Comment guider le travail de normalisation terminologique?

Nouha Omrane, Adeline Nazarenko, Sylvie Szulman

Comment guider le travail de normalisation terminologique?

Nouha Omrane, Adeline Nazarenko, Sylvie Szulman

LIPN (LABORATOIRE D'INFORMATIQUE DE PARIS NORD)
UNIVERSITÉ PARIS 13 & CNRS (UMR 7030)
nouha.omrane@lipn.univ-paris13.fr
adeline.nazarenko@lipn.univ-paris13.fr
sylvie.szulman@lipn.univ-paris13.fr

Résumé :
La normalisation d'un réseau terminologique est une étape importante de la construction d'ontologies dans la méthode TERMINAE ou simplement pour la construction de thesaurus. Nous montrons comment ce travail de normalisation peut être guidé par des critères de pertinence relatifs au domaine et au discours analysé.
Mots-clés : critères de pertinence des termes, normalisation sémantique, réseau terminologique

La méthode TERMINAE (Aussenac-Gilles *et al.*, 2008) repose sur trois niveaux de connaissances terminologique, termino-conceptuel et conceptuel pour la construction d'ontologies à partir de textes. Le niveau terminologique représente le vocabulaire mentionné dans les documents et sert de point de départ pour la construction d'un modèle du domaine. La première étape de normalisation permet de transformer le réseau terminologique initial en un réseau termino-conceptuel. C'est un réseau de termes non ambigus dont les noeuds sont interconnectés à travers des relations taxonomiques et sémantiques. Il décrit une terminologie du domaine stabilisée et documentée qui peut servir de base pour la construction d'une ontologie de domaine à partir de textes ou de thesaurus pour l'annotation des documents. Une deuxième étape, de formalisation, permet ensuite de transformer le réseau termino-conceptuel en un réseau conceptuel représenté sous la forme d'une ontologie. Cette méthode permet de faire le lien

entre la terminologie du domaine et le modèle conceptuel.

Dans la méthode TERMINAE, l'étape de normalisation est essentiellement manuelle bien que guidée par les interfaces de l'outil TERMINAE. C'est une étape difficile pour l'ingénieur de la connaissance qui se retrouve face à une masse d'unités à traiter, dont certaines sont ambiguës et qui ne sont pas toutes pertinentes pour le domaine. Nous proposons un dispositif de normalisation du réseau terminologique qui permet de construire à partir de ce dernier un réseau *termino-conceptuel* normalisé.

Au niveau terminologique, des unités, essentiellement des termes et des entités nommées, ainsi que des relations sont extraites d'un corpus d'acquisition par des outils de TAL (extracteurs de termes, reconnaisseurs d'entités nommées (REN) et outils d'extraction de relations à partir d'un corpus d'acquisition). Le réseau terminologique, construit à partir de ces éléments, est décrit à travers un graphe $G_T(UT, RT)$ étiqueté et orienté où les noeuds UT décrivant des unités terminologiques (terme, entité nommée) sont interconnectés par des arcs RT décrivant des types de relations terminologiques (syntaxiques, lexicales et sémantiques). Ce réseau terminologique est peu contraint au sens où il n'y a pas de limites sur le nombre et la nature des relations terminologiques.

Au niveau termino-conceptuel, le réseau sémantique est décrit par un graphe étiqueté et orienté $G_{TC}(TC, RTC)$ défini par un ensemble de noeuds termino-concepts TC et un ensemble d'arcs RTC décrivant des types de relations termino-conceptuelles. Les termino-concepts sont des unités terminologiques désambiguïsées qui sont pertinentes pour la modélisation du domaine. Une relation termino-conceptuelle est une relation taxonomique ou associative. Le réseau termino-conceptuel est plus contraint que le réseau terminologique (moins d'unités, pas de relations syntaxiques, différenciation entre relations taxonomiques et associatives).

Nous définissons un dispositif pour guider le travail de normalisation :

1. La désambiguisation des unités terminologiques repose sur l'étude des occurrences de ces dernières dans le texte ou des relations qu'elles entretiennent avec d'autres unités au sein du réseau terminologique. Les structures de connaissances et les interfaces de l'outil sont conçus pour supporter ce travail d'analyse.
2. Pour le filtrage et sélection des unités pertinentes, nous proposons plusieurs critères qui permettent de trier la liste des unités par ordre de pertinence et de mettre l'accent sur les termes susceptibles de dénoter des notions du domaine et mis en valeur dans le discours. Ces

critères dépendent du corpus d'acquisition et du domaine à modéliser. Un terme a un **poids de domaine** élevé s'il figure à proximité d'indices de domaine comme des entités nommées (Omrane *et al.*, 2011). Le **poids de discours** d'un terme augmente s'il figure dans un passage définitoire ou un passage saillant au regard de l'application visée (si toutefois de tels passages peuvent être identifiés). Le poids du terme dépend aussi de son **degré** (*i.e.* du nombre de relations dans lequel il entre).

Le travail de normalisation se fait localement : l'ingénieur de la connaissance sélectionne, désambiguise et regroupe des noeuds et des arcs qu'il a validés. Un mécanisme permet cependant de majorer le poids des unités normalisées et de propager des opérations de normalisation dans le réseau. Cela permet d'orchestrer le travail de normalisation à un niveau global pour assurer la cohérence.

Il n'y pas de critère formel pour détecter la fin du travail de normalisation. Seul l'ingénieur de la connaissance peut le décider en fonction de l'application qu'il vise. Il peut cependant mesurer l'évolution de son travail à travers des indicateurs qui lui sont fournis et qui permettent de mesurer, par exemple, la connectivité du réseau, la complétude de la hiérarchie, le taux des unités validées et/ou normalisées et de leur couverture par rapport au corpus d'acquisition.

Références

AUSSENAC-GILLES N., DESPRÉS S. & SZULMAN S. (2008). The terminae method and platform for ontology engineering from texts. In P. BUITELAAR & P. CIMIANO, Eds., *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, p. 199–223. IOS Press.

OMRANE N., NAZARENKO A. & SZULMAN S. (2011). Le poids des entités nommées dans le filtrage des termes d'un domaine. In *9th International Conference on Terminology and Artificial Intelligence*, p. 80–86.