

Notion d'ontologie et construction d'ontologie à partir de corpus de textes

Synthèse de lectures

de

Benoit Lavoie

benoit@benoit-lavoie.ca

Programme de Doctorat en Informatique Cognitive

Université du Québec à Montréal

8 février 2007

Table des matières

1. DESCRIPTION DE LA PROBLÉMATIQUE	1
2. LE CONCEPT D'ONTOLOGIE EN INFORMATIQUE	1
3. CONSTRUCTION D'ONTOLOGIE À PARTIR DE CORPUS DE TEXTES.....	5
4. REMERCIEMENTS.....	11
5. RÉFÉRENCES.....	11

1. Description de la problématique

Ce document décrit la notion d'ontologie en informatique et les rôles des statistiques et des grammaires rationnelles dans la construction d'ontologie à partir de corpus de textes non structurés.

La section 2 décrit le concept d'ontologie en informatique (le texte de cette section est basé principalement sur le premier chapitre du livre de Gómez-Pérez et al., 2004). La section 3 décrit la construction d'ontologie à partir de corpus de textes à l'aide de statistiques textuelles et de grammaires rationnelles. Finalement la section 4 contient la liste des principales références utilisées.

Ce document est une synthèse de lectures préparée dans le cadre du programme de Doctorat en Informatique Cognitive de l'Université du Québec à Montréal (UQAM).

2. Le concept d'ontologie en informatique

Introduction

Le concept d'ontologie en informatique est lié à celui du concept d'ontologie en philosophie (Gómez-Pérez et al., 2004). D'une part, le concept d'ontologie en informatique repose sur des notions développées informellement en philosophie depuis l'Antiquité: la notion de catégorisation développée par Aristote et ses prédécesseurs, la notion de symbolisme développée par William d'Ockam au Moyen Age, la notion de relativité des observations développée par Kant durant l'Age Moderne, etc. D'autre part, le concept d'ontologie en informatique repose sur la notion d'ontologie formelle développée en philosophie par Edmund Husserl: les ontologies formelles visent au développement d'une logique systématique, formelle et axiomatique afin de classer les entités du monde (objets, évènements, etc.) et les catégories qui les modélisent (concepts, propriétés, etc.).

Aujourd'hui, les travaux sur les ontologies constituent un aspect de recherche important en informatique (Gómez-Pérez et al., 2004; Bouchard & Obaid, 2005). Ces travaux portent sur différents domaines d'informatique incluant l'ingénierie des connaissances et l'intelligence artificielle. Les applications informatiques liées aux ontologies sont multiples: gestion des connaissances, traitement du langage naturel, commerce électronique, etc.

L'ingénierie ontologique est une discipline de l'informatique référant aux activités reliées au processus de développement des ontologies ainsi qu'aux méthodes, outils et langages pour développer ces ontologies. Plusieurs ontologies ont déjà été développées avec différentes méthodes. Un des buts de l'ingénierie ontologique est de rendre compatible la diversité de ces ontologies.

Trois développements récents ont contribué de façon marquée à l'avancement des travaux en ingénierie ontologique (Gómez-Pérez et al., 2004):

- Le programme de recherche “Knowledge Sharing Effort” de DARPA (1991): ce programme proposa de développer des connaissances déclaratives et des techniques de solution de problèmes qui soient réutilisables pour la construction de systèmes à base de connaissances afin de permettre aux développeurs d’applications de se concentrer que sur les connaissances et algorithmes de raisonnement qui soient spécifiques à leur domaine;
- Plusieurs projets (1988 - 1998) qui jetèrent les fondations des notions de méthodologies en ingénierie des connaissances: e.g. un de ces projets en 1993 a été à la base du développement de Protégé ¹ (un des environnements de développement d’ontologie les plus populaires aujourd’hui);
- Le Web Sémantique (1999); selon son fondateur, Berners-Lee, le Web Sémantique est une extension de l’Internet où les informations sont définies sémantiquement afin de faciliter le partage et la coopération.

Définitions du concept d’ontologie pertinentes à l’informatique

Plusieurs définitions du concept d’ontologie pertinentes pour l’ingénierie ontologique ont été proposées. Ces définitions sont souvent des raffinements de définitions déjà proposées et/ou sont complémentaires avec elles. La définition d’ontologie la plus citée dans la littérature est celle proposée par Gruber (1993 — citation Gómez-Pérez et al., 2004):

An ontology is an explicit specification of a conceptualisation.

