Parcourir, reconnaître et réfléchir. Combinaison de méthodes légères pour l'extraction de relations sémantiques

Mathieu Lafourcade[1] Nathalie Le Brun[2]

(1) LIRMM, 860 rue de St Priest, 34095 Montpellier cedex 5, France
(2) Imagin@t, 34400 Lunel, France
lafourcade@lirmm.fr, imaginat@imaginat.name

RESUME

La capture de relations sémantiques entre termes à partir de textes est un moyen privilégié de constituer/alimenter une base de connaissances, ressource indispensable pour l'analyse de textes. Nous proposons et évaluons la combinaison de trois méthodes de production de relations lexico-sémantiques.

ABSTRACT

Browse, recognize and think. Combination of light methods applied to semantic relations extraction

Extracting semantic relations from texts is a good way to build and supply a knowledge base, an indispensable resource for text analysis. We propose and evaluate the combination of three ways of producing lexical-semantic relations.

MOTS-CLES : cooccurrences, inférences, patrons et relations sémantiques, combinaison

KEYWORDS: cooccurrences, inferences, semantic schemas and relations, combining.

1. Introduction

Les relations sémantiques, qu'elles soient ontologiques (hyperonymes, hyponymes, parties/tout), lexicales (synonymes), ou encore qu'elles relèvent des rôles sémantiques (agent, patient, instrument, manière, lieux, etc.) sont d'un intérêt majeur pour la quasi-totalité des applications du TALN où le système doit « comprendre » de quoi il retourne : traduction automatique, indexation, résumé, détection de textes similaires, etc. La création de procédures pour produire ce type de ressources répond donc à de multiples besoins en TALN.

Les approches peuvent être variées, manuelles comme pour WordNet, (Miller, 1995), plus ou moins automatiques (BabelNet, Navigli and Ponzetto, 2010), ou encore contributives (Lafourcade, *et al.*, 2015). De nombreuses méthodes automatiques d'extraction de relations à partir de textes ont été proposées, mais les performances restent très inégales. En effet, certaines approches se veulent très précises, ce qui suppose une analyse lexico-sémantique profonde, donc coûteuse. La vitesse de traitement des textes, et donc le taux d'extraction de nouvelles relations, s'en trouve

ralenti. A l'inverse, on trouve des méthodes statistiques qui n'incluent quasiment aucun traitement linguistique du texte d'entrée.

Nous proposons dans cet article d'évaluer les bénéfices d'une approche combinant plusieurs stratégies. Cette approche s'inscrit dans le cadre d'un apprentissage sans fin (never ended learning) au sein du projet JeuxDeMots (JDM). L'idée est de mettre en place des boucles d'extraction/exploitation où un système d'extraction automatique joue le rôle de contributeur au sein du réseau. Les joueurs, par leur activité ludique ou de contribution directe, valident ou invalident les contributions. Ainsi, il est possible de façon holistique d'évaluer l'évolution des performances de notre système Schémas – Inférences- Cooccurrences (SIC), qui alimente le réseau et l'exploite pour mener sa tâche à bien.

Dans ce qui suit, nous présentons rapidement les travaux précédents en extraction automatique de relations sémantiques, en nous limitant à ceux dont l'approche est en rapport avec notre proposition. Ensuite, nous détaillons trois stratégies d'extraction ainsi que la façon dont nous les combinons. Enfin, nous détaillons et discutons les résultats obtenus.

2. Etat de l'art

L'utilisation de schémas lexico-sémantiques a été proposée par (Hearst, 1992) pour extraire des relations de synonymie et d'hyperonymie dans des textes. Les schémas sont par exemple de la forme « A est un type de B ». (Herbelot et Copestake, 2006) ont utilisé ce genre de schémas afin d'extraire des relations dans le domaine de la biologie à partir de pages Wikipédia, avec une précision excellente (88%) mais un rappel assez faible (20%).

