Amorçage d'une sémantique lexicale dans une population d'agents autonomes, ancrés et situés

Luc Steels (1,2) et Frédéric Kaplan (1,3)

(1) Sony CSL Paris - 6 rue Amyot, 75005 Paris
(2) VUB AI Laboratory - Pleinlaan 2, 1050 Brussels
(3) LIP6 - OASIS - UPMC - 4, place Jussieu F-75252 Paris
steels@arti.vub.ac.be, kaplan@csl.sony.fr
Web site: talking-heads.csl.sony.fr

Résumé

Cet article décrit l'amorçage d'une ontologie et d'un lexique partagé dans une population d'agents robotiques dotés de capacités visuelles. Cette évolution a lieu alors que les agents jouent un jeu de langage, appelé "guessing game". Nous étudions les dynamiques d'un tel système et montrons, en particulier, comment la synonymie et l'ambiguïté du système sémantique, qui émergent dans un premier temps, sont progressivement réduites au fur et à mesure que l'environnement physique se complexifie.

1. Introduction

Durant les dernières décennies, la linguistique computationnelle a atteint un haut degré de sophistication dans les méthodes qu'elle met en oeuvre, migrant peu à peu vers des applications industrielles. Néanmoins certains enjeux de base dans le traitement du langage naturel restent à ce jour des problèmes ouverts, dans la mesure où très peu d'outils théoriques ou pratiques existent pour les traiter. Deux de ces problèmes sont étudiés dans cet article.

La plupart des systèmes du traitement automatique du langage naturel sont des systèmes formels symboliques non ancrés dans la réalité par des senseurs et des actuateurs. Ceci limite nécessairement leur degré de compréhension du langage naturel et donc leurs possibles applications. Par exemple, beaucoup de concepts ancrés perceptuellement sont différents d'une langue à l'autre. Ceci rend la traduction très difficile quand le sens ne peut être relié à une expérience sensorielle. De la même manière, beaucoup de phrases produites dans un certain contexte ne peuvent être vraiment désambiguïsées que si le contexte peut être perçu, interprété et intégré dans le processus de compréhension du langage.

La plupart des systèmes de traitement automatique du langage naturel suppose que le langage est fixe. Le lexique et la grammaire sont précodés (souvent à la main) et ne sont pas supposés changer au cours des conversations. Pourtant, les conventions linguistiques sont en flux constant. Locuteur et interlocuteur transforment leur langage en fonction du succès qu'ils obtien-

FIG. 1 – *Deux talking heads et leur moniteur associé montrant ce que chaque caméra perçoit.*

nent lorsqu'ils communiquent ou des informations extra-linguistiques complémentaires quand leurs systèmes linguistiques sont incompatibles ou incomplets. Ceci suggère que les locuteurs humains ont des stratégies de réparation de leur connaissance linguistique en cas d'échec en communication et des stratégies de construction pour entendre l'étendue de leur langage quand cela est nécessaire.

Nous avons examiné ces deux problématiques en utilisant de la modélisation formelle, des simulations informatiques et des expériences avec des agents robotiques. Nous explorons ainsi une démarche alternative jusqu'ici peu pratiquée en traitement automatique du langage naturel. Nous étudions les dialogues en langage naturel sous la forme de *jeux de langage adaptatifs* au cours desquels locuteur et interlocuteur conversent au sujet de scènes du monde réel perçues au travers de leur appareil sensoriel. Dans nos expériences, des groupes d'agents autonomes commencent sans connaissance linguistique et doivent amorcer un langage et un système de concepts pour décrire le monde qu'ils percoivent, sans intervention humaine. Jusqu'à présent, nous avons étudié principalement la sémantique lexicale, mais de premiers résultats dans le domaine de la syntaxe (Steels, 1998) ont également été obtenus.

Cet article décrit nos expériences sur la formation d'un lexique. Il présente succintement les "Talking Heads", la plateforme expérimentale avec laquelle nous travaillons, et les stratégies utilisées par les agents pour construire leur langage. Nous montrons que des dynamiques sémiotiques très riches apparaissent avec l'élimination de la synonymie pure et des phénomènes de désambiguïsation.