[Traduction: *Une ontologie est une spécification explicite d’une conceptualisation.*]

Cette courte définition est indépendante du processus de construction de l’ontologie ou de son intégration avec une base de connaissances. Plusieurs auteurs ont proposé des raffinements à cette définition incluant:

- Borst (1997 — citation Gómez-Pérez et al., 2004) qui y ajouta la notion de conceptualisation *partagée*;
- Studer et collègues (1998 — citation Gómez-Pérez et al., 2004) qui la détaillèrent et restreignirent son application aux ontologies décrites avec des *langages formels* (voir plus bas).

Selon Gómez-Pérez et collègues (2004), les ontologies visent à capturer les connaissances consensuelles de façon générique afin de faciliter leur réutilisation et leur partage d’une application à une autre et d’un groupe de chercheurs à un autre. Les ontologies sont généralement construites de façon coopérative par des gens localisés à différents endroits.

Niveaux de formalisation des ontologies

Les ontologies peuvent être modélisées avec des langages de différents niveaux de formalité: (i) *hautement informel*: défini en langage naturel; (ii) *semi-informel*: défini dans une forme structurée et restreinte de langage naturel; (iii) *semi-formel*: défini dans un langage formel artificiel; et (iv) *rigoureusement formel*: défini avec un langage dont les termes sont associés avec une sémantique formelle, des théorèmes, et des preuves pour vérifier les propriétés telles que la validité et la complétude.

¹ <http://protege.stanford.edu>

Ontologies légères et ontologies lourdes

On distingue les ontologies légères (*lightweight ontologies*) et les ontologies lourdes (*heavyweight ontologies*) (Gómez-Pérez et al., 2004). Les ontologies légères incluent des concepts comprenant des propriétés et qui sont organisées en taxonomies avec des relations conceptuelles (e.g. *Yahoo! Directory*); certains auteurs considèrent les taxonomies comme des ontologies parce qu'elles fournissent des conceptualisations partagées pour des domaines donnés. Les ontologies lourdes ajoutent aux ontologies légères des axiomes et des restrictions clarifiant le sens. Les ontologies lourdes modélisent un domaine de façon plus profonde avec plus de restrictions basées sur la sémantique du domaine.

Classifications des ontologies

Plusieurs classifications des ontologies ont été proposées incluant la classification de Lassila et McGuinness (2001 — citation Gómez-Pérez et al., 2004) et celle de Gómez-Pérez et collègues (2004).

Lassila et McGuinness proposent une classification des ontologies selon l'information dont l'ontologie a besoin et la richesse de sa structure interne. Cette classification consiste en un continuum allant d'ontologies légères aux ontologies lourdes:

- *Vocabulaire contrôlé*: liste de termes.
- *Glossaire*: liste de termes avec leur sens spécifié en langage naturel.
- *Thésaurus*: glossaire contenant des descriptions sémantiques entre les termes.
- *Hiérarchie informelle de généralisation (is-a)*: hiérarchie du type *Yahoo! Directory* dont la structure est basée non pas sur des relations de généralisation mais sur la proximité des concepts.
- *Hiérarchie formelle de généralisation (is-a)*: hiérarchie dont la structure est déterminée par des relations de généralisation.
- *Hiérarchie formelle de généralisation (is-a) avec instances du domaine*: similaire à la catégorie précédente mais incluant des instances.
- *"Frame"*: ontologies incluant des classes avec propriétés pouvant être héritées.
- *Ontologies avec restrictions de valeur*: ontologies pouvant contenir des restrictions sur les valeurs des propriétés.
- *Ontologies avec contraintes logiques*: ontologies pouvant contenir des contraintes entre constituants (e.g. relations) définies dans un langage logique.

Gómez-Pérez et collègues (2004) propose une classification basée sur celle de Van Reijst et collègues (1997 — citation Gómez-Pérez et al., 2004) où les ontologies sont classées selon deux critères: (i) la quantité et le type de structure dans la conceptualisation, et (ii) le sujet de la conceptualisation.

- *Ontologies de représentation de connaissances*: modélisent les représentations primitives utilisées pour la formalisation des connaissances sous un paradigme donné.

