Compte rendu pour le CST de la thèse
Consolidation de grands réseaux lexicaux

Thèse de Manel ZARROUK
supervisée par Mathieu LAFOURCADE
Date de début: 01/11/2012 Date prévue de fin/ 30/11/2015

1 Contexte et problématique

Développer un réseau lexico-sémantique pour le TAL est l'un des enjeux majeurs du domaine. La plupart des ressources existantes ont été construites à la main, comme dans le cas de WordNet (Miller *et al.*, 1990). Bien entendu, quelques outils sont généralement utilisés pour la vérification de la cohérence, mais cependant la tâche reste coûteuse en temps et en prix. Les approches entièrement automatisées sont généralement limitées à la co-occurrence des termes car l'extraction des relations sémantiques précises entre termes à partir d'un texte reste difficile. De nouvelles impliquant l'externalisation ouverte (*crowdsourcing*) émergent en TAL spécialement avec l'avènement de Amazon Mechanical Turk ou plus largement avec Wikipédia et le Wiktionnaire pour ne citer que les plus connus. Wordnet ((Miller *et al.*, 1990) et (Fellbaum et Miller, 1998)) est un réseau lexical basé sur des synsets pouvant être globalement considérés comme des concepts. (Vossen, 1998) avec EuroWordnet, une version multi-langues de Wordnet et (Sagot et Fier, 2008) avec WOLF, une version française de Wordnet, ont utilisé des croisements automatiques de Wordnet avec d'autres ressources lexicales suivi d'une vérification manuelle partielle. (Navigli et Ponzetto, 2012) a construit BabelNet, un grand réseau lexical multilingue à partir de l'encyclopédie Wikipédia mais en se basant sur les co-occurences entre termes. Dans le domaine de l'intelligence artificielle, Cyc (Lenat, 1995) est un exemple de base de connaissances très redondante ayant demandé un effort manuel particulièrement important. Hownet (Dong et Dong, 2006) est un autre exemple d'une grande base de connaissances bilingue (anglais et chinois) contenant des relations sémantiques entre les formes de mots, les concepts et les attributs.
La construction collaborative d'un réseau lexical peut être catégorisée selon deux stratégies. Premièrement, comme un système contributif du type Wikipédia où des volontaires complètent les entrées (cas du Wiktionnaire). Dans un second cas, les contributions sont faites indirectement par l'entremise de jeux, connus sous le nom de GWAP (Game With A Purpose) (von Ahn et Dabbish, 2008).

L'expérience montre que les joueurs/contributeurs complètent le réseau sur ce qui leur parait intéressant. Ce faisant, un grand nombre de relations *triviales* ne sont pas présentes bien qu'elles demeurent pourtant nécessaires à l'obtention d'un réseau de qualité visant à être utilisé dans diverses applications du TAL dont notamment l'analyse sémantique (*comprendre* automatiquement un texte et en tirer des relations fines).

2 Contribution

Le but de notre recherche de thèse est d'enrichir, évoluer et consolider ce type de réseaux lexico-sémantiques construits par peuplonomie en produisant des inférences de nouvelles relations à partir de celle existantes.
Dans nos travaux nous nous fondons sur le réseau lexical issu du projet JeuxDeMots (Joubert et Lafourcade, 2008) pour développer les aspects théoriques ainsi que pour mener des expériences pratiques. Afin de consolider ce réseau, j'utilise une approche par inférence qui permet de déduire de nouvelles relations à partir de celles existantes. L'approche est uniquement endogène en ce qu'elle ne s'appuie sur aucune ressource externe au réseau. Les relations inférées sont soumises pour vote aux contributeurs et par la suite proposées à une validation ou invalidation par un expert. Une grande majorité des inférences se révèle correcte. Toutefois, une part non négligeable se révèle fausse et il convient de déterminer pourquoi. Ce processus d'explication constitue la réconciliation entre le moteur d'inférences et le validateur, mené à l'aide d'un dialogue lui permettant d'expliciter en quoi la relation considérée est incorrecte. Les causes possibles sont de trois natures : erreur dans une des prémisses, exception, ou confusion liée à la polysémie.

