TALN 2004, Session Poster, Fès, 19–21 avril 2004

Repérage de relations terminologiques transversales en corpus

Natalia Grabar[1,2,3], Véronique Malaisé[1,4], Aurélia Marcus[5], Aleksandra Krul[3]
[1] STIM/Assistance Publique – Hôpitaux de Paris, ERM 202 INSERM
[2] Département de Biomathématiques, Université Paris 6
[3] CRIM/INaLCO
[4] Université Paris 7 & DRE de l'Institut National de l'Audiovisuel
[5] Sinequa
ngr@biomath.jussieu.fr, vmalaise@ina.fr, aurelia_m@noos.fr, ola.krul@voila.fr

Résumé - Abstract

Les relations transversales encodent des relations spécifiques entre les termes, par exemple `localisé-dans`, `consomme`, etc. Elles sont très souvent dépendantes des domaines, voire des corpus. Les méthodes automatiques consacrées au repérage de relations terminologiques plus classiques (hyperonymie, synonymie), peuvent générer occasionnellement les relations transversales. Mais leur repérage et typage restent sujets à une conceptualisation : ces relations ne sont pas attendues et souvent pas connues à l'avance pour un nouveau domaine à explorer. Nous nous attachons ici à leur repérage mais surtout à leur typage. En supposant que les relations sont souvent exprimées par des verbes, nous misons sur l'étude des verbes du corpus et de leurs divers dérivés afin d'aborder plus directement la découverte des relations du domaine. Les expériences montrent que ce point d'attaque peut être intéressant, mais reste pourtant dépendant de la polysémie verbale et de la synonymie.

Transversal relations describe specific information existing between terms, for instance `consumes`, `located-in`, etc. They are often dependent on domains and even on corpora. Automatic methods, conceived to detect classical terminological relations (hyperonymy, synonymy), can occasionnally generate transversal relations. But their detection and typology depend on their conceptualisation : these relations are not expected and often not known for a newly explored domain. We aim here at their detection, but mainly at their typology. Since we suppose these relations are often expressed with verbs, we concentrate our investigation on the study of verbs and their derivatives to attack directly their discovering. Experiences show that this approach proposes interesting results which are nevertheless dependent on verbal polysemy and synonymy.

Mots-clefs – Keywords

Terminologie, corpus spécialisés, structuration de terminologies, relations transversales, verbes
Terminology, Specialised Corpora, Terminology Structuring, Transversal Relations, Verbs

1 Introduction

Les produits terminologiques ont la vocation de décrire la connaissance d'un secteur d'activité. Cette description peut être plus ou moins détaillée et fine : elle va d'une liste de termes « à plat » à une terminologie structurée. La structure d'une terminologie peut être assurée avec trois types de relations (Grabar & Hamon, 2004) :

- Les *relations hiérarchiques* (ou `est-un`) :

  pneumonie `est-un` *bronchopneumonie*
  bronchopneumonie `est-un` *maladie de l'appareil respiratoire*

  Notons que la *relation partitive* (*méronymie* ou `partie-tout`), qui sert souvent à décrire les artefacts et les organismes vivants à travers l'énumération de leurs parties constituantes, peut également assurer un rôle structurant à côté de la relation hiérarchique :

  poumon `partie-de` *appareil respiratoire*

- Les *relations lexicales* : la *relation d'équivalence*, ou de *synonymie*, relie les termes équivalents : *pneumonie* et *pneumopathie inflammatoire* ; l'*antonymie* relie les termes signifiant les notions opposés ou contraires.
- Les *relations transversales* relient les concepts qui se trouvent dans différentes branches hiérarchiques de l'arbre du domaine. Elles peuvent avoir une portée définitoire : un diagnostic (*pneumonie*) est localisé dans une partie du corps (*poumon*) et subit une atteinte morphologique (*inflammation*) :

  pneumonie → `localisé-dans` → *poumon*
  ↘ `atteint-par` → *inflammation*

De ces trois types de relations, les relations transversales, étant dans la majorité des cas sous-spécifiées, sont les plus occultées dans les produits terminologiques. Elles regroupent pourtant plusieurs types de relations sémantiques potentiellement utiles en description de la connaissance du domaine. Ces relations sont exprimées par les verbes ou bien articulées autour d'eux. Nous distinguons les relations transversales suivantes :

- Les *relations prédicatives ou actancielles* reflètent l'organisation syntaxique dans les énoncés et concernent les rôles des groupes nominaux (termes potentiels) autour des verbes (`objet`, `acteur`, `patient`, `instrument`, etc.) ;
- Les *relations causales*, que nous distinguons à part entière, car elles sont potentiellement présentes dans tout domaine et tout type de corpus ;
- Les *relations domaniales* décrivent les relations désignant les actes, les processus ou autres, spécifiques au domaine et au corpus. Leur transposition sur un autre domaine ou corpus n'est souvent pas possible, sinon sujette au changement de leur sémantique.