- *Ontologies générales ou communes* : modélisent les connaissances de sens commun réutilisables d'un domaine à l'autre. Ces ontologies incluent un vocabulaire relatif aux choses, évènements, temps, espace, causalité, comportement, fonction, etc.
- *Ontologies de niveau supérieur (top-level, upper-model)*: modélisent les concepts très généraux auxquels les racines des ontologies de plus bas niveaux devraient être liées. Cependant, il existe plusieurs ontologies de niveau supérieur et qui sont divergentes. Afin de résoudre ce problème, l'organisation de standardisation IEEE tente de développer une ontologie de niveau supérieur qui soit standard.
- *Ontologies de domaine* : modélisent les connaissances réutilisables dans des domaines précis. Ces ontologies fournissent les concepts et les relations permettant de couvrir les vocabulaires, activités et théories de ces domaines. Les concepts des ontologies de domaine sont souvent des spécialisations de concepts définis dans des ontologies de niveau supérieur.
- *Ontologies de tâches* : modélisent les vocabulaires relatifs à une tâche ou une activité générique en spécialisant certains termes des ontologies de niveau supérieur.
- *Ontologies de tâches de domaine* : ontologies de tâches réutilisables dans un domaine spécifique mais pas d'un domaine à l'autre et qui sont indépendantes de l'application.
- *Ontologies de méthodes* : modélisent les définitions des concepts et des relations pertinentes pour le processus de raisonnement afin d'effectuer une tâche spécifique.
- *Ontologies d'applications* : modélisent les connaissances requises pour des applications spécifiques. Les ontologies d'applications spécialisent souvent le vocabulaire des ontologies de domaine et des ontologies de tâches.

Modélisations des ontologies

Les ontologies lourdes peuvent être modélisées avec (i) des méthodes d'intelligence artificielle basées sur les "*frames*" et sur la logique de premier ordre (e.g. Cyc ² et Ontolingua ³ développés dans les années '90) ou avec (ii) la logique de description (Baader et al., 2003 — citation Gómez-Pérez et al., 2004) utilisée pour développer le langage ontologique OWL (Web Ontology Language) ⁴ du Web Sémantique.

Les ontologies légères peuvent être modélisées à partir des modèles utilisés en génie logiciel: la modélisation UML (Unified Modeling Language; Rumbaugh et al., 1998 — citation Gómez-Pérez et al., 2004) et la modélisation par diagramme Entité/Relation (Chen, 1976 — citation Gómez-Pérez et al., 2004).

² <http://www.cyc.com/>

³ <http://ontolingua.stanford.edu>

⁴ <http://www.w3.org/TR/owl-features/>

3. Construction d'ontologie à partir de corpus de textes

Motivations pour la construction d'ontologie à partir de corpus de textes

Différentes ressources sont disponibles pour la construction d'ontologie (Biemann, 2005, p.79; Gómez-Pérez & Manzano-Macho, 2003, p. 11): ontologies existantes, bases de données, documents semi-structurés (e.g. schémas XML), dictionnaires, corpus de textes non structurés, etc.

L'utilisation de corpus de textes non structurés offre différents avantages: (i) les textes non structurés sont généralement facilement accessibles (e.g. voir la plupart des textes trouvés sur Internet); (ii) les textes non structurés contiennent des connaissances générales et/ou des connaissances de domaine qui peuvent être pertinentes pour les ontologies générales et/ou les ontologies de domaine; et (iii) plusieurs méthodes et outils existants permettent la construction semi-automatique d'ontologie à partir de corpus de textes non structurés (Biemann 2005; Buitelaar et al, 2005; Gómez-Pérez & Manzano-Macho, 2003).

Principales tâches reliées à la construction d'ontologie

Buitelaar et collègues (2005, p.6) identifient différentes tâches génériques reliées à la construction d'ontologie dans le contexte de l'extraction semi-automatique de connaissances à partir de corpus de textes non structurés. La Figure 2 illustre l'organisation de ces tâches: (1) identification des termes pertinents pour l'ontologie développée (la terminologie variera selon qu'il s'agit d'une ontologie générale ou d'une ontologie de domaine); (2) identification des relations de synonymie entre les termes; (3) identification des concepts et de leurs attributs; (4) identification des relations taxonomiques (structure de la hiérarchie) entre concepts; (5) identification des relations non taxonomiques (génériques) entre concepts; et (6) identification des règles spécifiant les contraintes sur les propriétés des concepts et les relations entre concepts.

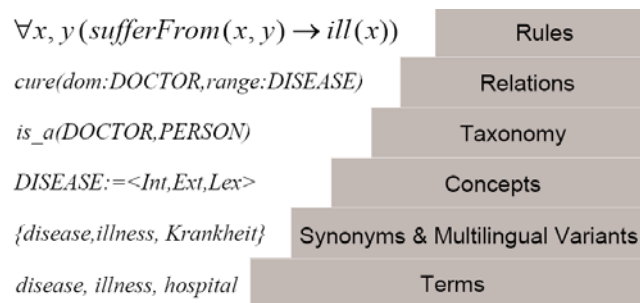


Figure 1. Tâches liées à la construction d'ontologie: source (Buitelaar et collègues, 2005, p.6)

