Caractérisation de marqueurs de relations par classification non supervisée

mots-clés : termes, marqueurs de relations, relation de cause, terminologie, classification, variation

Dans ce travail nous proposons une étude des marqueurs de relation en français, sous l'angle d'une approche non supervisée. Comme leur nom l'indique, les marqueurs de relation sont définis comme des éléments lexicaux, syntaxiques ou dispositionnels (Meyer 2001) qui permettent d'expliciter, de manière quasi-systématique, une relation sémantique entre deux éléments lexicaux. Bien que souvent interrogée, la systématicité du lien marqueur/relation a rarement été étudiée (Aussenac-Gilles & Condamines 2012). Dans le domaine de la terminologie, les travaux existants proposent une combinaison d'analyse lexicométrique et d'analyse manuelle de l'influence de paramètres tels que la langue, le domaine ou le genre textuel sur le fonctionnement linguistique des marqueurs de relation (notamment Marshman *et al*. 2008, Condamines 2008). Dans le domaine du Traitement Automatique des Langues, les travaux ont pour objectif d'extraire de manière quasi-automatique des relations sémantiques pour la recherche d'information, et ne prennent pas toujours en compte les paramètres de variation évoqués plus haut (par exemple, Girju *et al*. 2006, ou encore Wang *et al*. 2013, Lafourcade & Ramadier 2016). Partant de ce constat, notre étude propose d'interroger, sous un nouvel angle, l'autonomie des marqueurs de relation de cause, c'est-à-dire leur capacité à indiquer, seul, une relation causale entre deux éléments lexicaux. Plus précisément, nous proposons une étude automatique visant à améliorer, pour chaque entrée, l'évaluation de l'autonomie du candidat-marqueur dans l'expression de la relation de cause.

Nous nous appuyons pour cela sur une ressource de candidats-marqueurs de relations de cause obtenue à partir d'une analyse fondée sur des outils de linguistique de corpus (Lefeuvre 2017). Cette ressource comporte plus de 9000 entrées annotées manuellement : l'autonomie de chaque candidat-marqueur de cause recensé a été évaluée sur un corpus variant du point du vue du domaine (Volcanologie vs. Oncologie) et du genre textuel (Scientifique vs. Vulgarisé). De plus, les candidats-marqueurs sont organisés selon une classification, déterminée *a priori*, des catégories de la cause (*création*, *modification*, *empêchement*, …). Les annotations obtenues ont montré que : (I) l'évaluation de l'autonomie des marqueurs de relation ne peut être simplement binaire ; il y a une gradation, et que (II) l'annotation nécessite une expertise linguistique importante ainsi que l'exploration d'un contexte large.

Pour ces raisons, nous explorons les candidats-marqueurs et leurs contextes au moyen de techniques d'apprentissage non-supervisé (ou clustering). Il s'agit de faire ressortir les caractères particuliers qui affectent l'évaluation de l'autonomie du marqueur en fonction d'un contexte d'énonciation.

Nous utilisons un certain nombre d'indices propres à chaque occurrence des candidats-marqueurs : le domaine et le genre du texte source, la classe du marqueur (sa catégorie) et la structure syntaxique des contextes gauches et droits. Pour identifier, parmi tous les indices disponibles, les indices les plus pertinents, nous opérons en deux phases.

Dans un premier temps, nous utilisons des techniques de classification supervisée pour déterminer les indices les plus fiables pour caractériser l'autonomie des candidats-marqueurs. Puis, dans un second temps, nous exploitons ces indices pour effectuer une classification non supervisée (clustering) des contextes et faire ressortir des éléments d'analyse plus riches linguistiquement que de simples "features". Notre objectif est de faire émerger des indices saillants permettant d'expliquer certaines caractéristiques des candidats-marqueurs. Par exemple, les candidats-marqueurs de cause « amener (à) » et « contenir » sont en eux-mêmes des marqueurs peu autonomes. Ces deux candidats-marqueurs partagent des caractéristiques communes : (I) les termes en relation ne prennent pas une vraie valeur terminologique (ils sont ambigus), et (II) les candidats-marqueurs apparaissent majoritairement dans des textes de vulgarisation.