Dans (Ruiz-Caasado, 2005 et 2007) il est fait état d'apprentissage automatique de tels schémas de façon à extraire des relations également depuis Wikipédia, et les insérer dans Wordnet ; on note là aussi la faiblesse des performances concernant le rappel. L'approche consistant à faire appel à de l'apprentissage automatique de schémas à partir de textes a aussi été explorée par (Snow, *et al.*, 2004), également pour identifier des relations d'hyperonymie ou d'hyponymie. Dans (Girju, *et al.*, 2003) une approche supervisée vise à déterminer des contraintes sémantiques pour extraire des relations de méronymie. Les contraintes sont issues de la relation *part-of* de Wordnet et servent de données d'entraînement. Dans (Ramadier et Lafourcade, 2016) une approche similaire est présentée, à ceci près qu'un grand nombre de types de relations sémantiques sont identifiés et que les contraintes sémantiques sont déterminées manuellement.

De nombreux auteurs tentent d'extraire des relations depuis Wikipédia en exploitant les informations de structure des pages. Par exemple, (Sumida et Torisawa, 2008) ont utilisé cette stratégie sur Wikipédia en japonais pour extraire 1.4 millions de relations d'hyponymie avec une précision de 0.75. De façon similaire, (Ponzetto and Strub, 2007) exploitent les liens de catégorie de Wikipédia pour identifier des relations d'hyperonymie. Les travaux de (Pachenko, 2013) présentent une analyse en profondeur de fonctions d'évaluation de relations sémantiques entre termes. Une des conclusions est qu'aucune des mesures d'évaluation de la relation sémantique ne surpasse les autres. Ce sont ces différentes mesures de similarité qui conditionnent la proposition de telles ou telles relations sémantiques, ontologiques pour la plupart (hyperonymes, co-hyponymes, etc.)

On remarquera que peu de travaux sur la question s'appuient sur l'utilisation de bases de connaissances pour extraire de nouvelles relations sémantiques dans une approche de type apprentissage permanent en boucle. Par contre, quand il s'agit d'apprentissage automatique, de telles bases sont utilisées pour l'entraînement. De plus, la plupart des approches se limitent à des

relations ontologiques, comme l'hyperonymie (isa), la synonymie (syn) et la méronymie ou l'holonymie (has parts / is part of). Des relations comme cause/conséquence, caractéristiques, lieux, agent, patient et instrument (pour des verbes) ne font que rarement l'objet d'extraction.

Dans l'approche que nous expérimentons, nous nous focalisons sur du texte pur, en français exclusivement, issu de Wikipédia ou autre, sans exploiter la structure du document source. Nous souhaitons aussi extraire des informations de textes non encyclopédiques (comme des romans, par exemple). Pour chacune des méthodes d'extraction, nous utilisons, à des degrés différents, le réseau lexico-sémantique JeuxDeMots.

3. Combinaison de méthodes d'extraction de relations

Nous présentons trois méthodes relativement simples d'extraction de relations sémantiques entre paires de termes. Nous indiquons ensuite les grandes lignes de la combinaison de ces méthodes.

3.1. Cooccurrences et relations

La première méthode pour détecter des relations sémantiques est de construire un réseau de cooccurrences. La méthode tient compte des termes composés et comprend un prétraitement sur le texte, qui consiste à :

- Remplacer un terme par son lemme, quand le terme est un verbe conjugué et uniquement cela. Par exemple, le segment : « les poules dorment » deviendra « les poules dormir ». Par contre le segment « les poules couvent » reste inchangé car *couvent* peut être une forme du verbe *couver* mais aussi le substantif.
- Repérer les occurrences de termes composés par confrontation avec le réseau JDM. Les espaces sont remplacés par des soulignés, ce qui permet d'éviter leur segmentation.
- Les ponctuations sont conservées mais sont décollées des termes qui les précèdent (*chat,* => *chat* ,).
- Les majuscules ne sont pas modifiées ;
- En dehors de ce qui est mentionné ci-dessus, ni *pos tagging* ni analyse syntaxique ne sont réalisés.