2. L'expérience des "Talking Heads"

La plateforme robotique utilisée dans les expériences décrites dans cet article consiste en un ensemble de "Talking Heads" connectées par l'Internet. Chaque Talking Head est constituée d'une caméra SONY EVI-D31 pouvant bouger verticalement et horizontalement (figure 1), d'un ordinateur qui implémente les fonctions cognitives (perception, catégorisation, recherche dans le lexique, etc.), d'un écran présentant les états internes de l'agent chargé dans le corps de la "Talking Head", d'un moniteur montrant la scène telle que la voit la caméra et des entrées et sorties audio. Un agent, chargé dans le corps "physique" d'une Talking Head, peut se téléporter dans une autre Talking Head connectée par l'Internet. Deux agents ne peuvent interagir ensemble que dans le cas où ils sont instanciés physiquement dans deux corps partageant un même environnement. Dans les expériences présentées dans cet article, l'environnement partagé consiste en un tableau magnétique sur lequel différentes formes géométriques sont placées: des triangles, des cercles et des rectangles de différentes couleurs. La simplicité de cet environnment permet une étude plus facile des dynamiques complexes dans la population d'agents.

Le "guessing game" Les agents interagissent dans le cadre d'un jeu de langage nommé "guessing game". Un "guessing game" peut être joué par deux agents dotés de capacités visuelles. Un agent joue le rôle de *locuteur* et l'autre, celui d'*interlocuteur*. Les agents alternent dans ces deux rôles et tous développent la capacité d'être locuteur ou interlocuteur. Les agents sont capables de ségmenter en "objets" l'image perçue par la caméra et de collecter un ensemble d'informations sensorielles concernant chacun de ces objets: couleurs (canaux R G B), niveau moyen de gris (canal GRAY), position (canaux HPOS et VPOS). L'ensemble des objets et les données les concernant constitue le *contexte*. Le locuteur choisit un objet dans ce contexte : c'est le *sujet* de l'interaction.

Le locuteur donne un indice linguistique à l'interlocuteur afin qu'il identifie le sujet par rapport aux autres objets du contexte. Par exemple, si le contexte contient [1] un carré rouge, [2] un triangle bleu, [3] un cercle vert, le locuteur peut alors dire quelque chose comme "le rouge" pour identifier [1]. Si le contexte contient aussi un triangle rouge, il doit être plus précis et dire quelque chose comme "le carré rouge". Bien sûr, si les agents n'interagissent qu'entre eux, ils ne diront pas "le carré rouge" mais utiliseront leur propre langue et concepts qui, a priori, ne ressembleront pas au Français. Par exemple, un agent pourra dire "Malewina" pour signifier [EN-HAUT A-GAUCHE NIVEAU-ELEVE-DE-ROUGE]. A partir de cet indice linguistique, l'interlocuteur essaie de deviner quel est le sujet choisi par le locuteur. Il lui indique son choix en pointant vers un des objets. Cette étape est réalisée par la transmission de la direction dans laquelle la Talking Head regarde. Le jeu est un succès quand l'interlocuteur a deviné juste. C'est un échec si l'interlocuteur pointe vers un autre objet du contexte, ou si un des deux agents n'a pas été capable de réaliser une des étapes précédentes du jeu. Dans le cas d'un échec, le locuteur indique à l'interlocuteur, de façon non verbale, le sujet qu'il voulait désigner et les deux agents adaptent leur structures internes afin d'être efficaces dans les jeux futurs.

L'architecture des agents est constituée par deux composants: [1] un module de conceptualisation qui permet la catégorisation du monde perçu et la recherche d'un référent dans l'image et [2] un module de verbalisation permettant d'associer une forme verbale à un concept et d'interpréter une forme pour reconstruire son sens. Les agents commencent sans lexique ni ontologie préconstruite. Une ontologie partagée et un lexique doivent donc émerger au cours d'un processus d'auto-organisation. Les agents construisent, étendent et adaptent leur ontologie et leur lexique, au fur et à mesure qu'ils participent à des jeux de langage.