En cherchant à généraliser ces observations sur le corpus, nous remarquons effectivement que le genre textuel est un paramètre important à prendre en compte pour mesurer l'autonomie d'un marqueur de relation, et que l'on ne peut se fier simplement à un extracteur terminologique pour considérer comme termes les éléments mis en relation. Il apparaît nécessaire de distinguer, pour un même candidat-terme, les occurrences qui sont véritablement terminologiques de celles qui ne le sont pas (Judea et al. 2014, Daille et al. 2016). En fonction du contexte, un candidat-terme va prendre ou non sa valeur terminologique. De fait, les occurrences non-terminologiques n'amènent pas du tout le même usage des candidats-marqueurs, et ont une influence sur l'interprétation sémantique des verbes présents dans les marqueurs lexico-syntaxiques. Pouvoir distinguer les occurrences terminologiques et non-terminologiques des candidats-termes pourrait permettre de mieux catégoriser les candidats-marqueurs ayant un fonctionnement similaire à celui de « amener (à) » et « contenir ». Pour d'autres marqueurs, le cas est différent. Par exemple, le verbe « stopper » est un candidat-marqueur très autonome dans les corpus étudiés : il est autonome ou quasi-autonome dans 80% des cas (20 occurrences). Le verbe « libérer » présente les mêmes caractéristiques avec, en plus, une couverture plus grande (146 occurrences). A l'opposé, certains marqueurs ont une autonomie très marquée et qui transcende les différences entre genres. Il en est ainsi des marqueurs de modification qui sont autonomes dans plus de 90 % des cas, de même que « résulter de ».

Nous proposons ci-dessous les statistiques des 20 marqueurs les plus fréquents (sur 268 marqueurs au total) de la ressource utilisée pour notre étude. Pour chacun des marqueurs, nous indiquons l'effectif dans la ressource (nombre d'exemples annotés) ainsi que la proportion d'occurrences pour chaque niveau d'autonomie (de 4, occurrence très autonome à 0, occurrence très peu autonome). Dans le cadre d'une présentation, nous pourrons présenter les résultats complets de notre étude.

MARQUEUR	**Nbr.**	**Auto = 4**	**Auto = 3**	**Auto = 2**	**Auto = 1**	**Vulgarisation**	**Scientifique**
entraîner	308	0.529	0.279	0.052	0.14	0.594	0.406
résulter de	228	0.754	0.224	0.022	0	0.469	0.531
augmentation	226	0	1.0	0	0	0.257	0.743
contenir	224	0	0.022	0	0.978	0.554	0.446
provenir de	207	0.256	0.111	0.014	0.618	0.44	0.56
diminuer	201	0.343	0.498	0.07	0.09	0.632	0.368
diminution	188	0.011	0.989	0	0	0.346	0.654
aide	184	0	0.842	0.071	0.087	0.696	0.304
augmenter	175	0.469	0.423	0.091	0.017	0.509	0.491
réduire	171	0.357	0.363	0.164	0.117	0.404	0.596
provoquer	161	0.646	0.317	0.025	0	0.615	0.385
développement	157	0.019	0.981	0	0	0.535	0.465
créer	148	0.486	0.324	0.162	0.027	0.743	0.257
libérer	136	0.603	0.294	0.044	0.059	0.713	0.287
limiter	131	0.214	0.412	0.191	0.183	0.534	0.466
modifier	128	0.391	0.469	0	0.008	0.453	0.547
reposer sur	122	0.123	0.27	0.016	0.59	0.279	0.721
modification	119	0	1.0	0	0	0.571	0.429
affecter	118	0.525	0.381	0.042	0.051	0.415	0.585
développer	117	0	0.06	0.419	0.521	0.547	0.453

Bibliographie

Aussenac-Gilles, N., & Condamines, A. (2012). Variation and Semantic Relation Interpretation : Linguistic and Processing Issues. In *Proceedings of TKE 2012 : Terminology and Knowledge Engineering*, 106–122. Madrid, Espagne: Aguado de Cea et al.

Condamines, A. (2008). Taking genre into account when analysing conceptual relation patterns. *Corpora*, 8, 115-140.

Daille, B., Jacquey, E., Lejeune, G., Melo, L.F. & Toussaint, Y. (2016). Ambiguity Diagnosis for Terms in Digital Humanities. *10th edition of the Language Resources and Evaluation Conference (LREC 2016),* 23-28 May 2016, Portorož (Slovenia).

Girju, R., Badulescu, A., & Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1), 83-135.

Judea, A., H. Schutze, H. & Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics.*

Lafourcade M. & Ramadier, L. (2016) Semantic Relation Extraction with Semantic Patterns Experiment on Radiology Reports. *10th of the Language Resources and Evaluation Conference (LREC 2016)*, 23-28 May 2016, Portorož (Slovenia).

Lefeuvre, L., (2017). Analyse des Marqueurs de Relations Conceptuelles en Corpus Spécialisé : Recensement, Évaluation et Caractérisation en Fonction du Domaine et du Genre Textuel. Thèse de doctorat en Sciences du Langage, Université Toulouse – Jean Jaurès.

Marshman, E., L'Homme, M.-C., & Surtees, V. (2008). Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora*, 3(2), 141-172

Meyer, I. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In D. Bourigault, M.-C. L'homme & C. Jacquemin, eds. *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp.279-302.

Wang, W., Besançon, R., Ferret, O., & Grau, B. (2013). Regroupement sémantique de relations pour l'extraction d'information non supervisée. In *Actes de la 20eme Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013),*17-21 juin 2013, Les Sables d'Olonne, France, pp. 353-366.