Principales approches de construction d'ontologie à partir de corpus de textes non structurés

Maedche et Stabb (2001 — citation Gómez-Pérez & Manzano-Macho, 2003, p. 11) proposent une classification des principales approches utilisées pour la construction d'ontologie à partir de corpus de textes non structurés: (i) *extraction à base de patron*: une relation est reconnue quand une séquence de mots dans un texte correspond à un patron textuel; (ii) *association par règles*: mise en correspondance de termes ou de concepts par le biais de règles de type implication (*si X alors Y*); (iii) *regroupement conceptuel*: les concepts sont regroupés selon la distance sémantique entre chacun afin de former des hiérarchies (quelques métriques pour déterminer la distance sémantique sont détaillées plus loin); (iv) *élagage*

d'ontologie: des textes généraux et spécifiques à un domaine donné sont utilisés afin d'éliminer des concepts qui ne sont pas spécifiques au domaine; et (v) *apprentissage de concepts*: une taxonomie donnée est mise à jour incrémentalement avec de nouveaux concepts dont les termes correspondants sont extraits à partir de textes.

Rôle des statistiques textuelles pour la construction d'ontologie

La *fréquence de terme* (Wikipedia, 2006e)⁵ est proportionnelle au nombre de fois qu'un terme apparaît dans un document. Ce nombre est généralement normalisé par le nombre de tous les termes dans un document afin d'éviter les biais pour de longs documents. La fréquence de terme est parfois utilisée pour (i) l'identification de termes pertinents d'un corpus ou (ii) directement pour l'identification de candidats de concepts (ou d'instances de concepts) et de relations (Kietz et collègues, 2000 – citation Gómez-Pérez & Manzano-Macho, 2003); les termes et couplages de termes dont les fréquences relatives sont plus grandes dans un corpus de domaine que dans un corpus général sont parfois proposés à l'expert respectivement comme candidats de concepts et comme candidats de relations.

TF-IDF (*Term Frequency - Inverse Document Frequency*) (Wikipedia, 2006e) est une mesure statistique utilisée pour évaluer l'importance (le poids) d'un mot dans un document d'un corpus. L'importance croît proportionnellement avec le nombre d'occurrences du mot dans le document mais est contrebalancée par la fréquence du mot dans le corpus. La fréquence inverse du document (IDF) pour un terme donné est une mesure de l'importance générale de ce terme (c'est le logarithme du nombre de tous les documents divisé par le nombre de documents contenant ce terme): pour un terme t_i , $TF-IDF = TF \times IDF$ où $IDF = \log (\text{nombre de documents} / \text{nombre de documents contenant le terme } t_i)$. Un haut poids TF-IDF est atteint avec une haute fréquence d'un terme dans un document donné et une faible occurrence de ce terme dans les documents du corpus. TF-IDF tend à filtrer les termes communs. TF-IDF est parfois utilisé comme métrique de similarité pour mesurer la distance entre termes (citation Gómez-Pérez & Manzano-Macho, 2003, p. 15): la métrique permet ainsi de regrouper les termes similaires en concepts communs. TF-IDF est également utilisé pour détecter les termes pertinents à un domaine donné (citation Gómez-Pérez & Manzano-Macho, 2003, p. 45).

Chi-square (χ^2) (Manning & Schütze, 1999) est un test statistique déterminant la similarité dans les données en comparant les fréquences observées et les fréquences espérées en supposant l'indépendance d'occurrence dans les données. Si la différence est grande entre les fréquences observées et celles espérées alors on peut supposer qu'il y a dépendance entre les données. Ce test additionne les différences au carré entre les fréquences observées (O) et celles espérées (E) pondérées par la grandeur des fréquences espérées: $\chi^2 = \sum_{ij} (O_{ij} - E_{ij})^2 / E_{ij}$ où ij sont les indices des données à comparer (typiquement, ce sont les indices d'une cellule dans une table où rangées et colonnes sont associées à des termes). Dans le contexte de la construction d'ontologie à partir de texte, χ^2 est utilisé de façon similaire à TF-IDF: χ^2 est utilisé comme métrique de similarité pour mesurer la distance entre termes (citation Gómez-Pérez & Manzano-Macho, 2003, p. 15) et il est également utilisé pour détecter les termes de domaine les plus pertinents (Buitelaar et al., 2004).

⁵ La référence (Wikipedia, 2006e) est utilisée pour décrire la fréquence de terme ainsi que TF-IDF.