Les termes composés sont identifiés par comparaison avec ceux existant dans le réseau JDM. En cas de conflit (par exemple, un segment A B C avec deux mots composés A_B et B_C), on applique une priorité à droite (on obtiendra A B_C). Ensuite, une segmentation est réalisée à partir des caractères espaces. Une fenêtre de k mots est utilisée pour établir les relations de cooccurrences, avec un poids décroissant de k (mot adjacent) à 1 (mot à une distance de k termes). Nous avons utilisé une fenêtre de 10 mots, afin de maximiser le rappel, ce qui est l'objet de cette méthode.

Nous utilisons la base de connaissances JeuxDeMots comme support pour la détermination des lemmes et catégories morphosyntaxiques mais aussi pour l'identification approximative des types de relations. Plus précisément, nous avons des règles de ce type, qui exploitent les parties du discours :

- Si X r_pos Verbe & Y r_pos Adv → X r_manner Y ;
- Si X r_pos Nom & Y r_pos Adj →X r_caracY ;
- Par défaut, si X est en cooccurrence avec Y → X r_assoc Y

Les règles sont strictes et doivent être entendues comme : si X est un verbe et uniquement un verbe et si Y est uniquement un adverbe alors X sera lié à Y par une relation de manière. Par exemple, soit la phrase suivante : « le chat attrapa rapidement le rat noir ». La phase de prétraitement nous fournit le texte : « le chat attraper rapidement le rat noir » (nous n'indiquons pas les poids). Ces relations sont pondérées par le poids de la cooccurrence entre les deux termes.

• chat r_assoc attraper • attraper r_manner rapidement • attraper r_assoc rat • attraper r_assoc noir • rapidement r_assoc rat • …	• attraper r_assoc rat • rat r_assoc noir • chat r_assoc rat • chat r_assoc noir • rapidement r_assoc chat

3.2. Schéma lexico-sémantique avec contraintes

Dans (Ramadier et Lafourcade, 2016) la méthode de (Hearst, 1992) et (Herbelot et Copestake, 2006) qui exploite des schémas lexico-sémantiques, a été étendue de façon à ce que les termes vérifient des relations sémantiques issues du réseau lexical de JeuxDeMots. Par exemple :

- X du Y avec X r_isa artéfact & Y r_isa personne → Y r_own X (le fusil du soldat)
- X du Y avec X r_isa partie du corps & Y r_isa personne → Y r_part X (le bras du soldat)
- X du Y avec X r_isa personne & Y r_isa lieu_humain → Y r_lieu X (la fille du coron)

Bien sûr, certains schémas ne sont pas associés à des contraintes, par exemple :

- X est localisé dans le/la/mes/un/une/des/ Y → X r_lieu Y
- X est un type de Y → X r_isa Y
- X fait partie de Y → X r_holo Y
- X se compose de Y → X r_has_parts Y

Une relation entre deux termes sera pondérée par le nombre de fois ou elle a été découverte par des schémas différents dans des segments textuels distincts.

3.3. Induction et Abduction sur un réseau lexico-sémantique

Dans (Zarrouk, *et al.*, 2014) et (Zarrouk and Lafourcade, 2015), des approches de production de nouvelles relations lexico-sémantiques par inférence ont été proposées. Ces approches sont strictement endogènes sur le réseau sémantique JeuxDeMots et se basent sur la déduction et plusieurs formes d'abduction. Aucun texte n'est donc utilisé pour cette méthode.

3.4. Combinaison de relations

La combinaison de deux méthodes consiste à ne retenir que les relations sémantiques trouvées conjointement par chacune des deux méthodes. Bien que la méthode par cooccurrences produise

des relations sous-spécifiées (relation de type *idées associées* entre les termes), qui n'ont pas d'équivalences dans les deux autres méthodes, ces relations neutres sont utilisées comme suit :

X r_t Y + X r_assoc Y → X r_t Y	Une relation neutre (de type *idées associées* r_assoc) valide une relation typée (r_t) pour la même paire de termes.