Jusqu'à présent, nous avons établi quatre schémas sur lesquels peut se baser ce système d'inférence sans compter d'autres schémas actuellement à l'étude.

2.1 La déduction

La déduction est un schéma d'inférence descendant fondé sur la transitivité de la relation ontologique *is-a*. Ce schéma transfert les relations du générique vers le spécifique. Comme illustré sur la figure 1 les relations (1) et (2) constituent les prémisses (relations déjà existantes) et la relation (3) la conclusion proposée et devant être validée pour être incluse dans le réseau.
Notre hypothèse peut se présenter comme suit : si le terme *A* est un type de *B* et que *B* a une relation de type *R* avec *C*, on peut supposer que le terme *A* possède une relation de type *R* avec *C*.

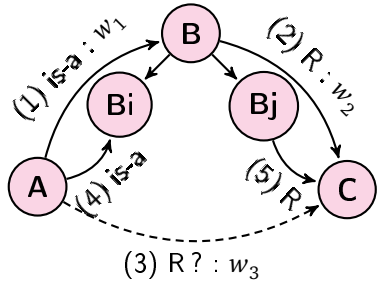

FIGURE 1 – Schéma d'inférence déductive triangulaire avec un bloquage logique se basant sur la polysémie du terme B du milieu. Les termes B_i et B_j sont des raffinements/usages de B.

$$\exists\, A \xrightarrow{is-a} B \quad \wedge \quad \exists\, B \xrightarrow{R} C \quad \Rightarrow \quad A \xrightarrow{R} C$$

Illustrons cela par l'exemple suivant :

$$chien \xrightarrow{is-a} canidé \wedge canidé \xrightarrow{has-part} croc \Longrightarrow chien \xrightarrow{has-part} croc.$$

La déduction est un schéma qui ne peut s'appliquer que sur les termes ayant au minimum un hyperonyme car fondé sur la relation d'hyperonimie (is-a / est-un).
Ce schéma simpliste s'est avéré naïf et a abouti à des inférences fausses à cause de la polysémie d'une part et du poids très faible de quelques relations d'une autre part. Afin de remédier à cela on a établi une stratégie de filtrage logique et statistique qui a amélioré nettement la qualité des inférences proposées.

On a appliqué le moteur d'inférence se basant sur ce schéma sur 25 000 termes ayant minimum un hyperonyme séléctionnés aléatoirement et cela a produit plus que 1 500 000 inférences avec un taux de 80-90% valides, aux alentours de 10% valides mais pas pertinentes (*arbre* $\xrightarrow{has-part}$ *atome*) et un taux réduit d'erreurs dans les prémisses.

Les articles (Zarrouk *et al.*, 2013b), (Zarrouk *et al.*, 2013c) détaillent ces travaux.

2.2 L'induction

L'induction est un schéma inverse de la déduction en ce qui est ascendant mais se base toujours sur la transitivité de la relation ontologique *is-a*. Il est considéré comme

schéma de généralisation, c'est-à-dire qu'il transfert les relations du spécifique vers le générique.

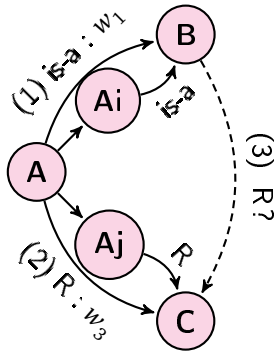

FIGURE 2 – Schéma d'inférence inductive triangulaire avec un bloquage logique se basant sur la polysémie du terme *A*. Les termes A_i et A_j sont des raffinements/usages de *A*.

Dans ce cas, si un terme *A* est un type de *B* et possède une relation *R* avec *C*, notre système va supposer que le terme *B* peut avoir la même relation *R* avec *C* (figure 2). Les mêmes stratégies de filtrage ont été appliquées tout en les adaptant au schéma.

En effectuant les mêmes expériences faites avec la déduction, nous obtenons aux alentours de 360 000 relations potentielles
Les résultats sont de 5% meilleures que celles déductives avec 80-95% de taux de validité de relations.