L'**analyse sémantique latente** (Manning & Schütze, 1999; Wikipedia, 2006d) est une technique d'extraction et de représentation de la signification contextuelle des mots par calculs statistiques sur un large corpus textuel. L'idée de base est que l'agrégation des contextes où les mots apparaissent et n'apparaissent pas fournit un ensemble de contraintes déterminant la signification des mots et la similarité avec les autres mots. L'analyse sémantique latente utilise une matrice de termes de documents décrivant les occurrences des termes dans les documents. La matrice est creuse avec les rangées correspondant à des documents et les colonnes correspondant aux termes. La métrique TF-IDF est souvent utilisée afin de pondérer les occurrences des termes (le poids d'une entrée donnée dans la matrice est proportionnel au nombre de fois qu'un terme apparaît dans un document). Les poids des termes rares sont ajustés de façon à refléter leurs importances relatives. L'analyse sémantique latente transforme la matrice de termes de documents en un espace de concepts latents (agrégation de contextes d'apparition des termes) permettant de mettre en relations termes et concepts, ainsi que documents et concepts. La notion de concept issue de l'analyse sémantique latente est supportée par l'hypothèse distributionnelle de Richard Harris (citation Biemann, 2005; Wikipedia, 2006c) selon laquelle les mots qui tendent à apparaître dans des contextes similaires ont des sens similaires. Dans les deux cas, la co-occurrence de termes est interprétée comme un indicateur de proximité sémantique. De façon générale, l'extraction de connaissances ontologiques à partir de textes non structurés repose sur l'hypothèse distributionnelle des mots dans les textes

Dans le contexte de construction d'ontologie à partir de textes, l'analyse sémantique latente permet ainsi de déterminer la similarité entre termes et de les mettre en relations: relations de synonymie (où différents termes réfèrent à un même concept) et relations de polysémie (où un même terme réfère à plusieurs concepts): ces informations peuvent servir à regrouper les termes dans une ontologie (Paaß & al., 2004). L'analyse sémantique latente permet également de regrouper et de classifier des documents selon leurs similarités conceptuelles (Biemann, 2005, p.86). Les documents regroupés portent sur des domaines similaires et peuvent servir de base à la collecte de termes pour la construction d'ontologie de domaine.

Rôle des grammaires rationnelles pour la construction d'ontologie

Une grammaire rationnelle est un ensemble d'expressions (couples, triplets, schémas, patrons, etc.) ordonnées définissant un langage (Serrano, 1991). Dans le contexte de l'analyse de texte, la grammaire spécifie la composition des phrases d'un langage naturel.

Les modèles **N-grams** (Manning & Schütze, 1999, pp. 192-195; Jurafsky & Martin, 2000, Chapter 6) sont considérés par certains comme des *grammaires rationnelles stochastiques* (Casacuberta & Vidal, 2004, p. 205); i.e. des modèles hybrides représentant les propriétés statistiques et compositionnelles des textes. Les modèles N-grams modélisent des séquences de termes (mots, lettres, etc.) avec des N-grams qui sont des sous-séquences de N termes. Ils permettent de déterminer la probabilité d'un mot étant donné les N-1 mots précédents. Les séquences à fortes probabilités permettent de déterminer les séquences de mots fortement associés, telles les collocations (e.g. "*cordons bleus*"), où chaque séquence de mots pourra être mise en correspondance avec un concept. Les N-grams permettent également de comparer les contextes d'occurrences

des mots et d'évaluer les similarités dans le cadre de construction d'ontologie (Sugiura & al, 2004, pp. 3-6).

Les **patrons d'expressions** (e.g. expressions régulières; Jurafsky & Martin, 2000, Chapter 2) contiennent des termes et des variables auxquelles peuvent être associées des contraintes. Les patrons d'expressions sont unifiés avec des textes de façon à instancier les patrons avec des fragments de textes satisfaisant aux structures et aux contraintes des patrons. Les patrons d'expressions permettent de spécifier des relations et/ou des arguments de ces relations afin d'extraire des mots correspondant aux relations ou aux arguments. Par exemple, ils sont parfois utilisés pour l'extraction des relations d'hyponymie (relation sémantique de subordination ou d'appartenance à une classe de plus bas niveau) (Biemann, 2005; Hearst, 1998 – citation Gómez-Pérez & Manzano-Macho, 2003): sachant que *Shakespeare* est un hyponyme de *poète*, à partir du patron correspondant à la séquence "*poète ... Shakespeare*" on peut trouver dans un texte l'expression "*poète tel que Shakespeare*" et faire l'hypothèse que "*X tel que Y*" indique une relation d'hyponymie entre *X* et *Y*. Inversement, à partir du patron "*X tel que Y*" on peut trouver dans un corpus les couples de mots *X* et *Y* qui sont possiblement en relations d'hyponymie. La même approche peut être appliquée avec d'autres types de relations lexicales: relations de hypernymie (inverse de l'hyponymie), synonymie, antonymie, méronymie (relation entre la partie et le tout), etc. Les patrons de mots sont souvent utilisés pour raffiner des ontologies existantes. Cependant, les taux d'erreurs sont parfois élevés et des vérifications par des experts sont souvent nécessaires (Gómez-Pérez & Manzano-Macho, 2003).

