Programme de recherche pour une délégation
à l'Institut de Recherche en Informatique de Toulouse
IRIT UMR 5505

Analyse automatique de la sémantique du langage naturel en théorie des types

Christian Retoré
(PR Université Bordeaux 1 – LaBRI UMR 5800)

Janvier 2012

1 Présentation

Je propose pour cette délégation un projet interdisciplinaire entre logique (théorie des types) et traitement automatique des langues sur une modélisation de la sémantique (compositionnelle et lexicale) conduisant à une analyse automatique à large couverture, via une intégration de cette modélisation dans la plateforme Grail (analyse automatique, syntaxique et sémantique du français à grande échelle). De plus une telle formalisation devrait aussi conduire à proposer des mesures de la complexité cognitive, permettant de quantifier notre difficulté à comprendre certains énoncés. Cette modélisation sémantique sera développée dans un cadre issu des travaux récents en logique : notamment en logique du second ordre, intuitionniste ou linéaire, en théorie des types et dans le cadre de la logique catégorique.

Ce programme de recherche s'inscrit dans deux projet impliquant le LaBRI (Bordeaux) et l'IRIT (Toulouse)

- Analyse et synthèse dans les grammaires catégorielles abstraites : du lexique au discours (ANR Contenus numériques et interactions)
- Projet ITIPY — Extraction automatique d'itinéraires dans des récits de voyages dans le corpus régional et historique de la médiathèque de Pau (projet des régions Aquitaine et Midi-Pyrénées)

Depuis longtemps l'IRIT a développé les liens entre les domaines connexes que sont la logique, l'intelligence artificielle, le le traitement automatique des langues,la linguistique formelle, ces deux derniers sujets étant développés en partenariat avec l'ERSS (CLLE UMR 5263, Toulouse). En ce qui concerne mon présent projet, sont particulièrement concernés les chercheurs de l'IRIT qui participent régulièrement aux rencontres de nos projets communs, Nicholas Asher (DR CNRS), Philippe Muller (MC), Sergeï Soloviev (PR), Nathalie Aussenac-Gilles (DR CNRS) ainsi que certains chercheurs de l'ERSS

Il est particulièrement intéressant de développer ce travail à l'IRIT car il poursuit des travaux de Nicholas Asher (DR) en particulier ceux décrit dans son livre de 2011 [Asher, 2011], utilise théories des types voisines de celles développées par Sergeï Soloviev (PR) [Soloviev and Luo, 2000]. Du point de vue du traitement automatique des langues pour une validation de ce travail sur

le corpus Itipy la collaboration avec Philippe Muller (MC) et Nathalie Aussenac-Gilles (DR) sera d'une grande utilité que ce soit pour l'annotation, le choix du lexique en liaison avec l'ontologie géographique. Finalement la description de la langue et la justesse linguistique du modèle proposé bénéficieront des interactions avec le CLEE ERSS, je pense notamment à Myriam Bras (PR), Anne Le Draoulec (CR) et Michel Aurnague (DR).

En contre partie, j'apporterais à l'IRIT ma connaissance des grammaires catégorielles (et notamment sur la plateforme Grail d'analyse syntaxique et sémantique du françaisde Richard Moot [Moot, 2010b]), du lambda calcul du second ordre, de la logique linéaire et de certaines questions sur lesquelles nous avons déjà des résultats [Moot and Retoré, 2011, Retoré, 2011, Moot et al., 2011a, Moot et al., 2011b] en utilisant la structure du lexique introduite dans [Bassac et al., 2010].

2 Logique et théorie des types pour la sémantique du langage naturel

Que ce soit pour le traitement automatique des langues ou dans l'exploration de la faculté de langage, la sémantique ainsi que l'interface entre syntaxe et sémantique nécessitent des modèles formels, notamment pour permettre une analyse sémantique automatique, et sans doute aussi pour une meilleure compréhension des processus cognitifs impliqués. En effet, pour le moment, force est de constater sur tous ces plans, le grand retard des travaux de sémantique sur ceux de syntaxe du langage naturel. Mon objectif global est de définir des modèles capables d'analyser automatiquement d'un énoncé isolé ou inscrit dans un contexte discursif voire dialogique, et, dans un second temps, de donner si possible une mesure de sa complexité, le tout dans le cadre de la théorie des types.

Dans l'analyse sémantique du langage naturel il convient de distinguer deux niveaux, lesquels peuvent également être traités dans le cadre de la théorie des types :

1. La métalogique (*glue logic*, logique d'assemblage, souvent décrite par un le lambda calcul) : elle permet d'obtenir le sens d'une unité complexe (par exemple la phrase) à partir du sens de chacune de ses unités qui la compose (par exemple le sujet, le syntagme verbal et les circonstants dans le cas d'une phrase). Traditionnellement, la métalogique est le lambda calcul simplement typé ou logique propositionnelle intuitionniste.
2. La logique dans laquelle on exprime le sens des unités simples ou complexes : traditionnellement c'est soit la logique du premier ordre, soit la logique d'ordre supérieur, et parfois la théorie des types.