Plus de détails ont été fournis dans nos articles (Zarrouk *et al.*, 2013b), (Zarrouk *et al.*, 2013c) et (Zarrouk *et al.*, 2013a).

2.3 L'abduction

L'abduction est un schéma fondé sur la similarité entre des exemples. La similarité est identifiée par le partage de relations sortantes entre des termes.
Le principe de ce schéma est de supposer que les relations sortantes d'un ensemble de terme *similaires* à un terme cible *A* peuvent être valables pour ce dernier (Figure 3)
Ceci est fait en trois étapes :

– Sélectionner un ensemble de termes similaires au terme cible *A* (ensemble d'exemples) ;

– Proposer les relations pas déjà partagées entre l'ensemble et A comme potentielles pour ce dernier ;
– Présenter les relations proposées pour validation/invalidation.

Contrairement à l'induction et la déduction, l'application de l'abduction sur des termes sans hyperonyme est possible. L'abduction génère même des inférences ontologiques pouvant être utilisées par la suite par les autres schémas.

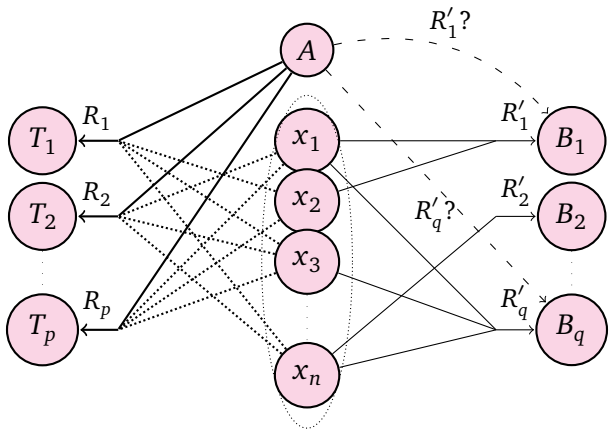

FIGURE 3 – Schéma d'abduction avec l'ensemble d'exemples x_i partageant des relations R_T avec A et identifiant de nouvelles relations abduitesR'_B.

Par exemple :
- terme cible : A
- A possède comme relations : $(bec \overset{has-part}{\longleftarrow} A)$ et $(nid \overset{location}{\longleftarrow} A)$.

- Il existe 3 exemples partageant ces relations avec A :
$(bec \overset{has-part}{\longleftarrow} \{ex_1, ex_2, ex_3\})$ $(nid \overset{location}{\longleftarrow} \{ex_1, ex_2, ex_3\})$

- On prend ces termes comme un ensemble d'exemples.

- Ces exemples possèdent des relations sortantes, ces relations vont être proposées comme relations potentielles pour A. Par exemple, si nous avons

$$\{ex_1, ex_2\} \overset{agent-1}{\longrightarrow} voler$$
$$\{ex_1, ex_2, ex_3\} \overset{has-part}{\longrightarrow} plumes$$
$$\{ex_2\} \overset{carac}{\longrightarrow} coloré$$
$$\{ex_3\} \overset{agent-1}{\longrightarrow} chanter$$

Le système infère que le terme A est éligible d'avoir ces relations et les propose pour validation/invalidation :

$A \xrightarrow{agent-1} voler\,?$ $A \xrightarrow{has-part} plumes\,?$
$A \xrightarrow{carac} coloré\,?$ $A \xrightarrow{agent-1} chanter\,?$

Naturellement l'application du schéma dans son état brut génère beaucoup de bruit (inférences erronées) et la stratégie de blocage logique et statistique joue un rôle afin d'éviter ou d'atténuer ce bruit.

On a appliqué le moteur d'inférence abductive systématiquement sur les termes contenus dans le réseau. Cela a produit 629 987 relations dont 137 416 n'existaient pas avant dans le réseau. Ces dernières concernent 10 889 entrées lexicales distinctes avec la moyenne de 12 nouvelles relations par entrée.

