Sémantique lexicale et distributionnalisme : éléments pour le repérage automatique du sens en corpus

Emmanuel Cartier
Université Paris 13 Sorbonne Paris Cité,
LIPN - équipe RCLN UMR 7030 CNRS

Dans cette présentation, nous chercherons, à la lumière des dernières intuitions de (Harris, 1988) et des avancées de la sémantique distributionnelle en Traitement automatique des langues (TAL), à vérifier dans quelle mesure il est possible de rendre compte du sens des unités lexicales de façon automatique, en s'appuyant sur les hypothèses du distributionnalisme.

L'une des caractéristiques distinctives du distributionnalisme harrissien est de ne prendre appui que sur les matérialisations écrites ou orales des langues, et de ne pas recourir à des hypothèses sur la pensée sous-jacente ou sur la relation de référence qui lie le langage au monde.

La sémantique distributionnelle (Baroni et al. 2010) se fonde sur une hypothèse qu'Harris énonce comme suit:

...if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution. (Harris, 1954, p.786)

La sémantique distributionnelle s'appuie en réalité sur trois hypothèses issues du distributionnalisme que nous formulerons ainsi :

Hypothèse 1 : les unités linguistiques sont repérables par leur répétition en corpus.

La première hypothèse donne la clé d'identification des unités linguistiques. La reprise de ces hypothèses par les statistiques lexicales, les linguistiques de corpus, et les grammaires de construction, a permis de revisiter la notion d'unité lexicale, et d'étendre la notion à celle de construction et de forme-sens (Goldberg, 2013).

Hypothèse 2 : leur usage-sens est repérable au moyen de la répétition des contextes dans lesquels ils sont pris.

La seconde hypothèse fonde une étude systématique des préférences sélectionnelles des parties du discours. (Ramish, 2015) montre ainsi que la combinaison entre les schémas syntaxiques de syntagmes nominaux les plus fréquents dans une langue donnée et un calcul de répétition de mots graphiques permet de récupérer un grand nombre de locutions. (Kilgarriff, 2004) déduit, sur cette base, les structures argumentales les plus fréquentes des verbes, qu'il appelle *Word Sketches*.

Hypothèse 3 : deux unités lexicales partageant un grand nombre de contextes sont dans une relation de similarité sémantique.

Cette hypothèse est celle qui fonde une étude sémantique sur des bases distributionnelles, et a donné lieu à de nombreux travaux en TAL, montrant qu'il est possible, en se basant uniquement sur la distribution des unités lexicales, d'identifier des lexies en relation de similarité sémantique.

De manière tardive, (Harris, 1988) va compléter sa méthode d'analyse des langues en explicitant quatre contraintes : l'ordre partiel, l'inégalité de probabilité d'occurrence, la réduction et la linéarisation.

Dans cet article, nous montrerons, au travers d'une étude en français sur gros corpus centrés sur le sens verbal, que les travaux actuels en TAL n'ont pas pris suffisamment en compte ces quatre contraintes, et notamment celles d'ordre partiel, de réduction et de linéarisation.

L'expérimentation que nous présenterons s'appuiera sur un corpus du français contemporain de 100 millions de mots issu du journal Le Monde. Nous étudierons une centaine de verbes du français, avec pour objectifs de montrer :

- d'une part, que les approches de la sémantique distributionnelle et des statistiques lexicales ont omis de prendre en compte trois contraintes explicitées par (Harris, 1988) à savoir les contraintes de réduction et de linéarisation, ainsi que celle d'ordre partiel.
- Qu'il est donc nécessaire, pour appliquer efficacement et automatiquement les hypothèses distributionnelles d'accès au sens, de tenir compte des contraintes des langues et de tenter de rétablir la forme canonique des phrases simples telle qu'elle peut être décrite dans l'ordre partiel non réduit et non contraint. Nous montrerons, dans le cadre d'une recherche sur les sens verbaux, quelles sont les opérations à effectuer sur le corpus brut pour s'approcher ces "phrases simples".
- que l'approche du sens par la méthode distributionnelle est une méthode efficace et automatique d'accès au sens, d'une part par le biais de l'explicitation des lexies « similaires » (et au mieux synonymes), d'autre part par le biais de l'explicitation du fonctionnement syntaxico-sémantique des unités lexicales. Nous expliciterons un certain nombre d'exemples de résultats de ces calculs distributionnels automatiques.
- Enfin, que le sens, comme tout ce qui a trait au langage, est l'empire du continu et que, s'il est possible d'expliciter des formes-sens « prototypiques », la réalité des discours explicitent également un continuum de formes-sens, en synchronie, et une stabilité relative des formes-sens, en diachronie, même courte.

Baroni M., Lenci A. (2010) Distributional Memory : A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36-4 (2010), 50

Goldberg A. (2013) Constructionist Approaches. *The Oxford Handbook of Construction Grammar*, Edited by Thomas Hoffmann and Graeme Trousdale, Oxford University Press.

Harris Z. S. (1954). Distributional structure. *Word,* 10(23), 146–162. [trad. 1970 : La structure distributionnelle", *Langages*, No.20, 14-34. Paris]

Harris Z. S. (1988). *Language and Information*. New York: Columbia University Press, ix, 120 pp. [Revised version of the Bampton Lectures given at Columbia University, New York City, in Oct. 1986.]

Kilgarriff A. et al. (2004). The Sketch Engine. *Proceedings of Euralex*, p. 105–116, Lorient.

Ramisch C. (2015) Multiword Expressions Acquisition: A Generic and Open Framework. *Theory and Applications of Natural Language Processing* series XIV, Springer, ISBN 978-3-319-09206-5, 230 p., 2015.