# Inférence de règles déductives par abduction

Mathieu Lafourcade, Manel Zarrouk, Alain Joubert

# Inférence de règles déductives par abduction

Mathieu Lafourcade, Manel Zarrouk, Alain Joubert
LIRMM, 161, rue ADA 34392 Montpellier Cedex 5
mathieu.lafourcade@lirmm.fr, manel.zarrouk@lirmm.fr,
alain.joubert@lirmm.fr

RESUME

Les réseaux lexico-sémantiques sont un type de ressources majeur en TAL. Indépendamment des stratégies de construction utilisées, inférer automatiquement des règles avec lesquelles de nouvelles relations peuvent être produites est une approche possible pour améliorer la couverture et la qualité globale de la ressource. Ce type de règle peut également être utilisé avec profit pour l'analyse sémantique de textes. Dans ce contexte, un moteur d'inférences a pour tâche d'explorer le contenu du réseau lexical et d'en extraire des règles potentielles. Seules les règles validées par des contributeurs ou des experts sont mémorisées et appliquées afin d'enrichir le réseau et en retour de découvrir de nouvelles règles. La nature polysémique des termes de la ressource et les nombreux silences sont autant d'obstacle à la validation automatique des règles candidates, cependant des évaluations statistiques combinées à des traits symboliques permettent d'obtenir et de filtrer les règles douteuses efficacement.

ABSTRACT

In NLP, lexical semantic networks represent a major and crucial type of resources. Nonetheless the construction strategy used, automatically inferring rules with which new relations can be produced is a possible way to improve coverage and global quality of the resource. Such rules can be used likewise profitably for semantic analysis of texts. Within that context, an inference engine has for goal to explore the contents of the lexical network and extract potential rules. Only rules validated by contributors or experts are memorized and applied latterly as patterns to enrich the network and so on discover new rules. The polysemy of terms in the resource and the presence of silence represent obstacles to the automatic validation of the candidate rules. However, statistical evaluations combined to some symbolic features, allow efficiently to have and to filter out the dubious rules.

MOTS-CLES: inférence de règles, enrichissement, réseau lexico-sémantique, traits symboliques, filtrage statistique.
KEYWORDS: rule inference, enrichment, semantic lexical network, symbolic features, statistical filtering.

## 1. Contexte et motivation

Le projet JeuxDeMots (Lafourcade, 2007) vise à produire un réseau lexico-sémantique à l'aide d'approches contributives dont notamment des jeux en ligne. L'inférence d'associations dans le réseau est un moyen efficace de le consolider, à savoir ajouter de nouvelles relations et de détecter celles possiblement erronées. Dans (Zarrouk et al, 2013) un schéma d'inférence déductif a été proposé afin de produire des associations dans un réseau lexical. Les règles d'inférence, qui sont des schémas avec une seule variable, peuvent être utilisées pour l'amélioration de la ressource, mais également avec profit lors d'une analyse sémantique de texte.

La plupart des ressources existantes de type réseau lexico-sémantique ont été construites manuellement. Des outils sont souvent utilisés pour la vérification de la cohérence, mais la tâche reste coûteuse en temps et en prix. Les approches automatisées à partir de corps sont généralement limitées à la cooccurrence des termes, l'extraction de relations sémantiques précises restant difficile. L'externalisation ouverte (crowdsourcing) émerge en particulier avec l'avènement de Amazon Mechanical Turk et plus largement avec Wikipédia. Wordnet ((Miller, 1990) et (Fellbaum, 1998)) est un réseau lexical de l'anglais fondé sur des pouvant être considérés comme des concepts. EuroWordnet (Vossen, 1998), une version multi-langues de Wordnet et WOLF (Sagot, 2008), une version française de Wordnet, sont issus de croisements automatiques de Wordnet avec d'autres ressources lexicales suivi d'une vérification manuelle partielle. Dans le domaine de l'intelligence artificielle, Cyc (Lenat, 1995) est une base de connaissances très riche ayant demandé un effort manuel particulièrement important. BabelNet (Navigli, 2012) est un réseau lexical multilingue construit automatiquement à partir des cooccurence de termes de Wikipédia. HowNet (Dong, 2008) est un autre exemple d'une base de connaissances bilingue (anglais et chinois) contenant des relations sémantiques entre des formes de mots, des concepts et des attributs. (Harabagui, 1998) a réalisé des inférences à partir de textes en se basant sur WordNet comme base de connaissance. L'évaluation de règles inférées depuis des corpus reste difficile comme le souligne (Zeichner, 2012). A notre connaissance, il n'existe pas de travaux d'inférence endogène de règles de déduction à partir de réseaux lexicaux construits par externalisation ouverte.

