Compréhension automatique du geste et de la parole spontanés en communication homme-machine : apport de la théorie de la pertinence

Frédéric Landragin, Nadia Bellalem, Laurent Romary
LORIA – UMR 7503
Campus scientifique – BP 239
54506 Vandœuvre-lès-Nancy CEDEX

Abstract

With an approach based on the spontaneous character of man-machine communication, we need to provide multimodal systems with intelligent comprehensive abilities. The design of such systems must rely on theories from social sciences. But these theories have to handle concepts which can be formalized in a computing point of vue. The exemple we present in this paper with Relevance Theory appears to fit well this concern.

1. Introduction

Maintenant que les utilisateurs de systèmes de dialogue homme-machine ne sont plus seulement des informaticiens, il s'avère important que les efforts nécessaires à la réussite de la communication ne reviennent non plus à ces utilisateurs, par des apprentissages et des contraintes d'utilisation, mais à la machine, par de nouvelles capacités de compréhension, d'adaptation, de raisonnement. Le développement informatique de systèmes de dialogue intelligents pose ainsi de nombreux problèmes, faisant intervenir de plus en plus la linguistique, la psychologie, les sciences cognitives en général. Si l'on veut que l'utilisateur puisse communiquer librement face à un écran, il faut le laisser utiliser ce qu'il connaît le mieux, c'est-à-dire ce qu'il utilise tous les jours en communication de face à face : le langage et les gestes conversationnels. Cette approche fondée sur la spontanéité de la communication – et non sur l'état actuel de la technologie – nécessite des algorithmes performants, basés sur des notions et des modèles issus de plusieurs disciplines : linguistique pour un traitement de la langue naturelle faisant intervenir syntaxe, sémantique et pragmatique ; psychologie pour la simulation de capacités telles que la perception du contexte visuel, support de la communication. Lors de l'élaboration de tels algorithmes, il nous semble judicieux d'être guidé plus particulièrement par les études sur le fonctionnement humain, donc de choisir une approche cognitive. L'objectif de ce papier est de faire état des problèmes liés à cette approche pluridisciplinaire, et d'en déduire des pistes pour une implantation informatique.

2. Dispositif d'acquisition et capacités de compréhension

Laisser libres les usages des modalités naturelles que sont la parole et le geste a deux types de conséquences, d'une part sur le choix du matériel d'acquisition, d'autre part sur les capacités de compréhension à simuler en priorité. Pour la parole, le choix du matériel est immédiat (microphones). En revanche, le système doit accepter et pouvoir traiter tout vocabulaire, même si le contexte applicatif réduit généralement de beaucoup le nombre de mots attendus et permet de maximiser les performances sur ceux-ci (Pierrel 1987). De même, toute construction syntaxique doit être traitée, en particulier celles relevant d'un niveau de langage parlé et comprenant hésitations, répétitions, corrections, agrammaticalités (Lopez 1999).

Pour le geste, les deux types de conséquences posent des problèmes. Si l'on souhaite une communication parfaitement naturelle, aucun dispositif intrusif ne doit être imposé à l'utilisateur, pas même un gant numérique. Il ne reste plus que le système de vision artificielle par caméras, lourd à mettre en œuvre. De plus, tout geste doit a priori être traité : les co-verbaux (référentiels, expressifs ou paraverbaux) aussi bien que les quasi-linguistiques ou les synchronisateurs (dans la suite nous nous baserons sur la classification des gestes de (Cosnier & Vaysse 1997)). Or les exploiter tous s'avère impossible, du fait de l'absence de standardisation formelle du geste, de sa grande sensibilité au contexte, et de l'impossibilité de segmenter le signal gestuel et d'identifier ses unités élémentaires (Cosnier & Vaysse 1997). En fait, pour le geste, les deux types de conséquences interagissent : il s'agit de trouver un dispositif qui permette une communication spontanée tout en limitant de lui-même l'éventail des gestes réalisables. Si le seul fait que l'interlocuteur soit une machine permet de supposer que certains types de gestes ne seront pas employés (synchronisateurs car la régulation de la conversation pose moins de problèmes avec une machine qui n'est pas sensée couper la parole ; expressifs car l'utilisateur sait bien que la machine n'y sera pas sensible), d'autres hypothèses sont nécessaires.