Les **grammaires morphologiques** (Jurafsky & Martin, 2000, Chapter 3) modélisent les constituants morphologiques des mots (morphèmes lexicaux et morphèmes grammaticaux). Elles permettent de déterminer la similarité des termes au niveau des morphèmes lexicaux et de faire abstraction des différences grammaticales (e.g. "*cheval*" et "*chevaux*" sont des termes dont les morphèmes lexicaux sont similaires). Les grammaires morphologiques sont souvent implémentées avec des automates à états finis (Jurafsky & Martin, 2000, Chapter 3). Dans le contexte de la construction d'ontologie à partir de textes, les grammaires morphologiques sont utilisées pour le prétraitement des textes afin d'obtenir des morphèmes lexicaux à partir desquels d'autres traitements sont effectués (Maedche & Staab, 2004, p. 177). Par exemple, les matrices de termes utilisées pour l'analyse sémantique latente peuvent contenir des morphèmes lexicaux obtenus après prétraitement des textes.

Les **grammaires morpho-syntaxiques** (Jurafsky & Martin, 2000, Chapter 8) modélisent les catégories syntaxiques des mots (nom, verbe, adjectif, etc.). Elles peuvent permettre de désambigüiser des termes similaires à partir des catégories syntaxiques (e.g. le terme "*couvent*" peut être désambigüisé syntaxiquement selon qu'il est un nom ou un verbe). Suivant les travaux de Chen (1983), les grammaires morpho-syntaxiques sont souvent utilisées dans les applications de modélisation conceptuelle (e.g. Overmyer et al., 2001) afin de déterminer les types d'information (concepts, attributs ou relations) à partir des catégories syntaxiques: Chen propose que des termes nominaux soient associés aux noms de concepts, que des termes adjectivaux soient associés aux attributs de concepts, et que des termes verbaux soient associés aux relations entre concepts. Des approches similaires sont utilisées pour la construction d'ontologie. Les étiqueteurs morpho-syntaxiques peuvent servir à identifier les termes pertinents pour la construction d'ontologie à partir des simples

catégories syntaxiques. Cependant, la plupart des approches de création d'ontologie n'utilisent que les noms communs afin d'identifier des candidats de concepts (Biemann, 2005, p. 79).

Les **grammaires syntaxiques** (Jurafsky & Martin, 2000, Chapters 9–12) modélisent les structures syntaxiques des phrases. Ces grammaires permettent de distinguer les relations de dépendances syntaxiques entre les mots qui sont généralement spécifiés par leurs propriétés morpho-syntaxiques: e.g. l'expression "*les poules couvent*" peut être analysée comme étant composée d'un syntagme verbal (le verbe *couver* à l'indicatif présent) ayant pour sujet syntaxique un syntagme nominal (*les poules*, où *poule* est un nom commun pluriel défini). Dans le contexte de construction d'ontologie à partir de textes, les grammaires syntaxiques permettent de regrouper les termes selon les similarités syntaxiques. Par exemple, dans l'outil SVETLAN (citation Gómez-Pérez & Manzano-Macho, 2003, p. 45), les termes nominaux qui dans le corpus ont des relations syntaxiques similaires avec les mêmes termes verbaux sont agrégés sous le même concept: l'hypothèse est que les verbes et leurs relations syntaxiques permettent de catégoriser les noms. Cependant les analyseurs syntaxiques sont relativement peu employés pour la création d'ontologie (Buitelaar, 2005, p. 4); les grammaires syntaxiques sont souvent peu accessibles pour la plupart des langues et leurs couvertures sont souvent insuffisantes pour de grand corpus de textes (Manning & Schütze, 1999).