Un réseau lexico-sémantique peut contenir des formes de mots, des termes, des concepts, des expressions composées et idéalement des sens de termes (des usages). Les usages ne sont pas pour autant nécessairement totalement désambigüisés et peuvent avoir eux-mêmes plusieurs raffinements. Par exemple, *frégate* peut être un *oiseau* ou un *navire*. Une *frégate>bateau* peut être distinguée comme un *navire moderne* ou un *navire à voiles ancien*. Les concepts présents dans le réseau peuvent prendre la forme de traits sémantiques généraux, par une relation nommée SEM non ambigus sous la forme de symboles (VIVANT, HUMAIN, ARTEFACT, OBJET NATUREL, LIEU, TEMPS, etc.). Dans le contexte d'une approche collaborative, un réseau lexical doit être considéré comme en cours de construction et l'est en pratique. Certains usages ainsi que de nombreuses relations peuvent manquer et il est difficile de connaître l'état de silence de chaque entrée.

Il s'agit alors de produire automatiquement des règles d'inférence déductives par inspection du contenu du réseau lexical. Une règle, une fois validée, sera appliquée afin de produire de nouvelles relations dans le réseau. La polysémie du réseau ainsi que les silences sont les principaux obstacles à l'identification statistique de règles correctes. Par contre, l'information symbolique quand elle est présente constitue une aide précieuse au filtrage des règles candidates.

## 2. Expérience et résultats

A partir du réseau lexical JeuxDeMots nous avons cherché à produire des règles inductives de la forme « *Si x est une sorte de B alors x a une relation R avec C* » (que l'on notera : $\forall x, is\text{-}a(x, B) \Rightarrow R(x, C)$ ). L'interprétation d'une telle règle est qu'elle doit être vraie pour la plupart des *x*, modulo les exceptions qui devront rester en nombre limité pour être considérées comme telles. Il s'agit depuis des informations potentielles présentes dans le réseau de déterminer celles qui sont systématiquement valides. L'approche abductive a consisté à énumérer systématiquement les propriétés existantes pour chaque terme ayant au moins un hyperonyme. Il s'agit donc de trouver dans le réseau lexical des exemples fournissant un support à l'abduction de règles. Pour chaque terme B, on effectue les comptages suivants :

- #E : le nombre de *x* (tel que is-a(x, B) > 0), il s'agit des termes concernés par la règle ;
- #ER$^+$ : le nombre de *x* (is-a B) renseignés pour R et ayant R(x,C) > 0 (les exemples positifs) ;
- #ER$^-$ : le nombre de *x* (is-a B) renseignés pour A et ayant R(x, C) < 0 (les contre-exemples).

On ne considère que les règles candidates où le nombre d'exemples #ER$^+$ dépasse un seuil donné (>= 1). L'approche produit un grand nombre de règles candidates, à savoir plus de 330 000, pour 91 896 relations x is-a B présentes dans le réseau.

| # exemples min (#min) | 1 | 3 | 5 | **10** | 20 | 40 |
|---|---|---|---|---|---|---|
| # candidats | 336 000 | 3 593 | 1 718 | **649** | 330 | 92 |

La quantité de silence ($\#S = \#E - (\#ER^+ + \#ER^-)$ ) pour une règle est le nombre d'exemples relevant de la règle candidate qui ne sont pas du tout renseignés pour la relation R. Nous en déduisons le taux de véracité TV ($\#S/\#ER^+$) de la règle candidate et un taux de fausseté TF ($\#S/ER^-$) . Par exemple, pour la règle candidate *Tout oiseau vole* $\forall x, is\text{-}a(x, oiseau) \Rightarrow agent(x, voler(déplacement\ aérien))$, nous avons TV = 87% et TF=0.7%. C'est à dire que pour les termes ayant pour hyperonyme *oiseau*, ceux renseignés pour la relation *agent* sont très majoritairement reliés à *voler* et qu'un petit nombre est négativement relié à *voler* (*autruche*, *émeu*, *casoar*, *pingouin*). Empiriquement, il a été constaté que la tolérance de 1% d'exceptions combinée à au maximum 50% de silence fournissait le F1-score optimal de 80% (maximum de règles fausses et minimum de règles correctes éliminées) pour un nombre minimal d'exemples #min = 10.