La théorie de la pertinence (Sperber & Wilson 1995), qui ne s'est pas intéressée spécifiquement au geste mais aux processus communicationnels en général, nous semble apporter un cadre théorique intéressant pour notre approche, d'une part parce qu'elle s'insère bien dans nos préoccupations cognitives, d'autre part parce que les hypothèses sur lesquelles elle repose et les concepts qu'elle manipule nous semblent formalisables du point de vue informatique. Ainsi, l'hypothèse de départ étant que le communicateur choisit le stimulus ostensif le plus pertinent de tous ceux qu'il peut utiliser, on suppose que l'utilisateur d'un système de dialogue homme-machine multimodal choisit les énoncés les plus pertinents dans le contexte courant. Un geste pertinent est un geste qui, avant tout, apporte de l'information utile à l'interlocuteur, ce qui élimine les paraverbaux, qui ne portent que peu d'informations et sont plus utiles à la production qu'à la compréhension. D'autre part, nous choisissons l'écran tactile comme dispositif d'acquisition, ce qui permet d'éliminer les quasi-linguistiques dont la forme ne se prête pas à une représentation sur une surface plane. Ne restent que les gestes référentiels, qui apportent effectivement de l'information nécessaire à la compréhension et qui sont spontanément réalisables sur écran tactile. De plus, les conditions d'usage, c'est-à-dire le contact nécessaire entre l'écran et le doigt ou le stylet, font que le système ne reçoit que la partie significative d'un geste, épuré de ses phases d'approche et de retrait.

3. Pertinence et référence

Ces choix mettent l'accent sur un point important de la communication, particulièrement présent en dialogue homme-machine avec l'écran comme support : la référence aux objets. Les expressions référentielles langagières possibles sont multiples, les gestes également (pointage, entourages, etc.), ce qui amène à une explosion de la combinatoire des expressions référentielles multimodales. La théorie de la pertinence nous donne un critère pour s'y retrouver, à travers les concepts que sont les effets contextuels et l'effort de traitement.

Les effets contextuels regroupent les conséquences de l'interprétation d'un énoncé sur le contexte, en considérant que le contexte se compose d'hypothèses caractérisées chacune par une certaine force. L'effacement de certaines hypothèses du contexte, la modification de leur force et la dérivation de

conclusions nouvelles sont les différents effets contextuels tels que définis par (Sperber & Wilson 1995). Ils dépendent de l'énoncé complet et ne s'appliquent pas au niveau des expressions référentielles. Les seuls effets contextuels de celles-ci sont : le fait nouveau qu'il y a eu référence, et le résultat de cette référence, c'est-à-dire les objets sur lesquels elle a porté.

L'effort de traitement dépend tout d'abord de la complexité de l'énoncé : plus la partie orale fait intervenir de traits sémantiques, plus il est élevé ; plus la trajectoire gestuelle est longue, plus il est élevé. Intervient de plus l'effort nécessaire pour associer geste et parole : plus les indices (synchronisation temporelle, adéquation sémantique) permettant de faire les liens sont faibles, plus il est élevé. L'effort dépend également de la complexité du contexte et de l'accessibilité des informations dans ce contexte. En dialogue multimodal, le contexte regroupe des informations provenant de deux sources principales : l'historique du dialogue et l'état de la scène affichée à l'écran à l'instant de la référence. Une méthode d'évaluation consiste à pondérer la complexité de l'énoncé par le taux de présence dans le contexte des informations requises : si par exemple l'énoncé oral fait intervenir un filtrage sur une propriété P, l'effort de traitement sera d'autant plus élevé que le contexte perceptif contiendra d'objets vérifiant P.

4. Perspectives

Dans (Landragin et al.), des scores d'évaluation de ces notions ont été proposés dans le cadre d'une application précise, et souffrent justement de cette spécificité. Notre objectif est maintenant de les généraliser, non seulement en compréhension pour détecter l'éventuelle incongruïté d'un énoncé, en génération pour choisir parmi plusieurs possibilités, mais également dans des manipulations de simulation comme indicateurs ou critères de validation.

Références bibliographiques

Cosnier J. & Vaysse J. 1997, Sémiotique des gestes communicatifs, Nouveaux Actes Sémiotiques, 53.
Landragin F., De Angeli A., Wolff F., Lopez P. & Romary L. (in press), Relevance and Perceptual Constraints in Multimodal Referring Actions, in: Van Deemter K. & Kibble R. (Eds.) Information Sharing: Givenness and Newness in Language Processing, CSLI Publications.
Lopez P. 1999, Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres, Thèse de l'Université Henri Poincaré, Nancy.
Pierrel J.-M. 1987, Dialogue oral homme-machine, Paris, Hermès.
Sperber D. & Wilson D. 1995, Relevance: Communication and Cognition (2nd edition), Oxford, Blackwell.
Wolff F. 1999, Analyse contextuelle des gestes de désignation en dialogue homme-machine, Thèse de l'Université Henri Poincaré, Nancy.