisé pour interpréter les gestes de désignation naturelle de l'utilisateur, et un troisième modèle, dit de représentation mentale, unifie l'ensemble des informations connues sur chaque composant d'une tâche. Un autre modèle tient compte des présupposés inclus dans les énoncés d'un utilisateur afin de mieux adapter les réponses de la machine et de gérer le dialogue.

Par ailleurs, nous cherchons des algorithmes qui exploitent les connaissances de ces modèles pour comprendre les énoncés multimodaux. Ainsi un de nos algorithmes interprète les différentes références incluses dans une tâche, notamment quand celle-ci est visualisée sur un écran, et intègre le fonctionnement des déterminants en français. S'appuyant sur un pavage de l'espace cognitif, par exemple, de l'écran, l'algorithme lève les ambiguïtés d'énoncés, tels «la figure», «cette figure», «une de ces figures». Nos algorithmes sont écrits en Java, le langage en vigueur sur le réseau *Internet*, assurant ainsi leur adéquation aux divers systèmes informatiques.

Enfin, nous avons proposé une nouvelle architecture, modulable et répartie, pour les systèmes de dialogues multimodaux, qui s'adapte d'une part au modèle de l'application à gérer, c'est-à-dire à la définition des objets à traiter, à leurs propriétés et aux fonctions qui leur sont applicables. D'autre part, l'architecture s'adapte aux spécificités lexicale, syntaxique, sémantique et d'usage des différents modes de communication ; ainsi, pour le langage, nous traitons la langue soutenue, technique, écrite et orale.

Bien que nos recherches soient encore fondamentales, des applications industrielles sont déjà en cours. Ainsi, avec la Société *Alcatel*, nous avons mis au point une interface de communication, répondant aux gestes et à la voix, qui équipe un terminal de télécommunication composé d'un écran tactile et d'un combiné téléphonique. Nous mettons également au point un environnement multimodal de consultation d'annuaires complexes pour ce type de terminal de télécommunication qui pourrait bientôt apparaître sur le marché.

À l'avenir, seule une communication multimodale nous permettra d'accéder de façon naturelle aux nombreuses applications interactives qu'offrent aujourd'hui le réseau *Internet* et la télévision numérique.

Jean-Marie PIERREL,
Université de Nancy et Équipe «Langue et Dialogue» du LORIA

transférés directement du *Net 21* sur les *Handy 21* et *Enviro 21*, en fonction des demandes, des erreurs et des mises à jour des utilisateurs. Les utilisateurs adapteront les machines qui les entourent à leurs besoins et à leurs habitudes.

Une conviction et un souhait

Oxygen nécessite des ordinateurs portables, des ordinateurs muraux et des ordinateurs embarqués, un réseau innovant, des systèmes de compréhension de la parole et de synthèse vocale, l'accès aux connaissances, la collaboration, l'automatisation et la personnalisation. *Oxygen* devra sa puissance à l'ensemble de ses éléments, qui sont tous tournés vers l'utilisateur. La communication avec les ordinateurs a commencé avec des nombres binaires. Puis on a utilisé des textes et, récemment, des icônes ; il est temps d'améliorer encore les interfaces homme-machine.

Je crois que l'informatique du futur dépendra essentiellement des échanges vocaux (ou à l'aide d'autres capacités sensorielles), de l'accès aux connaissances, de l'automatisation, de la collaboration et de la personnalisation. Ces cinq aspects, à mon avis incontournables, sont le volant, l'accélérateur et les freins que l'informatique utilisera pour nous en donner plus.

Je propose que tous ceux qui souhaiteront tirer profit du nouveau monde de l'information exploitent les potentialités du système *Oxygen*. Les médecins, par exemple, seraient plus efficaces s'ils disposaient à tout instant des données de *Medline* (qui regroupent des articles médicaux de très nombreuses revues), des dossiers médicaux de leurs patients et des avis de collègues spécialisés.

L'informatique, avec *Oxygen* et d'autres systèmes de même nature, doit maintenant nous libérer des détails techniques que nous supportons depuis 40 ans et nous faire entrer dans une nouvelle ère : celle des ordinateurs au service de l'utilisateur. Après la charrue, le moteur et l'ordinateur, la quatrième révolution concerne la ressource terrestre la plus précieuse : nous-mêmes.

Michael DERTOUZOS est directeur du Laboratoire d'informatique de l'Institut de technologie du Massachusetts.

Conversation avec votre ordinateur

VICTOR ZUE

Comment commander les ordinateurs sans lever le petit doigt.

Les auteurs de science-fiction savent depuis longtemps que le langage parlé est, pour l'homme, le moyen le plus pratique de communiquer avec les machines. Croire qu'ils donnent simplement leur version du mythe de Pygmalion est insuffisant. La parole est surtout naturelle : nous savons parler avant de savoir lire ou écrire. La parole est rapide : nous parlons environ cinq fois plus vite que nous ne tapons à la machine, et peut-être dix fois plus vite que nous n'écrivons. Enfin, le langage parlé est flexible : nous n'avons besoin d'aucun autre support que l'oral pour mener une conversation, les yeux ou les mains sont libres.

