
MODÉLISATION MULTI-AGENTS DES ÉCHANGES LANGAGIERS : APPLICATION AU PROBLÈME DE LA RÉFÉRENCE ET À SON ÉVALUATION

Andrei POPESCU-BELIS

Groupe Langage et Cognition

tel.: 33 1 69 85 80 58

fax.: 33 1 69 85 80 88

BP 133, F-91403 ORSAY Cedex, FRANCE

Mel : popescu@limsi.fr

<http://www.limsi.fr/Individu/popescu>

Le traitement automatique des langues utilise souvent les techniques symboliques de l'intelligence artificielle, bien que différentes extensions ou solutions de remplacement aient été suggérées. Cette thèse propose de réaliser des modèles multi-agents de l'interaction linguistique, dans un cadre appelé " pragmatique simulée ".

Pour reproduire le caractère situé du langage, il est essentiel d'observer que celui-ci dépend d'un support biologique individuel, ainsi que d'un environnement. Le langage possède également une fonction et une évolution collectives. À ce changement de perspective par rapport à l'approche symbolique correspondent ici des propositions concrètes de modélisation, provenant d'un examen raisonné des bases naturelles du langage et de son fonctionnement. En particulier, il faut souligner l'importance du phénomène de la référence, à savoir le lien réalisé à travers le langage entre les entités de l'environnement et les représentations que possèdent les locuteurs.

L'argumentation et la description des réalisations s'organisent en trois grandes parties. La première partie met en place le paradigme de la " pragmatique simulée " et décrit deux réalisations. Tout d'abord, plusieurs tentatives de dépassement de l'approche symbolique en traitement de la langue sont présentées : les modèles statistiques, les modèles continus, le connexionnisme symbolique. Ces propositions dessinent un aperçu cohérent des points susceptibles d'être améliorés dans le traitement automatique de la langue, points auxquels la " pragmatique simulée " tente aussi de répondre.

L'étude des sources d'inspiration naturelles est développée selon deux directions clé : l'ancrage biologique du langage dans l'individu, et sa dimension collective. Sur ce plan, trois cas d'émergence du langage dans le monde naturel sont mis en avant : l'apparition du langage dans l'espèce humaine, son acquisition par l'enfant, et la situation de créolisation (selon la description de D. Bickerton). Les modèles informatiques de ces phénomènes montrent la difficulté qu'il y a, au niveau individuel, à modéliser des agents qui soient à la fois autonomes et doués de capacités linguistiques ; d'ailleurs, celles-ci s'avèrent actuellement rudimentaires. Les agents inspirés par la théorie de G. M. Edelman, et réalisés dans son équipe, représentent une direction prometteuse, mais en pratique encore éloignée des aspects communicatifs. Les modélisations des échanges langagiers à l'aide de sociétés d'agents utilisent des agents plus simples que les précédents. Dans cette direction de recherche relativement récente, plusieurs simulations abordent l'évolution d'un langage, sous un aspect lexical, syntaxique ou sémantique.

Les principes tirés du passage en revue de ces nombreux domaines conduisent à une esquisse de la " pragmatique simulée " dans un modèle multi-agent. Des contraintes sur les perceptions, actions et valeurs des agents sont ainsi énoncées, puis un schéma d'émergence d'un " langage " entre agents est proposé, en trois stades : actes de langage simples (innés), communication asyntaxique mais utilisant la référence aux objets, passage à la communication syntaxique référentielle grâce à un modèle de créolisation. Une analyse des microstructures possibles des agents vient clore ces développements théoriques.

Deux modèles informatiques réalisés comme éléments de base de notre modélisation par agents sont ensuite détaillés. Au niveau individuel, le premier modèle adapte et implémente des idées neurobiologiques de G. M. Edelman dans un agent adaptatif. Sont notamment étudiés la représentation de valeurs, les protocoles de connexions réentrants et la programmation à base d'acteurs. Cela représente un premier pas vers la conception d'un support structurel réaliste du langage dans un agent.

Au niveau collectif, la simulation d'une société d'agents purement communicants montre l'intérêt d'utiliser une grammaire d'arbres adjoints lexicalisée. Celle-ci permet aux agents d'établir de façon incrémentale des conventions lexicales et syntaxiques pour dialoguer à propos des situations qu'on leur présente. Toutefois, ces deux réalisations n'abordent pas de façon réaliste le lien entre le code de communication et les entités qu'il peut dénoter ; ce lien essentiel est toutefois rarement modélisé pour les agents communicants.

La deuxième partie de la thèse étudie donc la question de l'ancrage du langage dans l'environnement par le biais de la référence linguistique, c'est-à-dire le lien de désignation, à travers les expressions du langage, entre les représentations des locuteurs et les entités du monde de référence.

Une vision d'ensemble de ce vaste domaine est d'abord donnée, allant des études philosophiques de la référence, en passant par les théories linguistiques et les modèles computationnels, jusqu'aux systèmes qui tentent de la traiter. Ce panorama indique la nécessité, dans le cadre de la " pragmatique simulée ", de s'orienter vers l'exploitation des études du fonctionnement concret de la référence chez les humains. Parmi celles-ci, sont exploitées ici les études de la capacité cognitive à distinguer des objets, puis à les catégoriser, ainsi que certains travaux de linguistique fonctionnaliste. À ceux-ci viennent s'ajouter deux expériences originales sur l'usage des articles en français, la première montrant une certaine variabilité de cet usage, et la seconde aboutissant à l'identification de onze " cas référentiels ".