Ce schéma est détaillé dans un article en cours de soumission que vous trouverez joint à ce rapport(About inferences in a Crowdsourced Lexical-Semantic Network)

2.3 Schéma qui vise à enrichir les raffinements [1] et les utiliser pour faire enrichir les termes connexes

Ce schéma d'inférences est actuellement en phase de tests qui pour le moment donnent des résultats assez prometteurs.
Pour enrichir le réseau, ce schéma se base sur les raffinements des termes et les synonymes ou hyperonymes de ces termes. Cet enrichissement se fait par le transfert des relations sortantes entre le raffinement du terme et le synonyme/hyperonyme /hyponyme (ou le raffinement s'il existe) de ce terme.
Pou éclaircir le principe, on prend par exemple un terme A ayant un raffinement A' et un synonyme/hyperonyme/hyponyme B. Notre schéma vérifie qu'il existe une relation quelconque entre A' et B et si c'est le cas il propose les relations sortantes de A' pour B et celle de B pour A' (figure 4).

1. Un terme polysémique peut avoir beaucoup d'usages substantiellement différents des définitions classiquement trouvées dans un dictionnaire. Un usage donné peut aussi avoir plusieurs raffinements. Par exemple, frégate peut être un oiseau ou un bateau. Une frégate>bateau peut être distinguée comme un bateau moderne ou un navire à voiles ancien.

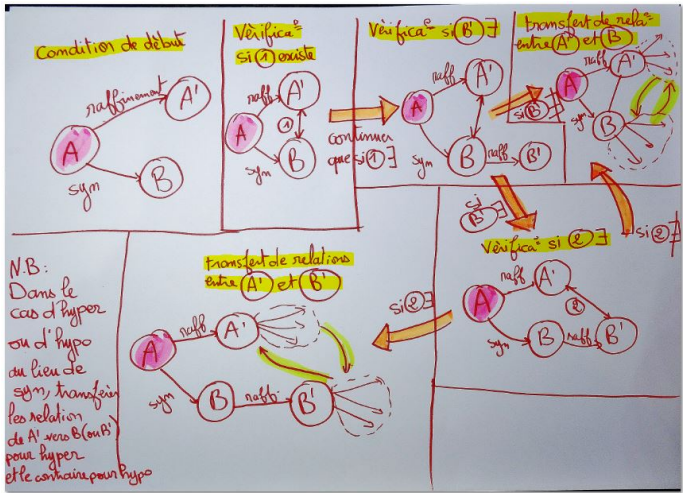

FIGURE 4 – Schéma explicatif du principe de fonctionnement du schéma

Le premier jet de résultats a généré en totalité > 300 000 relations candidates.
La figure 5 présente les taux de validité des relations générées par chaque version de ce schéma (syn/hyper/hypo).

Grib	Grib(syn)		Grib(hypo)		Grib(hyper)	
	valid	¬ valid	valid	¬ valid	valid	¬ valid
pourcentage	90.76%	9.24%	66.24%	33.76%	72.69%	27.31%

FIGURE 5 – Tableau de résultats préliminaires présentant le taux de validation des inférences de ce schéma

Ces relations proposées sont présentées comme pour tous nos schémas à un processus de validation/invalidation et réconciliation en cas de rejet.

2.4 Résumé de notre contribution

Pour résumer, le résultat de cette première année de thèse se présente sous forme d'une première version d'un système de consolidation endogène de réseau lexico-sémantique se basant sur un moteur d'inférences et un moteur de réconciliation. Le moteur d'inférences utilise des schémas d'inférences pour proposer de nouvelles relations potentielles à partir de celles déjà existantes dans le réseau. Ces relations en cas de validation sont insérées dans le réseau et dans le cas contraire sont présentées au moteur de réconciliation qui essaie de comprendre la provenance

de l'erreur dans la relation et la corriger par le biais d'un dialogue avec les utilisateurs.

Notre système, mis à part d'accomplir sa tâche principale qui consiste à consolider le réseau en densifiant les relations, s'avère être aussi un bon détecteur d'erreurs présentes dans le réseau, un classificateur par abduction, un identificateur de polysémie, marqueur d'exception et un annotateur de relations valides non pertinentes.