Face aux travaux consacrés au repérage des relations hiérarchiques (Hearst, 1992; Morin, 1999) ou synonymiques (Hamon *et al.*, 1998; Jacquemin, 1997), les relations domaniales ne sont pas étudiées. La situation est d'autant plus compliquée qu'elles sont dépendantes des domaines, des corpus et même des applications. Nous consacrons donc ce travail au repérage de relations transversales et, en particulier de relations domaniales. La principale question que nous nous posons est de savoir s'il est possible de faciliter cette tâche et, si oui, comment. Nous commençons par une présentation des approches en structuration de terminologies qui peuvent déboucher sur les relations transversales ou bien être adaptées à leur recherche (sec. 2). Nous présentons ensuite la méthode suivie (sec. 3) et le matériel requis (sec. 4) ; puis analysons et discutons les résultats (sec. 5) et tirons les conclusions (sec. 6).

2 Présentation et discussion des travaux en repérage automatique de relations transversales

Relations prédicatives ou actancielles. Le repérage de relations prédicatives ou actancielles peut profiter de l'existence des analyseurs syntaxiques. Les outils comme `Shallow parser` de Xerox ou `LinkParser`[1] (Brian, 2000) effectuent une analyse syntaxique des phrases et détectent les rôles sémantiques. Dans la phrase : *« Pour combattre les effets des déchirures dans la couche d'ozone la terre est recouverte d'une serre alimentée par un rayon laser. »*, le `Shallow parser` indique que *effets* est l'objet du verbe *combattre*, que *ozone* est un objet prépositionnel de *couche*, etc.

La morphologie constructionnelle (Corbin, 1987) conduit également à la détection de ces relations. Par exemple, *-eur* forme des noms d'agent à partir des verbes : {transport routier, transporteur routier} (Daille, 2003).

Relation causale. Les verbes de la relation causale ont été recensés et classés dans le travail de (Garcia, 1997); ils peuvent ainsi servir de marqueurs lexicaux pour la détection de différents types de causalité.

Relations domaniales. Par définition, les relations domaniales sont dépendantes des domaines. Les approches automatiques que nous présentons ici peuvent occasionnellement générer des relations domaniales, sinon suite à une adaptation. Les modèles distributionnelles et d'association (approches *a posteriori*), donnent des résultats potentiellement riches en relations transversales. Ainsi, l'analyse distributionnelle appliquée par (Grefenstette, 1994) peut faire émerger plusieurs types de relations : antonymie (*large*, *small*) ; synonymie (*large*, *important*, *great*) ; méronymie (*patient*, *group*) ; hyperonymie (*patient*, *child*, *woman*) et des relations domaniales (*patient*, *treatment*). Les règles d'association (Toussaint & Simon, 2000) proposent les relations d'hyperonymie (*histamine* `est-un` *biogenic amine*), de co-hyponymie (*spermidine* `est-un-frère` *putrescine*) et domaniales (*acids* `se-transforment-en` *esters*, *silica* `utilisé-pour` *chromatography*). Le typage de ces relations est fait manuellement ; pour un typage automatique l'utilisation de ressources sémantiques externes est nécessaire, mais s'avère insuffisante (Nazarenko *et al.*, 2001).

La projection de marqueurs lexicaux et de patrons lexico-syntaxiques est une approche *a priori*, elle est basée sur une étude initiale des corpus et du domaine. Le travail de (Garcia, 1997) propose des marqueurs pour le repérage de la causalité. (Séguéla & Aussenac, 1997) définissent de nouveaux patrons, spécifiques aux domaines et corpus étudiés, et y adaptent d'anciens patrons.

L'étude de la forme interne des termes peut conduire au repérage de relations de synonymie (Hamon *et al.*, 1998; Jacquemin, 1997), et de relations hiérarchiques (Grabar & Zweigenbaum, 2003). Dans ce dernier cas, il s'agit de l'inclusion lexicale des termes : le terme incluant est l'hyponyme du terme inclus (*sténose de l'aorte* `est-un` *sténose*). Lorsque le terme inclus se trouve en position de tête syntaxique, il s'agit potentiellement d'une relation hiérarchique, sinon la relation peut « cacher » une relation transversale (Grabar & Zweigenbaum, 2003). En reprenant l'exemple précédent : *sténose de l'aorte* `localisé-dans` *aorte*.

Avec les approches citées, les relations domaniales sont les plus hasardeuses au repérage. De plus, l'accès à ces relations n'est pas direct mais sujet à l'étude du domaine et du corpus.

[1] Disponible à l'adresse `http://search.cpan.org/dist/Lingua-LinkParser/`.

