Politiques d'action : de la validation psychologique et linguistique à la programmation d'agents cognitifs en intelligence artificielle

Andreas Herzig, Jérôme Lang

Politiques d'action : de la validation psychologique et linguistique à la programmation d'agents cognitifs en intelligence artificielle

Responsable scientifique : A. HERZIG & J. LANG

Andreas HERZIG & Jérôme LANG
IRIT (UMR 5505)
118 route de Narbonne
31062 Toulouse Cedex 4
Tél. 05 61 55 63 44 ;
E-mail : {herzig, lang}@irit. fr

Équipes partenaires

- Laboratoire Travail et Cognition (LTC), Université Toulouse II, UMR 5551
- Équipe de Recherche en Syntaxe et Sémantique (ERSS), Université Toulouse II, UMR 5610
- Centre d'épistémologie et d'ergologie comparatives (CEPERC), Université Aix-Marseille, ESA 6059

Rappel des enjeux et objectifs

Les nouvelles technologies de l'information et de la communication rendent cruciale la construction d'artefacts capables d'agir de manière flexible et autonome dans des environnements dynamiques et partiellement imprédictibles. La spécificité de ces artefacts de nouvelle génération réside dans leur capacité à s'adapter au monde extérieur (comprenant aussi bien le monde « physique » que les croyances et intentions des autres agents), en fonction des actions entreprises et des observations effectuées.

L'objectif de ce projet est la conception de modèles et de langages pour la programmation d'agents cognitifs complexes. Ces agents doivent être en mesure de poursuivre un but en planifiant des actions de façon autonome, d'acquérir des informations au moyen d'actions épistémiques et de communiquer avec des interlocuteurs (humains ou machines) au moyen d'actions de communication qu'il s'agira de produire aussi bien que d'interpréter. Un agent cognitif est naturellement destiné à être interfacé d'une manière ou d'une autre avec des opérateurs humains : il s'agit alors d'interpréter de la façon la plus fiable possible les croyances, intentions, préférences et buts de l'utilisateur, ce qui nécessite que les modèles et les langages manipulés par l'agent cognitif soient suffisamment proches des raisonnements humains et du langage naturel, et puissent donc être validés d'un point de vue psychologique et linguistique.

Rappel du calendrier des travaux

- Étude critique des langages de représentation d'actions ; formalisation logique des langages d'action à l'œuvre dans l'architecture cognitive ACT-R (année 1).
- Liens formels entre actions épistémiques et révision des croyances (année 1).
- Formalisation logique des actions de communication et intégration dans le processus de planification (année 1).
- Développement d'un petit ensemble d'acteurs humains simulés individuellement dans l'architecture ACT-R. Conception d'un recueil expérimental permettant de fournir des données sur la gestion du conflit entre stratégies de prise d'information et stratégies d'action immédiate (année 1).
- Simulation de l'interaction entre les différents opérateurs modélisés et comparaison des interactions émergeant du groupe d'agents simulés et des interactions observées chez les individus humains. Portage du modèle à une situation de travail réelle : la médecine d'urgence (année 2).

N.B. : Par rapport à notre proposition d'origine structurée en 4 axes, Cognitique nous avait demandé de limiter le projet à l'axe 3 « La pertinence dans le choix et l'interprétation d'actions ». Le présent calendrier en tient compte, et correspond à un recentrage du programme d'origine dans ce sens, que nous avons effectué suite à l'acceptation du projet.

État d'avancement à mi-parcours (septembre 2002)

Le projet d'origine ayant été limité à l'axe 3 « La pertinence dans le choix et l'interprétation d'actions », nous nous sommes focalisés en conséquence sur les deux composantes de cet axe, à savoir d'une part « la pertinence dans le choix d'actions », et d'autre part « la pertinence dans l'interprétation d'actions de communication ».

Les recherches ont été menées jusqu'à présent principalement sur quatre axes : philosophico-linguistique, logique, informatique et psychologique.

En ce qui concerne l'aspect philosophico-linguistique, nous nous sommes focalisés sur l'interprétation des actes de langage. Nous nous sommes basés sur une étude qui a été menée par Jacques Virbel (pragmatique linguistique, IRIT) sur la communication non littérale afin de recenser les différentes manières d'accomplir des actes illocutoires indirects (par exemple, poser une question comme « Peux-tu me passer le sel ? » peut parfois être interprété comme un ordre du type « Passe-moi le sel »). Parallèlement à cela, une étude des travaux de Grice (maximes de conversation et principe de pertinence) a été menée afin de comprendre le mécanisme de compréhension de ces actes illocutoires indirects. Les travaux de Grice ne portant en fait que sur les actes illocutoires de type assertif, il s'avère qu'il est indispensable de considérer les travaux de Daniel Vanderveken (philosophe du langage, Université du Québec à Trois-Rivières) sur une généralisation des maximes de qualité et de quantité à tous les types d'actes illocutoires (assertifs, directifs, engageants, déclaratifs, et expressifs). Lors de son séjour de quinze jours à Toulouse en juillet dernier, nous avons eu l'occasion de discuter de ces aspects de façon approfondie.

