Lexique organisé pour la composition sémantique

Bruno Mery

Lexique organisé pour la composition sémantique – Résumé

Bruno Mery[1]

Récemment, nombreux ont été les travaux visant à intégrer à une sémantique compositionnelle Montagovienne des éléments d'inférence lexicale. Ainsi, [Pustejovsky, 1995] pose les bases d'un système linguistique complet pour les polysèmes logiques. D'autres études, dont [Asher, 2008], ont exposé des extensions très raffinées à la logique compositionnelle sous-tendant l'analyse sémantique de la phrase pour y intégrer ces mécanismes.

Notre propos est de définir un système complet d'analyse sémantique prenant en compte ces raffinements tels qu'exposés dans [Bassac et al., TBP] ; pour ce faire, nous proposons ici les éléments nécessaires à la construction du lexique d'un tel système.

Situation

Les polysèmes logiques sont des termes lexicaux dont l'usage admet plusieurs dénotations distinctes à partir d'un même signifiant originel (par exemple, *Paris* en tant que lieu géographique, population, ou organe de pouvoir Français). Un des problèmes classiques de cette classe de phénomènes est de caractériser les constructions co-prédicatives valides (*Paris, ville cosmopolite, s'étend autour des méandres de la Seine*) ou invalides (***?*** *Paris, ville cosmopolite, a signé un accord commercial avec le nouveau gouvernement Thaïlandais*).

Afin de les traiter dans un système logique compositionnel, il nous faut déterminer un *langage de termes* associés aux entrées lexicales en permettant le calcul, ciblant une certaine classe de *formules logiques* résultant de l'analyse sémantique, et évoquer les *modèles de l'interprétation* de telles formules.

1 Langage de types et de termes

Les termes, composant les entrées lexicales et destinés à la composition sémantique, doivent permettre d'effectuer l'ensemble des calculs voulus. Le langage employé est le λ-calcul typé à n sortes (n dépendant du lexique). Concrètement, il s'agit d'un système proche de la proposition originale de [Montague, 1974], où les types utilisés sont e (pour les entités) et t pour les valeurs de vérité, ainsi que tous les types fonctionnels pouvant être construits ; avec n sortes, on distingue n types atomiques se substituant à e. Le calcul proposé par Montague correspond donc au calcul à 1 sorte.

De plus, les types sont variables à la manière du Système F de [Girard et al., 1989] ; le langage de types est donc basé sur une logique d'ordre supérieur.

Chaque terme lexical est un k-uple de termes, le premier d'entre eux décrivant le sens premier du lexème et les suivants, les opérations applicables pour l'accession aux sens dérivés. À chacun de ces éléments peuvent être associées des contraintes d'utilisation.

Dans notre exemple, *Paris* est associé à l'entrée lexicale suivante :

$$\left(\text{Paris}^T\ ,\ \frac{\lambda x^T.(f_L^{T\to L}\ x)}{\varnothing}\ ,\frac{\lambda x^T.(f_P^{T\to P}\ x)}{\varnothing}\ ,\ \frac{\lambda x^T.(f_G^{T\to G}\ x)}{global}\right)$$

Ici : T est le type des villes, L celui des lieux géographiques, P des groupes de population, G des corps gouvernementaux ; les f_* sont des *morphismes optionnels* permettant de mettre en œuvre les mécanismes de sémantique lexicale. Le système de calcul distingue deux modes d'application de ces morphismes, *local* et *global*, et la contrainte *global* sur le dernier morphisme permet d'exclure les co-prédications inélégantes telles qu'évoquées précédemment.

[1]Université de Bordeaux, LaBRI/CNRS, équipe Signes/INRIA

Pour un lexique optimal en espace, ces informations peuvent être hiérarchisées suivant une ontologie de types ; en particulier, toute entrée de type T sera très semblable à celle de *Paris* (cette structure d'héritage ontologique est décrite en détail dans [Pustejovsky, 1995]).

Pour construire ce lexique, on peut se baser sur un lexique génératif existant : les *qualia* et *dot objects* définis dans cette structure deviennent, dans notre formalisme, l'objet des morphismes à intégrer au lexème correspondant. On peut également construire manuellement un tel lexique, en utilisant la même démarche que pour un lexique génératif : chaque entrée lexicale doit répondre à la question "quelles facettes de ce mot peuvent être mises en œuvre pour autre chose que son sens principal" ? Par ailleurs, le lexique doit également intégrer une description de l'interprétation des termes, qui dépend du modèle choisi. Quelle que soit la valeur (dénotation, symbole…) choisie pour les signifiants, il est toujours intéressant de préciser le rôle des morphismes intégrés au lexique (e.g., f_P peut être glosé par "habitants de…").

2 Formules logiques

Après composition et réduction des termes lexicaux, l'analyse sémantique résulte en une formule logique. De multiples *modus operandi* sont possibles. Nous avons choisi de dégager des formules de la logique classique du premier ordre, sans modalités (les évaluations modales ou temporelles pouvant être reléguées à l'interprétation).