Le module de conceptualisation Les "sens" sont des catégories qui permettent de distinguer le sujet des autres objets du contexte. Les catégories sont organisées en arbres de discrimination (figure 2) dans lesquels chaque noeud contient un discriminateur capable de filtrer l'ensemble des objets en un sous-ensemble qui satisfait une certaine contrainte. Par exemple, il peut exister un discriminateur basé sur la position horizontale (HPOS) du centre d'un objet (normalisée entre 0.0 et 1.0) qui classe les objets du contexte en deux sous-ensembles: l'ensemble "gauche" quand HPOS < 0.5 (que nous designerons par [HPOS-0.0,0.5]) et l'ensemble droit quand HPOS > 0.5 ([HPOS-0.5,1.0]). Des sous-catégories supplémentaires sont créées en subdivisant la région de chaque catégorie. Par exemple, la catégorie "très à gauche" (ou [HPOS-0.0,0.25]) s'applique aux objets dont la valeur HPOS est dans la région [0.0, 0.25].

Un ensemble de catégories distinctives est trouvé en filtrant les objets du contexte en partant du sommet de chaque arbre de discrimination de façon à isoler uniquement le sujet. Il est, par exemple, possible que l'ensemble ([HPOS-0.5,1.0] et [VPOS-0.0, 0.25]) identifie le sujet sans ambiguïté car aucun autre objet du contexte n'a de valeur dans le domaine ainsi défini.

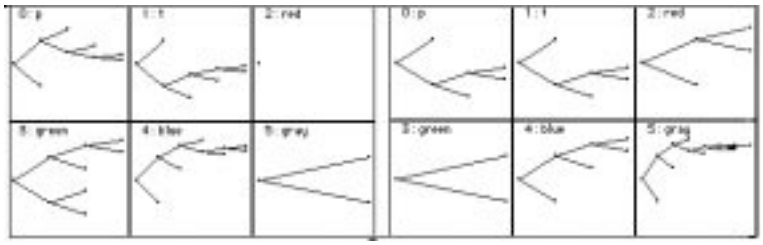

FIG. 2 – *Les arbres de discriminination de deux agents.*

Souvent, il y a plus d'une solution possible mais toutes les solutions sont transmises au module de verbalisation.

Les arbres de discrimination de chaque agent sont formés en utilisant des dynamiques de croissance et d'élagage couplées avec l'environnement. Les arbres de discrimination croissent de façon aléatoire par l'addition de nouveaux noeuds subdivisant les régions de catégories existantes. Les différents noeuds des différents arbres sont en compétition les uns avec les autres. L'utilisation et le succès de chaque noeud sont mesurés et les noeuds non pertinents pour les environnements rencontrés par les agents sont élagués. De plus amples détails au sujet des jeux de discrimination peuvent être trouvée dans (Steels, 1997).

Le module de verbalisation Le lexique de chaque agent consiste en une mémoire associative où sont rangées des associations entre des formes (ici des mots) et des sens (ici des catégories simples). Chaque association a un score. Les mots sont des combinaisons aléatoires de syllabes. Quand un locuteur doit exprimer une catégorie, il sélectionne tous les mots associés à cette catégorie, les ordonne et choisit celui avec le score le plus élevé pour le transmettre à l'interlocuteur. Quand l'interlocuteur doit interpréter un mot, il sélectionne tous les sens possibles de ce mot et teste ceux qui sont utilisables dans le présent contexte, i.e. ceux qui identifient un référent unique. Parmi ceux-ci, l'interlocuteur choisit celui avec le score le plus élevé.

En fonction du résultat du guessing game, le locuteur et l'interlocuteur modifient le score de certaines associations. Quand le jeu est un succès, ils augmentent chacun le score de leur association gagnante et diminuent celui des associations en compétition, implémentant ainsi un processus d'inhibition latérale. Quand le jeu est un échec, chacun diminue le score de l'association qu'il a utilisée. De nouvelles associations peuvent être mémorisées. Un locuteur crée un nouveau mot quand il n'a pas encore de mot pour le sens qu'il veut exprimer. Un interlocuteur peut rencontrer un nouveau mot qu'il n'a jamais entendu auparavant et dans ce cas, créer une nouvelle association entre ce mot et son sens probable. Pour determiner ce sens, l'interlocuteur utilise d'abord l'information non verbale donnée par le locuteur pour identifier le référent et catégorise ce dernier en utilisant ces propres arbres de discrimination. Ces mécanismes d'amorçage du lexique ont été décrits et validés dans des papiers plus anciens (Steels & Kaplan, 1998) et sont similaires à ceux étudiés par Oliphant (Oliphant, 1996).