Du point de vue logique, nous avons exploité les travaux de Virbel afin de définir la notion de *forme d'indirection*, en correspondance avec la notion syntaxique qu'il donne. En isolant et en représentant au sein des croyances d'un agent (décrit au sein d'une logique modale de la croyance, de l'intention, et de l'action) certains *faits pertinents du contexte*, nous avons montré comment écrire des lois d'actions (décrivant les effets et les préconditions des actions) conditionnelles au contexte d'énonciation. En collaboration avec Maud Champagne (du Laboratoire de Neuropsychologie Jacques Lordat), nous avons ainsi obtenu une manière d'inférer des actes illocutoires non littéraux sans « traiter » l'acte littéral *a priori*, c'est-à-dire sans en dériver les effets et les préconditions. Ce modèle répond donc à des exigences issues des connaissances actuelles en neurosciences du langage tendant à montrer que le sens non littéral n'est pas issu d'un traitement du littéral. Comme étape suivante à l'interprétation proprement dite d'un énoncé, nous avons également établi dans notre logique un certain nombre d'axiomes de coopération en vue d'adopter les croyances et intentions d'autrui, le but étant d'arriver à les anticiper. Les résultats vont paraître dans (Champagne et al., à paraître).

Du point de vue psychologique. Afin d'établir une collaboration étroite entre informaticiens et psychologues, Dominique Longin (informaticien de formation) a été détaché au Laboratoire Travail et Cognition (UMR 5551) en tant que post-doctorant depuis le 1er février 2002 (retard dû à des problèmes administratifs). Dans un premier temps, afin d'étudier l'architecture cognitive ACT-R/PM, nous avons élaboré un modèle ACT-R de l'effet Stroop ayant déjà fait l'objet d'une expérimentation sur des sujets. Par elle-même l'étude de cet effet ne rentre pas dans le cadre du projet mais la disponibilité de ces données expérimentales au LTC a rendu possible la phase initiale d'acquisition du savoir-faire en modélisation dans l'environnement d'ACT-R. Ce savoir-faire permettra de répondre à la fois aux exigences de plausibilité psychologique en modélisation cognitive et aux exigences de formalisation de l'intelligence artificielle. Les résultats de ces travaux seront reportés dans (Longin et Raufaste, en préparation). Dans un second temps, nous avons élaboré un modèle ACT-R empirique de l'interprétation d'actes de langage indirects de type directif. Les stimuli ayant servi à valider le modèle sont issus de ceux utilisés par Maud Champagne lors de sa thèse en neuropsycholinguistique. L'intérêt de ce modèle ACT-R est qu'il constitue une alternative aux modèles normatifs issus de la linguistique pragmatique et de la philosophie du langage (et, par voie de conséquence, aux formalismes logiques en découlant). En implémentant dans ACT-R une théorie psychologique du traitement des actes indirects, on tend à s'assurer du réalisme psychologique de cette théorie et de son bien-fondé.

Sur le plan des langages d'action, nous avons travaillé sur la représentation logique des préférences d'un agent, ainsi que sur la formalisation logique des actions de communication et leur intégration dans le processus de planification d'actions. En ce qui concerne le dernier point, nous avons étudié le délicat problème de l'intégration de la révision des croyances. Une telle révision est nécessaire lorsqu'un agent apprend des faits (soit par une action épistémique, soit par la communication avec un autre agent) qui sont en contradiction avec ses croyances antérieures. Après avoir mis en évidence les problèmes que rencontrent les propositions précédentes, nous avons fait une proposition originale (Herzig, Lang et Longin, 2002 ; Herzig et Longin, 2002 ; Herzig et Longin, 2002a). Ces travaux s'inscrivent dans une perspective où le concept de révision est rapproché de celui de mise à jour, suivant en cela à des travaux avec des philosophes sur les différences et similarités entre ces deux concepts (Crocco et Herzig, à paraître). En ce qui concerne les préférences des agents, les critères qui sont sous-jacents au choix d'une action par un agent n'ont de sens que par rapport à ses buts, ses préférences. Afin de pouvoir construire des modèles logiques du choix d'une action par un agent, il faut donc se donner un langage logique de représentation de préférences, dans lequel les objectifs de l'agent sont exprimés de manière concise et intuitivement satisfaisante. Cette problématique est importante tant en ce qui concerne la programmation d'agents autonomes artificiels (à qui il faut « transmettre » les objectifs de l'agent humain pour lequel il travaille) qu'en ce qui concerne l'interprétation automatisée d'actions entreprises par un agent humain. Ce sujet a fait l'objet d'une publication récente (Lang, van der Torre et Weydert, 2002).

Programme de travail prévu pour l'année 2003

Du coté de la linguistique, nous pensons établir un certain nombre d'énoncés contenant des actes non littéraux en vue de disposer de stimuli pour une expérimentation sur des sujets (aspect psychologie). Les énoncés devraient couvrir au moins trois (si ce n'est quatre) catégories d'actes indirects (assertifs, directifs, engageants, et peut-être déclaratifs), divers violations de maxime, ainsi que divers cas d'ironie. Ces stimuli devront être élaborés de concert avec les psychologues, afin de ne pas introduire de biais dans l'expérimentation.