3 Perspectives

Ce que nous visons à faire est d'améliorer notre système qui est un premier pas vers un raisonneur autonome ou semi-autonome et de le rendre capable de découvrir automatiquement et mémoriser des règles d'inférences qui sont des schémas d'inférence à une seule inconnue et plusieurs prémisses.
Ces règles qui sont aptes à instancier une conclusion (figure 6) ou indiquer une information potentielle (figure 7).

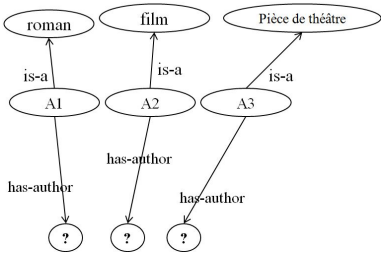

FIGURE 6

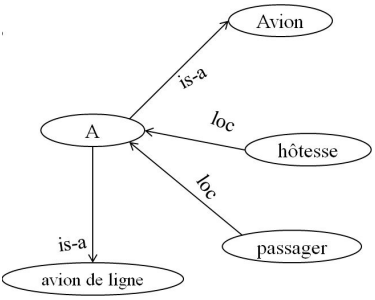

FIGURE 7

D'une autre part, nous souhaitons de conceptualiser un langage de modélisation spécifique. Les idées la dessus restent pas très claires à présent.

4 Rappel des publications

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013a). Inductive and deductive inferences in a crowdsourced lexical-semantic network. 9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013), page 6.

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013b). Inference and reconciliation in a lexical-semantic network. 14th International Conference on Intelligent Text Processing and Computational Linguistic (CILING-2013), page 13.

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013c). Inférences déductives et réconciliation dans un réseau lexico-sémantique. 20éme conférence du Traitement Automatique du Langage Naturel 2013 (TALN 2013), page 14.

LAFOURCADE, M., ZARROUK, M. et JOUBERT, A. (2013). Inférence de règles déductives par abduction. Méthodes mixtes pour l'analyse syntaxique et sémantique du français (Mixer), Atelier (TALN-2013), page 4.

5 Autres activités pendant ces 12 premiers mois

- Module doctoral : Prise de parole en public, pédagogie interactive niveau 1 (21 heures validées)
- Organisation DOCTISS 2013 journée conférence des doctorants de l'école doctorale i2s (40 heures de module doctorale validées)
- école d'été : ESSLI 2013 (The 25th European Summer School in Logic, Language and Information (ESSLLI 2013)Heinrich Heine University in Dûsseldorf, Germany, August 5-16, 2013.)
- Co-encadrement d'un groupe TER M1 (sujet :Construction d'un programme qui joue à jeuxdemots)
- Relectures CICLING 2013 et TALN 2013
- Participations aux conférences CICLING2013 à SAMOS et TALN2013 aux SABLES D'OLONNES
- Participation aux Workshops :
  • Workshop on The logic of the lexicon January 2013 : Toulouse
  • Atelier Méthodes mixtes pour l'analyse syntaxique et smantique du français (Mixer) (TALN-2013)

Références

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

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

JOUBERT, A. et LAFOURCADE, M. (2008). Jeuxdemots : un prototype ludique pour l'émergence de relations entre termes. *In proc of JADT'2008, Ecole normale supérieure Lettres et sciences humaines , Lyon, France, 12-14 mars 2008*, page 8 p.

LAFOURCADE, M., ZARROUK, M. et JOUBERT, A. (2013). Inférence de régles déductives par abduction. *Méthodes mixtes pour l'analyse syntaxique et sémantique du français (Mixer), Atelier (TALN-2013)*, page 4.

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

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 fran ?ais ? 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. et JOUBERT, A. (2013a). Inductive and deductive inferences in a crowdsourced lexical-semantic network. *9th International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, page 6.

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013b). Inference and reconciliation in a lexical-semantic network. *14th International Conference on Intelligent Text Processing and Computational Linguistic (CILING-2013)*, page 13.

ZARROUK, M., LAFOURCADE, M. et JOUBERT, A. (2013c). Inférences déductives et réconciliation dans un réseau lexico-sémantique. *20éme conférence du Traitement Automatique du Langage Naturel 2013 (TALN 2013)*, page 14.