La première génération de systèmes intégrant des techniques vocales apparaît aujourd'hui. Certains d'entre eux reconnaissent des dizaines de milliers de mots. Des logiciels de dictée sont déjà fabriqués par plusieurs sociétés : *IBM*, *Dragon Systems*, *Lernout & Hauspie* ou *Philips*. D'autres systèmes, initialement étudiés par les Laboratoires *AT&T Bell*, puis mis au point par les Sociétés *Nuance*, *Philips* et *SpeechWorks*, traitent des discours transmis par téléphone sans étalonnage préalable. Ces techniques, intégrées dans des services d'assistant virtuel, permettent aux utilisateurs d'accéder par téléphone à des informations de presse, à des cotations en Bourse ou au courrier électronique. Cependant, le projet *Oxygen* requiert des systèmes de reconnaissance de la parole plus efficaces.

Avec les systèmes de la nouvelle génération, les hommes communiqueront avec les ordinateurs presque aussi facilement qu'entre eux. Les ordinateurs ne se limiteront pas à comprendre ce qu'on leur dit, mais devront tenir des conversations : les dispositifs classiques de reconnaissance vocale, qui convertissent des signaux sonores en symboles numériques, seront associés à des logiciels de compréhension du langage, afin que l'ordinateur saisisse la signification des mots prononcés.

L'ordinateur sera capable de parler, c'est-à-dire de trouver des documents sur le réseau *Internet*, d'en extraire l'information appropriée et de la présenter sous la forme de phrases correctement construites. À chaque étape du processus, la machine dialoguera avec l'utilisateur pour clarifier certains points et rectifier d'éventuelles erreurs. Il posera des questions : «Vous avez dit Rodez, dans l'Aveyron, ou Rodes, dans les Pyrénées-Orientales ?»

Galaxy vous parle

Nous étudions de tels systèmes depuis dix ans au Laboratoire d'informatique de l'Institut de technologie du Massachusetts. Jusqu'à présent, les machines qui ont vu le jour ont des domaines de compétence très limités, mais elles sont utiles parce qu'elles donnent une information à jour, par téléphone. Par exemple, le système *Jupiter* fournit des prévisions météorologiques de 500 villes à travers le monde (ce service gratuit est disponible au 00 1 881 573 8255) ; le système *Pegasus* donne les horaires de 4 000 vols aux États-Unis ; le système *Voyager* informe sur le trafic routier. Aujourd'hui, ces systèmes parlent l'anglais, l'espagnol et le chinois. Si l'on ne tient pas compte du temps nécessaire à la récupération des informations sur le réseau *Internet*, nos systèmes répondent à la vitesse d'une conversation normale. En France, la *SNCF* met au point le système *ARISE*, qui permettra de réserver des billets de train par téléphone, et le système *MASK*, qui sera intégré dans un kiosque de vente de billets *(voir l'encadré)*.

Nos applications de reconnaissance et de synthèse vocale sont fondées sur une architecture, nommée *Galaxy*, que nous avons construite il y a cinq ans. C'est une architecture répartie, c'est-à-dire que différents calculs sont exécutés par plusieurs ordinateurs. *Galaxy* extrait de l'information de différents domaines

de connaissances pour répondre à la requête d'un ou de plusieurs utilisateurs simultanément. On accède à *Galaxy* par téléphone ou par ordinateur, avec une connexion au réseau *Internet*, qui transmet les données.

Galaxy assure cinq fonctions : la reconnaissance de la parole, la compréhension du langage, l'extraction de l'information, l'élaboration du langage et la synthèse vocale. Quand on pose une question, un serveur nommé *Summit* analyse ce qu'il entend à l'aide d'une bibliothèque de phonèmes (les plus petites unités sonores qui constituent les mots, dans une langue), puis il dresse une liste ordonnée de phrases possibles. Un deuxième serveur, nommé *Tina*, applique les règles grammaticales, décompose la phrase en éléments (sujet, verbe, complément...) et traduit alors l'énoncé le plus plausible en un bloc sémantique, c'est-à-dire une série de commandes que la machine comprend. Par exemple, *Tina* reformulera la question : «Où se trouve le musée d'Orsay ?» en la commande suivante : «Localiser le musée nommé musée d'Orsay.»

Ensuite, un troisième serveur, *Genesis*, convertit le bloc sémantique en une requête adaptée au format de la banque de données la plus pertinente. *Tina* réorganise alors l'information extraite en un nouveau bloc sémantique, que *Genesis* convertit en une phrase dans la langue de l'utilisateur : «Le musée d'Orsay se trouve au numéro 1 de la rue de la Légion d'honneur, à Paris, dans le VII^{e} arrondissement.» Enfin, un synthétiseur vocal dit la phrase.