Quelles sont les structures adéquates à la description de la sémantique ? quelle logique et quelle métalogique conviennent à la description du langage naturel ? Quelle est la structure du lexique où est stockée la sémantiques des unités élémentaires, c'est-à-dire des mots ?

3 Sémantique et pragmatique lexicale : le sens des mots en contexte

Le système de types sert initialement à contrôler que la composition assemble bien des sens compatibles, et que *Le cinq court.* ne soit pas possible. Mais on voit que le contexte peut forcer certaines interprétation, par exemple la phrase ci-dessus devient sémantiquement acceptable dans le contexte d'un match.

Il s'agit, lors de l'analyse de choisir le sens du mot approprié dans le contexte de la phrase, et aussi de rendre compte des prédications possibles et impossibles :

- possible : *Barcelone est un port cosmopolite.* (lieu et habitant)
- impossible : *Barcelone a battu Madrid et refuse un statut transfrontalier à la Cerdagne.* (club et institution)

Ce sera la principale question à travailler lors de ma délégation à l'IRIT si elle m'est accordée. Le modèle que nous avons proposé dans se distingue de celui de Nicholas Asher [Asher, 2011] en ce sens que ce n'est pas le type qui autorise ou non certains glissements de sens mais le mot lui-même qui se voit attribuer un ou plusieurs lambda termes dont l'utilisation est optionnelle et qui codent ces changements de sens.

Une autre différence est notre utilisation du lambda calcul du second ordre usuel comme métalogique tandis que Nicholas Asher [Asher, 2011] utilise des règles spécifiques issues de la logique catégorique. Quelle est la différence entre les logiques utilisées ? Laquelle décrit plus précisément les phénomènes linguistiques présents en corpus ? Du point de vue de la complexité des algorithmes de calcul des représentations sémantiques, quelle logique est la plus efficace ?

D'autres chercheurs de l'IRIT comme Sergeï Soloviev (avec Zhaohui Luo de Londres) ont construit une théorie des types avec sous-typage très pertinente comme langage des représentations sémantiques avec coercition [Soloviev and Luo, 2000] C'est très similaire à notre modèle, et nous aimerions là encore pouvoir comparer ces approches logiques de la sémantiques, notamment sur la question du sous-typage, très fréquente en analyse sémantique (les prédicats applicables aux animaux le sont des chiens, par exemple). Les connaissances de Nathalie Aussenac-Gilles sur les liens entre lexique et ontologies seront très appréciées, en particulier pour traiter de ces questions.

4 Quelques autres phénomènes sémantiques subtils mais très présents en corpus

Utilisant le modèle ci-dessus basé sur ΛTy_n, nous avons d'ores et déjà entrepris, souvent avec Richard Moot, de traiter des phénomènes sémantiques particuliers et de les intégrer dans la plateforme Grail. Précisons qu'il s'agit de construire les représentations sémantiques de la phrase et non de dite si elles sont vraies ou fausses dans une situation donnée (ce qui est fort difficile en présence de quantificateurs et/ou de prédicats vagues).

Le voyageur fictif Très présent dans notre corpus Itipy de récits de voyage, c'est le phénomène par lequel une route peut mettre en scène une personne qui la emprunte ladite route, personne qui n'existe pas forcément. *Ce chemin monte pendant deux heures, il vaut mieux en prendre un autre.* : le *pendant deux heures* nécessite de considérer quelqu'un qui emprunterait cette route, mais comme le montre la suite de la phrase, personne ne l'emprunte... [Moot et al., 2011a, Moot et al., 2011b] Sur tous les phénomènes spécifiques à la sémantique du temps et de l'espace, la proximité des chercheurs de CLLE ERSS sera très appréciée (M. Bras, A. Le Draoulec, M. Aurnague).

Les pluriels Evidemment très répandus, ils sont une source d'ambiguïtés entre les lectures collectives et distributives *Les benjamins ont gagné.* (ensemble, lecture collective), *Les étudiants ont un numéro d'anonymat.* (chacun a le sien, lecture distributive). Le même modèle permet de calculer automatiquement les lectures possibles [Moot and Retoré, 2011]

La quantification Qu'elle soit universelle (*tous les, chacun*) ou généralisée (*la plupart, la majorité*), la quantification est très répandue, mais par quelles formules logiques décrire les énoncés quantifiés ? comment construire automatiquement ces formules à partir de l'analyse syntaxique de la phrase ? La réponse donnée dans [Retoré, 2011] permet d'envisager des interprétations différentes en termes de preuves et de réfutations comme dans [Abrusci and Retoré, 2011a, Abrusci and Retoré, 2011b]

Le traitement de ces phénomènes classiques dans un cadre unique, celui de la théorie des types, a bien sûr retenu l'attention des collègues toulousains, et nous projetons les développement suivants : comment acquérir les entrées lexicales correspondantes ? quelles interprétations donner au formules logiques construites (puisque les interprétations standard ne fonctionnent pas en raison des prédicats vagues et des quantifications vagues).