La polysémie est un facteur de règles erronées dont l'effet est grandement atténué par la prise en considération de traits symboliques présents dans le réseau. Par exemple, nous avons obtenu la règle candidate fausse $\forall x, is\text{-}a(x, métier) \Rightarrow partie(x, cou)$ car la plupart des termes de métier fusionnent à la fois l'activité et la personne. En considérant les traits sémantiques majoritaires dans le réseau lexical, il est possible d'abduire la règle $\forall x, is\text{-}a(x, métier) \wedge sem(x, \text{HUMAIN}) \Rightarrow partie(x, cou)$, qui s'interprète de la façon suivante *«si A est un terme de métier et a le trait sémantique HUMAIN alors A peut avoir comme partie COU »*. Pour le terme *métier à risque*, le trait sémantique HUMAIN ne sera pas trouvé et la règle ne sera pas appliquée.

La prise en compte de symboles sémantiques associés aux termes dans le réseau lexical de JeuxDeMots permet d'obtenir un F1-score de 93%. Les cas d'échec sont les cas rares non renseignées aboutissant à des généralisations abusives, et qui concernent un nombre de règles réduit (8%) où il est possible de solliciter les contributeurs afin d'en énumérer les exceptions. Au final, l'abduction consiste ici à débusquer et transformer des informations indiquant une possibilité *(des oiseaux peuvent voler)* en règles généralement vraies *(tout oiseau peut voler).*

## 3. Références


BLANCO, E. et MOLDOVAN, D. (2011). A model for composing semantic relations. *Ninth International Conference on Computational Semantics (IWCS'11),* Oxford, United Kingdom, pages 45–54.

DONG, Z. et DONG, Q. (2006). *HowNet and the Computation of Meaning.* WorldScientific, London.

FELLBAUM, C. et MILLER, G. (1998). *(eds) WordNet.* The MIT Press.

HARABAGIU, S. et MOLDOVAN, D. (1998). Knowledge processing on an extended wordnet. *WordNet : An Electronic Lexical Database, MIT Press.*, pages 381–405.

LAFOURCADE, M. (2007). Making people play for lexical acquisition. *In Proc. SNLP 2007, 7th Symposium on Natural Language Processing.* Pattaya, Thailande, 13-15 December 2007, page 8 p.

LENAT, D. (1995). Cyc : A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

LIEBERMAN, H., SMITH, D. A. et TEETERS, A. (2007). Common consensus : a web-based game for collecting commonsense goals. *In Proc. of IUI*, Hawaii., page 12 p.

MARCHETTI, A., TESCONI, M., RONZANO, F., MOSELLA, M. et MINUTOLI, S. (2007). Semkey : A semantic collaborative tagging system. *in Procs of WWW2007*, Banff, Canada, page 9 p.

MIHALCEA, R. et CHKLOVSKI, T. (2003). Open mindword expert : Creating large annotated data collections with web users help. *In Proceedings of the EACL 2003, Workshop on Linguistically Annotated Corpora (LINC 2003)*, page 10 p.

MILLER, G., BECKWITH, R., FELLBAUM, C., GROSS, D. et MILLER, K. (1990). Introduction to wordnet : an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

NAVIGLI, R. et PONZETTO, S. (2012). Babelnet : Building a very large multilingual semantic network. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010, pages 216–225.

SAGOT, B. et FIER, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. *TALN 2008*, Avignon, France, 2008., page 12.

von AHN, L. et DABBISH, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8):58–67.

VOSSEN, P. (1998). *Eurowordnet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, Norwell, MA, USA, page 200.

ZARROUK, M., LAFOURCADE, M and JOUBERT, A. (2013) Inference and reconciliation in a lexical-semantic network *In proc of 14th International Conference on Intelligent Text Processing and Computational Linguistic (CILING-2013)*, University of the Aegean, Samos, Greece, March 24–30, 2013, 13 p.

ZEICHNER, N., BERANT, J. and DAGAN, I. (2012) Crowdsourcing Inference-Rule Evaluation *ACL 2012*, 5 p.

ZESCH, T. et GUREVYCH, I. (2009). Wisdom of crowds versus wisdom of linguists measuring the semantic relatedness of words. *Natural Language Engineering, Cambridge University Press.*, pages 25–59.