3 Méthode

Face à ces difficultés de repérage de relations domaniales, nous nous demandons si les données elles-mêmes ne peuvent pas guider, de manière (semi)automatique, la tâche. Nous restons sur l'hypothèse que ces relations sont articulées autour des verbes et des noms déverbaux d'action. Les noms déverbaux sont repérés par rapport à leurs suffixes dérivationnels les plus productifs et fréquents : *-ation* {*cogénérer*, *cogénération*}, *-ment* {*développer*, *développement*} et *-age* {*relever*, *relevage*}.

Nous explorons deux pistes dans le repérage de relations domaniales et vérifions à chaque fois si les relations déjà découvertes dans notre corpus lors d'une étude précédente, correspondent aux verbes qui y sont usités. La première prévoit un investissement minimal et consiste en un examen de la liste des verbes du corpus. La deuxième piste cerne les verbes d'une manière plus contrainte. Elle est inspirée des travaux d'Emmanuel Morin (Morin, 1999) : à partir d'une liste de termes en relations transversales projetés sur le corpus nous repérons et analysons les patrons qui leur correspondent. Nous utilisons les ressources sémantiques (séries de synonymes) provenant d'un dictionnaire de langue générale (le Robert d'une édition des années 70) afin d'établir les liens sémantiques entre différents verbes pouvant être considérés comme équivalents. Si la correspondance entre les verbes utilisés dans le corpus et les verbes exprimant les relations déjà validées existe, l'acquisition des relations domaniales pourrait être guidée par les verbes et les noms déverbaux d'action qu'un corpus contient.

4 Préparation et sélection du matériel

Corpus. Le corpus sur la cogénération contient les documents provenant de l'Internet, des bases de données internes d'EDF et des actes de conférences. Tous ces documents ont été triés et filtrés. L'étiquetage morphosyntaxique est effectué avec `TreeTagger` et corrigé avec `Flemm`. Le corpus contient 772 003 occurrences.

Termes en relations transversales. L'analyse et validation (1) des relations de synonymie générées avec Synoterm (Hamon *et al*., 1998), (2) de relations hiérarchiques générées à travers les inclusions lexicales (Grabar & Zweigenbaum, 2003) et (3) de relations hiérarchique, méronymiques et lexicales trouvées avec les patrons lexico-syntaxiques de (Séguéla & Aussenac, 1997), nous ont conduit finalement à un ensemble de 19 relations terminologiques. Parmi les relations domaniales détectées, nous sélectionnons pour l'analyse les relations suivantes :

- `conduit` (12 paires de termes) : l'objet nommé par le 1^{er} terme effectue le transport ou la conduite de la substance nommée par le 2^{eme} terme :
  réseau de transport de l'électricité `conduit` *électricité*,
  conduit d'échappement `conduit` *gaz d'échappement*.
- `consomme` (52 paires de termes) : l'objet nommé par le 1^{er} terme consomme la substance nommée par le 2^{eme} terme :
  centrale bagasse `consomme` *bagasse*, *microturbine* `consomme` *biogaz*.
- `produit` (234 paires de termes) : signifie des processus de production concrète (naturelle ou artéfactuelle) ou abstraite :
  organisme d'accréditation `produit` *accréditation*,
  digesteur `produit` *biogaz*,
  cogénération `produit` *chaleur*.

5 Résultats et discussion

Liste des verbes du corpus. L'étiquetage morphosyntaxique montre que le corpus contient 1 922 verbes (sans épurer les erreurs d'étiquetage). Nous en sélectionnons 1 028, qui apparaissent au moins quatre fois : ces verbes très ou moyennement fréquents sont peut-être caractéristiques du domaine. Leur analyse peut conduire aux types suivants : production de l'énergie (*produire*, *réaliser*, *appliquer*, ...), sa distribution (*fournir*, *conduire*, *consommer*, *alimenter*, *accorder*, ...), etc. Par contre, si nous replaçons ces verbes dans leurs contextes, il s'avère (1) qu'à côté des noms déverbaux obtenus par suffixation, on doit également examiner ceux obtenus par conversion {*conduire/V*, *conduit/N, conduite/N*} ; (2) que les verbes sont polysémiques et ne peuvent pas être utilisés tels quels.

Projection de termes en relations domaniales sur le corpus. Les paires de termes en relations domaniales projetées sur le corpus n'ont pas toutes donné de contextes verbaux ou déverbaux. Ainsi, les paires de termes qui n'apparaissent pas dans l'espace phrastique, ne peuvent pas conduire à l'établissement de patrons lexico-syntaxiques.