Nous combinons les approches deux à deux car une combinaison des trois approches, bien qu'augmentant la précision, réduirait trop drastiquement le nombre de relations retenues. On retiendra donc les relations produites par au moins deux des trois méthodes. Le poids de la combinaison est la moyenne géométrique des relations.

4. Expérimentation et Discussion

Nous avons fondé notre expérimentation sur le réseau lexical JeuxDeMots pour l'approche par inférences. Pour les deux autres approches, qui sont basées sur des textes, nous avons utilisé un corpus comprenant Wikipédia en français (pour les schémas et les cooccurrences) et l'œuvre d'Émile Zola via Wikisource (pour les cooccurrences). Le choix de ce corpus pour les cooccurrences s'explique par la volonté de ne pas nous limiter à des textes de nature encyclopédique, mais d'enrichir la collecte de relations sémantiques en exploitant les avantages de la littérature romanesque : offrir (1) une plus grande diversité de relations entre les termes, et (2) plus d'informations de sens commun et de la vie quotidienne.

Les relations extraites sont les suivantes : synonymie (pour les verbes, noms, adjectifs, adverbes), agent, patient, instrument, manière, a pour lieu (pour les verbes), hyperonymie, hyponymie, instance, caractéristique, a pour lieu, est un lieu pour, parties, tout, cause, conséquence (pour les noms). Certaines relations sont clairement symétriques (comme l'hyperonymie et l'hyponymie) cependant le niveau de renseignement est très différent, en particulier au niveau des poids. Dans le cas du réseau JeuxDeMots, ces relations sont volontairement non symétrisées.

La *productivité* est la capacité d'une approche à produire des relations. Nous utilisons cette mesure en lieu et place du traditionnel rappel, que nous ne sommes pas en mesure d'évaluer. En effet, nous ne disposons pas des moyens humains pour déterminer l'ensemble des relations qu'il faudrait extraire sur notre corpus. Ce travail est lourd dans la mesure où il faut lire chacun des textes. De plus, l'accord inter-annotateur n'est généralement pas très élevé (moins de 50% en moyenne). Nous prenons comme référence l'approche par inférence et nous lui attribuons une productivité de 1. En pratique, l'approche par inférence a produit environ 60 millions de relations (qui restent potentielles tant que non validées) entre le 1er novembre 2016 et le 30 mars 2017. La *précision* est la mesure classique du rapport entre les relations évaluées comme justes et l'ensemble des relations proposées. Dans notre approche, on cherche bien entendu à maximiser ce critère critique, tout en restant raisonnablement productif. Enfin, la *pertinence* est le rapport entre les relations pertinentes et les relations justes. Décider si une relation est pertinente reste relativement subjectif, mais en général les personnes interrogées (par crowdsourcing et GWAP) tombent globalement d'accord. Par exemple, la relation : *souris r_has_part atomes*, est correcte mais peu pertinente, car la propriété d'être constitué d'atomes est universelle. La pertinence a un

rapport avec la spécificité d'une relation, en général, plus une relation est spécifique à une classe de termes, plus elle est pertinente.

L'évaluation a été réalisé par deux méthodes conjointes 1) validation manuelle d'un échantillon aléatoire et 2) croisements de réponses de joueurs via JeuxDeMots. Ces évaluations se font en continu (pour les chiffres ci-dessous sur la période de novembre 2016 à mars 2017), les résultats affichés ci-dessous sont ceux de fin mars 2017. La validation manuelle consiste à proposer la relation à un joueur (jeu Askit, http://jeuxdemots.org/askit.php), ce dernier devant se prononcer sur sa validité et sa pertinence. La méthode par croisement fait appel au jeu classique JeuxDeMots où des parties sont offertes aux joueurs avec comme mot-cible le terme premier de la relation trouvée et comme consigne le type de la relation. Par exemple, si la relation *: rat r_carac noir* a été extraite, alors des parties seront proposées pour le terme *rat* et la relation *r_carac*. Si parmi les réponses se trouve le terme *noir*, alors la relation est validée. Cette approche équivaut à interroger des personnes afin de voir si la relation extraite émerge ou pas. Les joueurs peuvent passer s'ils ne savent pas. Nous avons sélectionné en priorité les relations dont le poids correspond au 2^{e} quartile (soit les 50% ayant les poids les plus élevés).