Les **grammaires sémantiques** (Jurafsky & Martin, 2000, Chapters 15) modélisent les informations sémantiques associées aux phrases. Au niveau sémantique, les termes sont généralement classifiés comme des objets, des événements ou des états. Les informations sémantiques portent sur les propriétés de ceux-ci (e.g. objet animé) et sur leurs relations (e.g. relation causale entre un objet et un événement). Les grammaires sémantiques sont parfois intégrées aux grammaires syntaxiques puisque ces premières dépendent souvent de ces dernières. Dans le contexte de construction d'ontologie à partir de textes, les grammaires sémantiques permettent de regrouper les termes selon leurs similarités sémantiques. Par exemple, OntoExtract (citation Gómez-Pérez & Manzano-Macho, 2003, pp. 37-38) génère des taxonomies (ontologies légères) à partir d'analyse basée sur des grammaires sémantiques. Cependant, comme pour le cas des analyseurs syntaxiques, les analyseurs basés sur des grammaires sémantiques sont encore relativement peu employés pour la création d'ontologie à cause des ressources insuffisantes pour la plupart des langues ou pour couvrir de grands corpus de textes.

Exemple d'outil de construction d'ontologie à partir de corpus de textes

La Figure 3 illustre l'architecture d'un système permettant l'extraction de connaissances ontologiques à partir de corpus de textes (Buitelaar et al., 2004). Le système est composé de trois composantes principales: (i) SCHUG, un outil d'annotation XML⁶ de corpus à partir d'analyse linguistique; (ii) Protégé (2006), un environnement de développement interactif d'ontologie; et (iii) OntoLT, un outil instanciant une ontologie à partir du corpus annoté.

⁶ Extensible Markup Language; <http://www.w3.org/XML/>

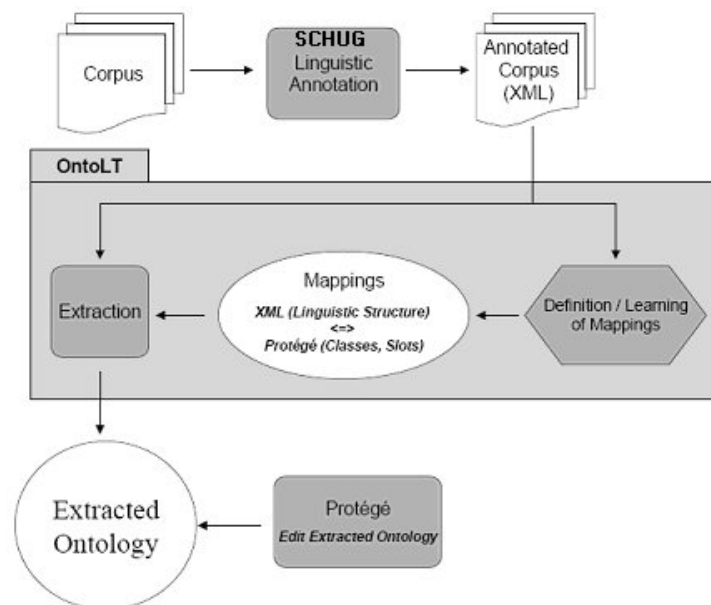


Figure 2. Système de construction d'ontologie à partir de corpus de textes: source (Buitelaar et al, 2004)

SCHUG utilise des grammaires syntaxiques pour annoter le corpus. À partir des annotations XML résultantes, OntoLT génère automatiquement un ensemble de règles d'extraction de format XPath (langage utilisé pour référer aux annotations XML et qui permet la spécification de préconditions d'application à partir des annotations)⁷. XPath est un exemple de langage à base de patrons d'expressions (voir plus haut). Pour la génération des règles d'extractions à partir des annotations, OntoLT utilise des heuristiques syntaxiques basées sur les catégories et les relations (e.g. un sujet de prédicat est associé à un concept⁸). OntoLT utilise également la métrique statistique Chi-square (χ^2 ; voir plus haut) afin de déterminer la pertinence des termes pour un domaine donné. Les règles d'extraction produites peuvent ensuite être modifiées manuellement par un développeur afin de raffiner l'extraction. L'application des règles permet d'instancier une ontologie superficielle qui peut ensuite être raffinée interactivement par le développeur avec Protégé.