Exemples Nous commencerons par un jeu de langage tout simple. Le locuteur, **a1**, a choisi comme sujet un triangle placé en bas de la scène. Il n'y a qu'un autre objet dans la scène, un rectangle, en haut. Dans ces conditions, la catégorie $[\text{VPOS-0.0,0.5}]_{\text{a1}}$, qui est valide quand la position verticale $VPOS < 0.5$, est applicable pour le triangle mais pas pour le rectangle. Si l'on suppose que **a1** a une association dans son lexique entre $[\text{VPOS-0.0,0.5}]_{\text{a1}}$ et le mot "lu", alors **a1** selectionne cette association et transmet le mot "lu" à l'interlocuteur qui est l'agent **a2**.

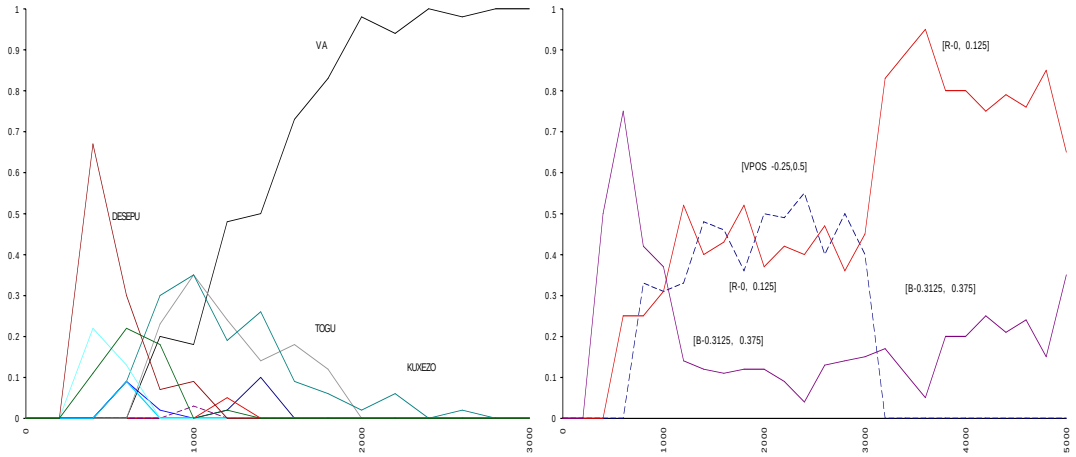

FIG. 3 – *Gauche: Diagramme RF, évolution de la fréquence de toutes les formes utilisées pour le même référent en 3000 jeux de langage. Droite: Diagramme FS, Evolution de la fréquence de tous les sens utilisés pour la même forme "va" sur 5000 jeux. Une situation au jeu 3000 provoque la perte d'un des sens.*

Si l'on suppose maintenant, que **a2** a rangé dans son lexique une association entre "lu" et $[\text{R-0.0,0.5}]_{\mathbf{a2}}$, il fait l'hypothèse que le sens de "lu" est $[\text{R-0.0,0.5}]_{\mathbf{a2}}$. Quand il applique cette catégorie à la scène, autrement dit quand il filtre les objets de la scène dont le niveau dans le canal rouge (R) n'est pas dans la région $[0.0, 0.5]$, il sélectionne un objet unique, le triangle. Ainsi **a2** conclut qu'il s'agit du sujet de l'interaction et pointe vers lui. Le locuteur reconnaît que l'interlocuteur a désigné le bon objet et le jeu est un succès. Ce jeu illustre une situation dans laquelle le locuteur et l'interlocuteur identifient le même référent avec deux sens différents. Le locuteur utilise la position verticale et l'interlocuteur utilise le niveau du canal rouge dans l'espace RGB.

Dans ce deuxième exemple, le locuteur est à nouveau **a1** et utilise la même catégorie et le même mot pour désigner le triangle. Mais l'interlocuteur, **a3**, interprète "lu" dans des termes de position horizontale $[\text{HPOS-0.0,0.5}]_{\mathbf{a3}}$ (partie gauche de la scène). Comme plusieurs objets dans la scène satisfont à cette catégorie, l'interlocuteur est incapable d'identifier un objet unique. Le locuteur pointe alors vers le sujet et l'interlocuteur crée une nouvelle association entre "lu" et $[\text{VPOS-0.0,0.5}]_{\mathbf{a3}}$. Cette association sera des lors en compétition avec celle qu'il a déjà.