	Schémas (S)	**Inférences (I)**	**Cooccurrences (C)**
Productivité	0.37 (22 M)	1 (60 M)	3,16 (190 M)
Précision	93 %	65 %	12 %
Pertinence	88 %	75 %	47 %

Tableau 1 : Evaluation des méthodes isolées. Nous prenons la méthode I comme référence pour le nombre de relations produite en 5 mois (60 millions pour I).

La méthode d'extraction par schémas (S) lexico-sémantiques avec contraintes produit très peu de relations fausses mais elle est relativement lente. Les 7% d'erreur correspondent à l'impossibilité d'appliquer des contraintes, quand au moins un des deux termes de la relation n'est pas suffisamment renseigné. Les relations extraites sont pertinentes, ce qui est normal puisqu'elles sont directement prélevées telles quelles dans des textes. La méthode par cooccurrence, comme attendu, est très productive, très rapide, et fort peu précise (beaucoup de déchets). Les relations correctes sont pertinentes une fois sur deux. Enfin, la méthode par inférence (qui ne s'appuie que sur la base de connaissances JeuxDeMots) montre des performances correctes. Les erreurs viennent essentiellement de l'impact de la polysémie, qui vient perturber les inférences déductives et abductives. La différence de productivité entre S et C s'explique essentiellement par la différence de vitesse des deux méthodes (S lente et précise, C rapide et floue).

La combinaison des méthodes deux à deux (tableau 2) consiste à ne retenir une relation que si elle est proposée par les deux méthodes. Il s'agit donc d'une intersection entre les propositions des deux méthodes.

	S+I	**S+C**	**I+C**
Productivité	0.22	0.35	0.78
Précision	96 %	94 %	87 %
Pertinence	93 %	84 %	88 %

Tableau 2 : Evaluation des méthodes par paires. S correspond à la méthode fondée sur les Schéma lexico-sémantiques, I à l'approche par Inférences, et C à l'approche par Cooccurrences.

Nous constatons que pour chaque paire la productivité baisse par rapport à celle obtenue pour chacune des méthodes prise isolément, ce qui est un résultat attendu. La combinaison S+I a une très faible productivité ce qui indique que peu de relations sont produites à la fois par les méthodes S et I.

L'approche par union des trois méthodes (tableau 3), permet non seulement d'obtenir une augmentation de productivité de 28 % par rapport à la méthode I (prise comme référence), mais surtout d'augmenter la précision et la pertinence. Si nous prenons comme référence la méthode I, la combinaison des approches permet de renforcer la précision avec la méthode S et la pertinence avec S et C. Combiner S+C permet (toujours en référence à I) de rajouter des relations qui n'ont pu être inférées. On rappelle que C est appliqué sur un corpus de textes plus large et plus général que la méthode S.

	(S+I) ∪ (S+C) ∪ (I+C)
Productivité	1.28
Précision	99.4 %
Pertinence	96 %

Tableau 3 : Évaluation de l'approche retenant les relations proposées par au moins 2 méthodes (union). La précision de 99.4 est supérieure à la meilleure précision du tableau 2, car en effet la combinaison de méthodes rend cela possible (union des paires de méthodes).