La relation `produit` est signifiée souvent par le verbe *produire* et ses nombreux dérivés, que le patron lexico-syntaxique devrait prendre en compte : les noms *production*, *producteur*, *produit*, les formes verbales *produire*, *produite*, *produisant*, *(il) produit*, *(ils) produisent*, *(ils) produisaient*. Tel n'est pas le cas pour le verbe *consommer* et ses dérivés, qui ne sont pourtant pas polysémiques dans le corpus. Ils se retrouvent souvent à proximité des termes « consommables » (*énergie*, *gaz*, *électricité*), mais ne marquent pas, dans les contextes relevés, cette relation. Dans d'autres cas, ces relations sont exprimées avec des verbes ou des expressions verbales « synonymes », comme par exemple pour la relation `consomme` : *dispose*, *utilise*, *fonctionne avec*. Les séries de synonymes de Le Robert n'indiquent que très partiellement ces relations synonymiques. Ce qui n'est pas surprenant : dans un corpus spécialisé, les moyens d'expressions divergent de ceux existant dans la langue générale. La constitution de classes de verbes avec une approche distributionnelle pourrait fournir des équivalences entre eux plus valables pour un corpus. D'autres relations, non décrites ici, ont un comportement similaire ou bien n'ont pas fourni de résultats intéressants. De manière générale, cette approche bute à la polysémie et la synonymie verbale. Le repérage et le typage de relations domaniales s'avèrent donc être délicats et nécessitent un travail supplémentaire en désambiguïsation des verbes.

6 Conclusion et perspectives

Nous avons présenté des expériences de repérage et de typage de relations domaniales en corpus de spécialité. Ces expériences sont basées sur l'exploitation des verbes et de leurs dérivés. Il apparaît ainsi que certaines relations (comme `produit`) sont exprimées par les verbes correspondants et leurs dérivés. Qui plus est, ces verbes se retrouvent parmis les plus fréquents du corpus et peuvent être considérés comme spécifiques. Mais pour la plupart des relations, ce sont les expressions et verbes équivants qui sont utilisés. La découverte de ces équivalents demandent l'utilisation de ressources sémantiques. Nous avons exploité les ressources provenant de la langue générale (séries de synonymes de Le Robert) et constaté qu'elles ne facilitent pas de beaucoup la découverte de relations domaniales. De manière plus sûre, ces expressions émergent avec la découverte du corpus et du domaine. D'autres relations encore ne peuvent pas être exprimées avec les patrons lexico-syntaxiques : les termes correspondants n'apparaissent pas à

l'intérieur des segments (phrases) à partir desquels les patrons pourraient être définis.

De manière générale, les relations domaniales sont plutôt imprévisibles lors de l'étude d'un nouveau domaine, et les verbes n'apportent qu'une solution partielle dans leur repérage et typage. Ils proposent par contre des informations qui peuvent permettre d'aborder assez rapidement les relations domaniales. À côté des méthodes automatiques déjà existantes, les verbes constituent une aide supplémentaire.

Références

BRIAN D. (2000). Parsing natural language with lingua::linkparser. *Perl Journal*, **19**, 60–66. Disponible à `http://www.improvist.com/Projects/Technology/Papers/LinkParser/`. Visité le 31/12/2003.

CORBIN D. (1987). *Morphologie dérivationnelle et structuration du lexique*, volume 1. Lille: Presse universitaire de Lille.

DAILLE B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16.

GARCIA D. (1997). Structuration du lexique de la causalité et réalisation d'un système d'aide au repérage de l'action dans les textes. In *Terminologie et Intelligence Artificielle (TIA)*, p. 7–26, Toulouse.

GRABAR N. & HAMON T. (2004). Les relations dans les terminologies structurées : de la théorie à la pratique. *Revue d'Intelligence Artificielle (RIA)*, (1). Sous presse.

GRABAR N. & ZWEIGENBAUM P. (2003). Lexically-based terminology structuring. In *Terminology*. Sous presse.

GREFENSTETTE G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *Proceedings of 17th International Conference on Computational Linguistics (COLING-ACL'98)*, p. 498–504, Université de Montréal, Montréal, Quebec, Canada.

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France. Disponible à `http://sern.ucalgary.ca/KSI/KAW/KAW99/papers/Gery1/index.html`. Visité le 26/08/99.

JACQUEMIN C. (1997). Guessing morphology from terms and corpora. In *ACM SIGIR*.

MORIN E. (1999). Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues (TAL)*, **40**(1), 143–166.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. & BOUAUD J. (2001). Corpus-based extension of a terminological semantic lexicon. In *Recent Advances in Computational Terminology*, p. 327–351. Benjamins.

SÉGUÉLA P. & AUSSENAC N. (1997). Un modèle de base de connaissances terminologiques. In *Terminologie et Intelligence Artificielle (TIA)*, p. 47–68, Toulouse.

TOUSSAINT Y. & SIMON A. (2000). Building and interpreting term dependencies using association rules extracted from galois lattices. In *Recherche d'Information Assistée par Ordinateur (RIAO)*, p. 1686–1692, Paris.