Soient les prédicats $\lambda x^P.(\text{cosmopolite}^{P\to t}\ x)$ et $\lambda x^L.(\text{fluvial}^{L\to t}\ x)$. Pour résumer, le mode de calcul fait appel à l'ensemble des combinaisons de termes accessibles aux prédicats et arguments concernés, pour sélectionner les termes de type attendu. Ainsi, la phrase *Paris est une ville cosmopolite située sur un fleuve* donne lieu au terme suivant :

$$\begin{array}{l}((\Lambda\xi\,\lambda x^{\xi}\,\lambda f^{\xi\to P}\lambda g^{\xi\to T}\ .\ (\text{et}\ (\text{cosmopolite}^{P\to t}(f\ x))\ (\text{fluvial}^{L\to t}(g\ x))))\\ \qquad\qquad \{T\}\ \text{Paris}^{T}\ \lambda x^{T}.(f_P^{T\to P}\ x)\ \lambda x^{T}.(f_L^{T\to L}\ x))\end{array}$$

Le tout, après réduction, permet d'obtenir la formule logique suivante :

$$\text{et}(\text{cosmopolite}(f_P(\text{Paris})), \text{fluvial}(f_L(\text{Paris})))$$

3 Modèles d'interprétation

Le but de l'analyse sémantique du texte est d'établir une (ou plusieurs) interprétations dans un modèle adapté. La modularité est complète, et tous les choix restent possibles : la théorie des modèles classique, une variante hyperintesionnelle, ou les modèles de Kripke, par exemple.

Afin d'obtenir une représentation à la fois synthétique, pour faciliter les raisonnements calculatoires, complète et pouvant être hiérarchisée pour tenir compte des contraintes discursives, notre choix se porte sur la sous-partie de l'univers de Herbrand pour la logique des prédicats du second ordre pour laquelle la sémantique calculée au cours de l'analyse est vérifiée, c'est-à-dire l'ensemble des *entités* et des *faits* établis au cours de l'énonciation et de l'analyse subséquente – les informations de type étant représentées par autant de prédicats adaptés.

Ainsi, la phrase *Paris est une ville cosmopolite située sur un fleuve* sera interprétée comme le singleton *Paris* représentant l'ensemble des entités analysées, et l'ensemble de faits $\{T(\text{Paris}), \text{cosmopolite}(f_P(\text{Paris})), \text{fluvial}(f_L(\text{Paris}))\}$.

On peut ensuite donner les interprétations de chacun des termes. En particulier, si le lexique comporte les définitions symboliques des morphismes :

Entités	Faits
Paris	T(Paris)
Habitants de Paris	P(Habitants de Paris), Habitants de Paris = Habitants de...(Paris)
	cosmopolite(Habitants de Paris)
Site de Paris	L(Site de Paris), Site de Paris = Site de...(Paris)
	fluvial(Site de Paris)

L'avantage d'une telle représentation est l'intégration immédiate dans des paradigmes connus du traitement des connaissances (les bases de données relationnelles, par exemple), permettant d'effectuer ensuite automatiquement et efficacement des recherches d'informations ou des confrontations de versions.

D'autre part, il est facile de modéliser les contraintes propres au dialogue, ou au récit à plusieurs acteurs. Pour ce faire, il faut conserver une représentation du lexique distincte par agent ; les modifications induites au fil de la composition lexicale resteront alors particulières à chacun, et pourront être intégrés aux différents points de vue, suivant le formalisme choisi pour représenter le discours.

Conclusion

Cette proposition d'organisation du lexique se veut avant tout fondée sur un formalisme bien défini, qui puisse facilement être intégré aux autres aspects de la sémantique formelle. En effet, il est souvent reproché aux théories comme [Pustejovsky, 1995] de ne pas pouvoir être imbriquée dans les formalismes classiques de sémantique compositionnelle, ou aux raffinements comme [Asher, 2008] d'être trop complexes à mettre en œuvre.

Sans prétendre être plus adaptée à la description du sens lexical que ces formalismes, notre approche reste aussi proche que possible des principes de la sémantique de Montague, et assez générique pour s'adapter à de nombreuses théories connexes.

Aussi, nous espérons que le système ici proposé puisse faire partie d'un formalisme complet et fonctionnel d'analyse sémantique, prenant en compte les problèmes particuliers de la polysémie logique.

Références

[Asher, 2008] Asher, N. (2008). A Type Driven Theory of Predication with Complex Types. *Fundamenta Informaticæ*, 84(2) :151–183.

[Bassac et al., TBP] Bassac, C., Mery, B., and Retoré, C. (TBP). Towards a Type-Theoretical Account of Lexical Semantics. *Journal of Language, Logic, and Information*. To appear.

[Girard et al., 1989] Girard, J.-Y., Taylor, P., and Lafont, Y. (1989). *Proofs and types*. Cambridge University Press, New York, NY, USA.

[Jacquey, 2001] Jacquey, E. (2001). *Ambiguïtés lexicales et traitement automatique des langues : modélisation de la polysémie logique et application aux déverbaux d'action ambigus en français*. PhD thesis, Université de Nancy 2.

[Montague, 1974] Montague, R. (1974). The proper treatment of quantification in ordinary English. In Thomson, R. H., editor, *Formal Philosophy*, pages 188–221. Yale University Press, New Haven Connecticut.

[Nunberg, 1993] Nunberg, G. (1993). Transfers of meaning. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 191–192, Morristown, NJ, USA. Association for Computational Linguistics.

[Pustejovsky, 1995] Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.