Les premiers essais de modélisation limités aux actes indirects ont révélé deux problèmes que la prochaine tranche devra traiter : d'une part la question de l'ordre de traitement des actes directs et indirects, d'autre part la question du caractère séquentiel ou « on-line » de ces traitements. Si l'on adopte une approche séquentielle, deux options de modélisation sont possibles : l'interprétation directe peut être construite d'abord, suivie de la recherche d'interprétations indirectes, ou bien l'inverse (auquel cas, l'interprétation directe peut apparaître comme le résidu d'une interprétation indirecte qui n'a pas abouti). Par souci de simplicité dans la démarche nous avons construit notre modèle d'interprétation des directifs indirects selon ce dernier abord. Une troisième voie, l'approche « on-line » consiste à faire mener simultanément la recherche d'interprétations directes et indirectes au fur et à mesure de l'acquisition des lexèmes. De très récents travaux empiriques en psychologie suggérant une plus grande plausibilité de la troisième voie, notre tâche immédiate sera de reformuler le modèle actuel en termes de sélection dynamique des interprétations pertinentes (directes ou non). Dans un second temps, nous étendrons le modèle aux autres cas d'énoncés non littéraux.

Du point de vue psychologique, une expérimentation sur des sujets humains fournira des données (en terme de type et de temps de réponse, et de pattern d'erreur) qui pourront être comparées à celles obtenues par le modèle ACT-R afin d'en régler les paramètres le plus finement possible. L'approche expérimentale visera aussi à apporter des informations quant aux choix de modélisation vus plus haut. À terme, nous espérons modéliser plusieurs agents sous ACT-R afin de les faire communiquer entre eux et observer leurs comportements respectifs, ce qui devrait conduire à une extension naturelle de nos travaux vers le dialogue entre agents.

Du point de vue logique, il semble que le modèle psychologique du traitement des actes non littéraux implémenté dans ACT-R puisse fournir la base à l'élaboration d'une logique de type autoépistémique. Idéalement, cette logique devrait permettre de décrire les mécanismes mis en œuvre dans ACT-R, afin d'en donner une caractérisation formelle. Nous pensons également établir une comparaison entre nos précédents travaux formels (utilisant des résultats issus de la pragmatique linguistique et de la philosophie du langage) et cette nouvelle logique.

Du point de vue des langages d'action, nous travaillons actuellement sur la prise en compte de la notion de pertinence d'actions par rapport à un but dans un contexte de décision séquentielle (planification) : informellement, une action A est pertinente (respectivement, immédiatement pertinente), pour un agent donné (avec ses buts, ses croyances initiales et l'ensemble des actions qui lui sont disponibles) s'il existe un plan minimal (au sens d'un critère donné - par exemple le coût, ou tout simplement la longueur) pour atteindre les buts, et qui contient A (respectivement dont la première action est A). Cette notion est intéressante à deux points de vue :

- dans un contexte de planification automatisée, un prétraitement (hors-ligne) du problème visant à identifier des relations de pertinence entre actions et objectifs, qui permettront de résoudre plus rapidement le problème une fois que toutes les données seront connues (la pertinence est vue ici comme une heuristique de choix d'action ; c'est pourquoi il sera intéressant de proposer un raffinement de la définition précédente en rendant graduelle la notion de pertinence) ;
- dans un contexte d'interprétation d'action, l'information qu'un agent a entrepris une certaine action, qui, s'il est rationnel, doit être pertinente par rapport à ses objectifs, permet d'inférer de nouvelles connaissances sur ses croyances, ses buts et ses intentions. Cette problématique est importante dans un contexte de communication homme-machine et/ou de coopération dans des environnements multi-agents.

Publications issues du projet

M. Champagne, A. Herzig, D. Longin, J.-L. Nespoulous, et J. Virbel (à paraître). *Formalisation pluridisciplinaire de l'inférence de certains types d'actes de langage non littéraux*. Revue I3, à paraître.

Andreas Herzig and Gabriella Crocco. *Les opérations de changement basées sur le test de Ramsey.* In Pierre Livet, editor, Revision. Hermès, à paraître.

A. Herzig et D. Longin (2002). *A logic of intention with cooperation principles and with assertive speech acts as communication primitives.* In C. Castelfranchi et W. L. Johnson (éditeurs), Proc. 1st Int. Joint Conf. on Autonomous Agents and Multi-Agent System (AAMAS 2002). ACM Press. pp. 920-927

J. Lang, L. van der Torre et E. Weydert (2002). *Utilitarian desires. Autonomous Agents and Multi-Agent Systems*, 5, 329-363, 2002 Kluwer Academic Publishers.

D. Longin, E. Raufaste et C. Lemercier (en préparation). Modélisation ACT-R d'une nouvelle interprétation de l'effet Stroop.