5 Intégration dans la plateforme Grail : utilisation et validation sur le corpus Itipy

L'analyse syntaxique profonde et à grande échelle est faite dans les grammaires catégorielles abstraites inventées et développées par M. Moortgat et R. Moot [Moortgat, 1996, Moot, 2002]. Elle fonctionnent avec une grammaire à large échelle acquise sur corpus, initialement pour une grammaire du néerlandais parlé acquise sur le NWO Dutch Spoken Corpus [Moot, 2007] et maintenant pour une grammaire du français acquise sur le corpus de Paris 7 et le corpus Itipy [Moot, 2010b, Moot, 2010a]. Cette grammaire produit la sémantique sous forme de représentations discursives qui sont évidemment bien adaptée à une analyse du discours ultérieure.

L'extension au lambda calcul du second ordre utilisé par notre modélisation de la sémantique et de la pragmatique lexicale a été implantée pour de petits lexiques tests mais il conviendrait de réfléchir à automatiser la construction de ces entrées lexicales pour que tout le lexique dispose de représentations sémantiques riches. C'est l'un des objectifs de ce projet.

Philippe Muller est particulièrement intéressé par l'annotation du corpus, l'extraction du lexique sémantique, et l'analyse automatique avec Grail. Bien évidemment, le corpus Itipy, propriété de la médiathèque de Pau, fournit un champ d'expérimentation idéal : il s'agit d'un corpus de 576 334 mots, constitué de récits de voyage dans les Pyrénées, où les phénomènes sémantiques mentionnés ci-dessus sont très présents. La mise en place de l'analyse sémantique automatique incluant les phénomènes susmentionnés sera un grand pas vers la reconstruction automatique d'itinéraires passant par certains lieux à partir de ces récits, objectif du projet régional Itipy.

6 Conclusion

Un an à Toulouse permettrait de mettre en commun nos avancées sur l'analyse sémantique automatique du langage naturel, de discuter du meilleur choix possible lorsque nos approches diffèrent et d'inclure les progrès de cette analyse sémantique automatique dans la plateforme Grail.

Références

[Abrusci and Retoré, 2011a] Abrusci, M. and Retoré, C. (2011a). Quantification and interaction. In *Rebuilding logic and rethinking language in interaction terms (CLMPS workshop)*, France.

[Abrusci and Retoré, 2011b] Abrusci, V. M. and Retoré, C. (2011b). Quantification in ordinary language : from a critic of set-theoretic approaches to a proof-theoretic proposal. In Schröder-Heister, P., editor, *14th Congress of Logic, Methodology and Philosophy of Sciences*.

[Asher, 2011] Asher, N. (2011). *Lexical Meaning in context – a web of words*. Cambridge University Press.

[Bassac et al., 2010] Bassac, C., Mery, B., and Retoré, C. (2010). Towards a Type-Theoretical Account of Lexical Semantics. *Journal of Logic Language and Information*, 19(2) :229–245.

[Moortgat, 1996] Moortgat, M. (1996). Categorial type logic. In van Benthem, J. and ter Meulen, A., editors, *Handbook of Logic and Language*, chapter 2, pages 93–177. North-Holland Elsevier, Amsterdam.

[Moot, 2002] Moot, R. (2002). *Proof nets for linguistic analysis*. PhD thesis, UIL-OTS, Universiteit Utrecht.

[Moot, 2007] Moot, R. (2007). Automated extraction of type-logical supertags from the spoken dutch corpus. In Bangalore, S. and Joshi, A., editors, *The Complexity of Lexical Descriptions and its Relevance to Natural Language Processing : A Supertagging Approach*. MIT Press.

[Moot, 2010a] Moot, R. (2010a). Semi-automated extraction of a wide-coverage type-logical grammar for French. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

[Moot, 2010b] Moot, R. (2010b). Wide-coverage French syntax and semantics using Grail. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

[Moot et al., 2011a] Moot, R., Prévot, L., and Retoré, C. (2011a). A discursive analysis of itineraries in an historical and regional corpus of travels. In *Constraints in discourse*, page http ://passage.inria.fr/cid2011/doku.php, Ayay-roches-rouges, France. Projet ITIPY de la région Aquitaine.

[Moot et al., 2011b] Moot, R., Prévot, L., and Retoré, C. (2011b). Un calcul de termes typés pour la pragmatique lexicale. In *Traitement Automatique du Langage Naturel, TALN 2011*, pages 161–166, Montpellier, France. Projet ITIPY de la région Aquitaine.

[Moot and Retoré, 2011] Moot, R. and Retoré, C. (2011). Second order lambda calculus for meaning assembly : on the logical syntax of plurals. In *Coconat*, Tilburg, Pays-Bas.

[Retoré, 2011] Retoré, C. (2011). Specimens : "most of" generic NPs in a contextually flexible type theory. In *Genius III*, Paris, France.

[Soloviev and Luo, 2000] Soloviev, S. and Luo, Z. (2000). Coercion completion and conservativity in coercive subtyping. *Annals of Pure and Applied Logic*, 1-3(113) :297–322.