5. Conclusion

Nous avons présenté trois méthodes d'identification de relations sémantiques entre termes. L'une d'elles est strictement endogène par rapport à une base de connaissances (le réseau JeuxDeMots dans notre cas), les deux autres se fondent sur des textes, mais exploitent également des informations sémantiques externes aux textes. Globalement, ces méthodes sont légères. Individuellement, elles ont des défauts (productive mais imprécise / précise mais peu productive). Leur union, en compensant leurs défauts respectifs, permet d'améliorer significativement productivité, pertinence, et précision.

L'approche proposée est générique et peut être appliquée à d'autres bases de connaissances, dans d'autres langues ou pour des domaines de spécialités. Toutefois dans ce dernier cas, il faut disposer d'un corpus de textes suffisamment important afin d'être capable d'utiliser des schémas lexico-sémantiques et d'extraire des cooccurrences dans des proportions statistiquement significatives.

Les relations strictement sémantiques (et non lexicales) étant indépendantes des langues, une piste d'amélioration du processus serait d'adapter cette approche à l'extraction de relations depuis des textes de langues différentes, en ayant recours soit à un processus de traduction, soit au réseau JDM multilingue actuellement à l'étude.

Références

GIRJU, R., BADULESCU, A., and MOLDOVAN, D. (2003) *Learning semantic constraints for the automatic discovery of part-whole relations.* In Proc. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology , NAACL '03, pages 1–8. Association for Computational Linguistics, 2003.

HEARST, M. A. (1992) *Automatic acquisition of hyponyms from large text corpora.* In Proc . 14th Conf. on Computational Linguistics, COLING '92, pages 539–545. Association for Computational Linguistics, 1992.

HERBELOT, A. and COPESTAKE, A. (2006) *Acquiring ontological relationships from Wikipédia using RMRS.* In Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies , 2006.

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, 8 p.

LAFOURCADE, M., LE BRUN N., and JOUBERT A. (2015) *Games with a Purpose (GWAPS)*, ISBN: 978-1-84821-803-1 July 2015, Wiley-ISTE, 158 p.

MILLER, G. A. (1995) *Wordnet: A lexical database for English.* Communications of the ACM , 38(11):39–41, November 1995.

NAVIGLI, R.and PONZETTO, S. P. (2010*) Babelnet: Building a very large multilingual semantic network.* In Proc. 48th Annual Meeting of the Association for Computational Linguistics , ACL'10, pages 216–225, 2010.

PANCHENKO, A. (2013) *Similarity Measures for Semantic Relation Extraction*. PhD Dissertation, Université catholique de Louvain & Bauman Moscow State Technical University, 193 p.

RAMADIER, L. ET LAFOURCADE, M. (2016) *Patrons sémantiques pour l'extraction de relations entre termes - Application aux comptes rendus radiologiques.* In 23rd French Conference on Natural Language Processing (JEP-TALN-RECITAL 2016), Paris, France, 4-8 July 2016, 6 p.

RUIZ-CASADO M., ALFONSECA E., AND CASTELLS P. (2005) *Automatic extraction of semantic relationships for wordnet by means of pattern learning from Wikipédia.* In Proc. 10th Int. Conf. Natural Language Processing and Information Systems, NLDB'05, pages 67–79. Springer, 2005

SNOW R., JURAFSKY D., AND ANDREW Y. NG. (2004) *Learning syntactic patterns for automatic hypernym discovery.* In Advances in Neural Information Processing Systems (NIPS), 8 p. 2004.

SUMIDA, A. AND TORISAWA, K. (2008) *Hacking Wikipédia for hyponymy relation acquisition.* In Proc. of IJCNLP 2008 , pages 883–888, 2008.

ZARROUK, M., LAFOURCADE, M., and JOUBERT A. (2014). About Inferences in a Crowdsourced Lexical-Semantic Network, *EACL 2014 (14th Conference of the European Chapter of the Association for Computational Linguistics)*, Gothenburg (Sweden), April 2014

ZARROUK, M. and LAFOURCADE, M. (2014) *Inferring Knowledge with Word Refinements in a Crowdsourced Lexical-Semantic Network.* In proc of the the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Irlande, 9 p.