Un modèle de la capacité référentielle d'un agent est alors proposé, qui, éloigné des théories logicistes de la référence, se fonde sur des structures perceptives compatibles avec la théorie d'Edelman. Ce modèle permet de relier l'expression linguistique des cas référentiels aux états perceptifs correspondants, par le biais de l'activation différenciée des différentes représentations.

Afin d'explorer plus avant le phénomène de la référence, cette fois-ci dans le cas de textes réels, un modèle plus représentationniste et symbolique est proposé. Ce modèle est compatible avec le point de vue de la pragmatique, et il adapte en vue d'une implémentation robuste la notion de " représentation mentale " (RM) des référents. Deux opérations indispensables à la construction des RM d'un texte sont retenues : la création d'une RM à partir d'une expression référentielle (ER) ou le rattachement d'une ER à une RM existante. L'implémentation du modèle aboutit à un " résolveur " de références dans les textes, qui utilise des contraintes de sélection et un mécanisme d'activation des référents pour décider de l'opération à effectuer afin de traiter chaque ER. Différents outils ont été réalisés pour cet " atelier de traitement de la référence ", notamment ceux permettant la réalisation, la manipulation et la conversion des ressources linguistiques, à savoir des textes où sont marquées les RM correctes, ou celles produites par un programme.

L'évaluation du résolveur, conçue comme la comparaison des RM correctes avec celles trouvées par celui-ci, nécessite une ou plusieurs mesures numériques de qualité (analysées plus loin). Les performances du résolveur sont analogues à celles des systèmes évalués dans les campagnes officielles MUC (Conférences sur la compréhension de messages). Les conditions expérimentales étaient ici quelque peu différentes, et les connaissances dont dispose le résolveur étaient relativement réduites.

La suppression de l'un ou l'autre des deux mécanismes du résolveur, ou bien la recherche des seuls liens de coréférence entre ER, diminuent nettement ses performances. L'analyse des scores permet d'identifier la taille optimale de la " mémoire à court terme " du résolveur, et d'améliorer le résolveur en optimisant les paramètres d'activation. On montre aussi, grâce à une façon de mesurer la pertinence de chaque règle, que la règle la plus importante est la compatibilité sémantique entre les RM et les ER. Par ailleurs, le caractère défini ou indéfini d'un groupe nominal ne permet pas de déterminer automatiquement l'opération à effectuer sur l'ER – ce qui confirme la nécessité des cas référentiels plus fins exposés auparavant.

La troisième partie étudie d'abord le problème de l'évaluation des systèmes de génie linguistique, et propose un cadre générique compatible avec plusieurs tâches et campagnes d'évaluation passées en revue. Ce modèle définit une évaluation en trois phases (mesure, appréciation, bilan) et suppose que l'on soit capable de mesurer pour chaque donnée la qualité de la réponse du système. Un cas fréquent est celui où l'on mesure une distance entre la réponse du système et l'ensemble clé, ou ensemble des réponses correctes. Plusieurs critères de cohérence sur les mesures sont alors exprimés, qui demandent que celles-ci parcourent de façon uniforme toute l'échelle des scores.

La compréhension des références d'un texte peut être évaluée, dans ce cadre, comme la mesure d'une distance entre deux partitions de l'ensemble des ER d'un texte. Ainsi posé, ce problème n'admet pas de solution canonique ; les mesures existantes sont examinées (mesures dites MUC, B3 et k) à l'aide des critères de cohérence précédents. Pour répondre à certaines limitations de ces mesures, plusieurs nouvelles mesures sont proposées : deux fondées sur le concept de " noyaux " (les embryons de RM trouvées par le programme), une mesure distributionnelle, et une mesure informationnelle. Cette dernière adapte le modèle classique du canal de communication pour définir la notion d' " information référentielle ", et mesurer la corrélation entre clé et réponse (source et récepteur) à l'aide de l'information référentielle perdue dans le canal, et respectivement des gains d'information injustifiés. Cette dernière mesure paraît donc aussi la plus adaptée à la communication entre agents.

Les mesures proposées satisfont certains des critères de cohérence, mais une utilisation plus fiable de ces mesures consiste à tenir compte de la variation simultanée de plusieurs d'entre elles, lorsque l'on souhaite comparer deux programmes ou deux états d'un programme.

La conclusion évoque plusieurs perspectives dans le cadre de recherche mis en place. Il faut tout d'abord réaliser l'intégration des deux expériences réalisées, l'une au niveau d'un agent, l'autre au niveau de la collectivité, et en particulier implémenter le modèle de la capacité référentielle d'un agent. Sur un plan différent, le résolveur de références dans les textes devra être progressivement enrichi de connaissances grammaticales – et ce sans perte de robustesse – puis intégré à d'autres applications. Enfin, l'évaluation devra être étendue aux modèles multi-agents de la communication langagière.