Pour passer d'un système à l'autre, l'utilisateur n'a qu'à demander : «Je veux parler à *Jupiter*» ou : «Passe-moi *Voyager*.» Depuis mai 1997, *Jupiter* a répondu à plus 30 000 appels, et le taux de compréhension des requêtes d'utilisateurs novices atteint 80 pour cent.

Le projet *MASK*

Dans les halls de gare, des billetteries automatiques à interface tactile sont destinées à faciliter l'accès de la clientèle au système de distribution. Malheureusement, la rigidité et le manque de convivialité de ce système n'ont pas séduit les voyageurs. Depuis 1994, dans le cadre d'un programme de recherche européen, en collaboration avec le CNRS, l'Université de Londres et les Sociétés *Mors* et *Signes particuliers*, la SNCF met au point une borne d'information et de distribution qui, à terme, remplacera les billetteries actuelles.

Il s'agit d'un kiosque de réservation et d'achats de billets de train à interface naturelle multimodale : la machine répond aux paroles ou au toucher de l'utilisateur par une voix de synthèse, des séquences vidéo ou des textes. Le kiosque, nommé *MASK*, est composé de deux ordinateurs : une station de travail *Silicon Graphics* qui gère les différents modules, ou logiciels, du système, et un ordinateur personnel classique, qui se charge de l'interface graphique. Lorsqu'un utilisateur énonce une requête, un premier module de reconnaissance de la parole, qui intègre les hésitations, traduit les signaux acoustiques en signaux numériques. Un deuxième module extrait de ces signaux les informations requises (date de départ, destination, gare de départ...) qu'un troisième module, le gestionnaire de dialogue, recherche dans une banque d'information, nommée *Riho*. Le lexique répertorie aujourd'hui plus de 1 500 mots, dont les noms de 600 villes françaises. Quand des informations sont encore nécessaires, un module de génération de phrase formule une question qu'un synthétiseur vocal prononce. Lorsque la requête est complète, une liste de trains est proposée sur l'écran, et le voyageur n'a plus qu'à choisir. L'achat d'un billet peut durer moins d'une minute.

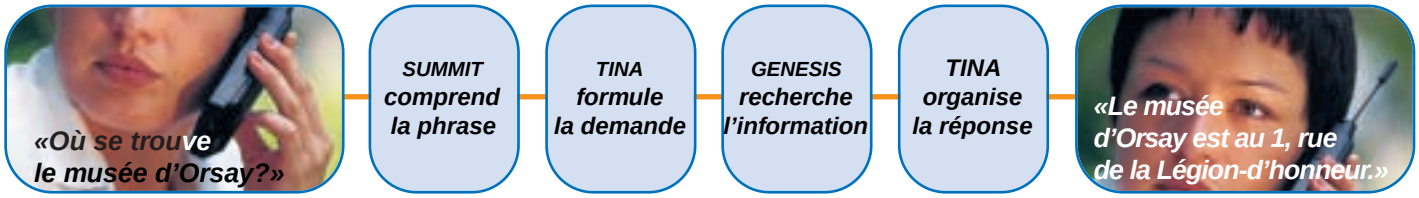
Le kiosque a été testé pendant un mois à la gare Saint-Lazare par des clients volontaires qui suivaient un scénario préétabli. On analyse ces essais préliminaires afin d'évaluer la machine en conditions réelles, et d'améliorer ses capacités de compréhension et de dialogue.

Nous enregistrons les appels, puis nous les analysons, afin d'améliorer les performances du système.

Nos systèmes de reconnaissance de la parole seraient parfaitement adaptés aux appareils portatifs *Handy 21* du projet *Oxygen* : le clavier, notamment, pourrait être supprimé, et le langage parlé permettrait aux utilisateurs de communiquer plus efficacement avec leur appareil. Ainsi, un homme d'affaires en déplacement pourrait demander à son ordinateur : «Dis-moi quand l'action Crédit lyonnais dépassera 210 francs.» La machine deviendra ainsi un assistant qui accomplit des tâches variées à partir d'un minimum d'instructions.

Nous devons encore progresser avant d'atteindre ce but. Notamment, nous devons réaliser des programmes capables de traiter de l'information dans des domaines complexes, nous devons augmenter le nombre de langues comprises par la machine et nous devons obtenir des systèmes qui réunissent des données variées sans y être explicitement invités. L'ordinateur ne doit pas seulement comprendre ce que l'on *dit*, mais plutôt ce que l'on *veut dire*.

Victor ZUE est directeur adjoint du Laboratoire d'informatique de l'Institut de technologie du Massachusetts.



LES APPLICATIONS À RECONNAISSANCE VOCALE utiliseront l'architecture *Galaxy* pour répondre à la question d'un utilisateur. Le système *Summit* établit une liste d'interprétations de la phrase, puis le système *Tina* traduit la proposition la plus plausible en instructions qui permettent au système *Genesis* d'extraire l'information d'une banque de données. Enfin, le système *Tina* formule une réponse.

Tom Draper Design ; photographies de P. Broze, Leo de Wys, Inc.