3. Dynamiques de coévolution entre catégorisation et lexicalisation

Nous pouvons maintenant illustrer certaines dynamiques du guessing game, lorsqu'il est joué par des agents ancrés et situés qui interagissent dans un environnement physique partagé. Nous prendrons comme exemple une série de 5000 jeux pour un groupe de 20 agents. La première tendance est la suppression de la synonymie, par un processus de rétroaction positive qui décide quelle forme va être associée préférentiellement à un référent particulier. Le diagrame RF de la figure 3 montre clairement ce phénomène: le mot "va" devient dominant pour l'expression de ce référent. Cette suppression de la synonymie est due à l'inhibition latérale et à la boucle de rétroaction entre le succès d'une association et son utilisation dans les jeux futurs.

Quand nous étudions les différents sens du mot "va", grâce au diagramme FS (figure 3), nous voyons de façon claire que même après 3000 jeux l'ambiguïté reste dans le langage. Trois sens stables de "va" ont émergé: [R-0,0.125], [B-0.3125,0.375], et [VPOS-0.25,0.5]. Ils sont tous

aussi efficaces pour distinguer le sujet désigné par ”va” et aucune situation pouvant lever cette ambiguïté ne s'est encore présentée.

Au jeu 3000, l'environnement produit une scène dans laquelle une des catégories, qui était distinctive pour l'objet désigné par ”va” n'est plus acceptable. Concrètement, nous avons, en tant qu'expérimentateurs, déplacé l'objet en question juqu'à une position très proche d'un autre objet : la catégorie faisant intervenir la position verticale ne permet plus de les distinguer. Cette modification de l'environnement est immédiatement suivi par un certain nombre de jeux non réussis. Les échecs se produisent car ”va” ne permet plus de sélectionner le bon objet pour les agents pour qui ”va” est associé à [VPOS-0.25,0.5]: ils doivent adopter un nouveau sens pour ”va” compatible avec la nouvelle situation. Le diagramme FM de la figure 3 montre que le sens [VPOS-0.25,0.5] a disparu. Les autres sens, correspondants à des informations sur les couleurs, sont encore possibles et non pas été affectés par le déplacement de l'objet.

4. Conclusions

L'approche que nous utilisons pour la formation du lexique est différente des méthodes plus traditionnelles inspirées de Quine (Quine, 1960). Quine suppose que les agents apprennent le sens des mots par des abstractions inductives successives à partir de situations au cours desquelles ils observent des relations particulières entre des objets et des mots. Les propriétés communes des référents constituent le sens d'un mot. Elles sont induites par l'étude des similarités sur de nombreux exemples. Ce point de vue est aussi sous-jacent dans les approches utilisant des réseaux de neurones pour l'acquisition d'un lexique (voir par exemple (Hutchins & Hazlehurst, 1995) ou (Regier, 1995)). Dans notre approche, au contraire, les agents inventent des mots et des sens au cours de jeux de langage, formulent différentes hypothèses sur le sens des mots utilisés par les autres et testent ces sens quand ils sont locuteurs. L'évolution vers une cohérence lexicale dans la population (où un mot est associé à un sens dominant et un sens est associé à un mot dominant) est un phénomène collectif dirigé par les réponses du système à de nouvelles situations où les différents sens d'un mot ne sont plus compatibles les uns avec les autres. Nous pensons qu'une compréhension fine de ces dynamiques ouvrira des voies alternatives intéressantes pour le traitement automatique du langage naturel.

Références

HUTCHINS E. & HAZLEHURST B. (1995). How to invent a lexicon: the development of shared symbols in interaction. In N. GILBERT & R. CONTE, Eds., *Artificial Societies: The Computer Simulation of Social Life*. UCL Press.

OLIPHANT M. (1996). The dilemma of saussurean communication. *Biosystems*, **37**(12), 31–38.

QUINE W. (1960). *Word and Object*. Cambridge Ma: The MIT Press.

REGIER T. (1995). A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics*, **6**(1), 63–88.

STEELS L. (1997). Constructing and sharing perceptual distinctions. In M. VAN SOMEREN & G. WIDMER, Eds., *Proceedings of the European Conference on Machine Learning*, Berlin: Springer-Verlag.

STEELS L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, (103), 1–24.

STEELS L. & KAPLAN F. (1998). Spontaneous lexicon change. In *Proceedings of COLING-ACL 1998*, p. 1243–1249, Montreal: ACL.