Nécessité de l'intervention humaine pour la construction d'ontologie

Aucun outil ou méthode ne permet aujourd'hui de créer de façon totalement non supervisée des ressources sémantiques de bonne qualité (Biemann, 2005, p.75); l'auteur sous-entend ici des ressources suffisamment profondes, dont la couverture soit suffisamment large et dont les erreurs dans les analyses soient suffisamment négligeables. La plupart des outils disponibles pour la construction d'ontologie à partir de textes (voir la compilation de Gómez-Pérez & Manzano-Macho, 2003) sont décrits comme nécessitant l'intervention humaine à différents niveaux:

- **Procurer des données ou connaissances initiales;** plusieurs méthodes et outils de construction d'ontologie requièrent que des experts fournissent des données ou connaissances pour initialiser la construction semi-automatique. Par exemple, certaines méthodes de construction d'ontologie (Aussenac-Gilles et collègues, 2000 –

⁷ <http://www.w3.org/TR/xpath>

⁸ Dans Protégé, les concepts sont appelés classes.

citation Gómez-Pérez & Manzano-Macho, 2003, p. 16) recommandent que des experts de domaine choisissent les documents utilisés pour la construction d'ontologie de domaine afin de s'assurer que toutes les notions pertinentes au domaine soient couvertes par ces documents.

- **Raffiner les informations extraites**; beaucoup de méthodes et outils ne permettent d'extraire que des ontologies superficielles qui peuvent souvent nécessiter des raffinements. Par exemple, la plupart des méthodes d'extraction pour la construction d'ontologie ne portent que sur l'extraction de concepts à partir de noms communs (Biemann 2005, p. 79). Dans ce cas, des raffinements peuvent être nécessaires afin d'ajouter des relations non taxonomiques aux ontologies par exemple. Un autre exemple présenté plus haut est OntoDL (Buitelaar et al., 2004) dont les règles d'extraction et les résultats d'extraction peuvent être raffinés par le développeur.
- **Valider les informations extraites**; les taux d'erreurs pour les informations extraites à partir de corpus sont parfois élevés et/ou peuvent nécessiter des validations par des experts. Par exemple le système On-to-Knowledge (Gómez-Pérez & Manzano-Macho, 2003, p. 22) extrait des relations d'hyponymie mais avec un fort taux d'erreurs.

Certaines méthodes et outils sont complètement interactifs et requièrent l'intervention d'experts tout au long du processus de création (e.g. *SubWordNet engineering Process Tool* – citation Gómez-Pérez & Manzano-Macho, 2003, pp. 43-44).

4. Remerciements

Ce travail a été financé par une bourse doctorale du Conseil de Recherches en Sciences Humaines du Canada (CRSH).

5. Références

- Biemann, Chris (2005). Ontology Learning from Text: A Survey of Methods. *LDV-Forum*, Vol. 20, No. 2, pp. 75–93. http://ariadne.coli.uni-bielefeld.de/gldv/site/2005_Heft2/Chris_Biemann.pdf
- Bouchard, Lorne; & Abdel Obaid (2005). *Construction d'ontologie*. Notes de cours (Séminaire spécial en informatique cognitive sur le Web Sémantique), Université du Québec à Montréal.
- Buitelaar, Paul; Philipp Cimiano; & Bernardo Magnini (2005). Ontology Learning from Text: An Overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications Series, Vol. 123, IOS Press, pp. 3–12. <http://www.aifb.uni-karlsruhe.de/WBS/pci/OL-Book-Intro.pdf>
- Buitelaar, Paul; Daniel Olejnik; & Michael Sintek (2004). A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In *Proceedings of the 1st European Semantic Web Symposium (ESWS)*, Heraklion, Greece, pp. 31–44. <http://dfki.de/~paulb/esws04.pdf>
- Casacuberta, Francisco; & Enrique Vidal (2004). Machine Translation with Inferred Stochastic Finite-State Transducers, *Computational linguistics*, Vol. 30, No. 2, pp. 205-225. <http://www.mitpressjournals.org/doi/abs/10.1162/089120104323093294>
- Chen, Peter (1983). *English Sentence Structure and Entity-relationship Diagrams*, Information Sciences, 29, pp. 127-149.
- Gómez-Pérez, Asunción; Mariano Fernández-Lopez; & Oscar Corcho (2004, 2^{ème} édition). Theoretical Foundations of Ontologies, Chapter 1 of *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer-Verlag, pp. 1–45.
- Gómez-Pérez, Asunción; & David Manzano-Macho (2003). *Deliverable 1.5: A survey of ontology learning methods and techniques*. OntoWeb Consortium. <http://www.deri.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf>

- Jurafsky, Daniel; & James Martin (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall.
- Maedche, Alexander; & Steffen Staab (2004). Ontology Learning. In *Handbook on Ontologies in Information Systems*. S. Staab & R. Studer (eds.), Springer, pp. 173-189.
<http://www.aifb.uni-karlsruhe.de/WBS/sst/Research/Publications/handbook-ontology-learning.pdf>
- Manning, Christopher; & Henrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- Overmyer, Scott; Benoit Lavoie; & Owen Rambow (2001). Conceptual Modeling through Linguistic Analysis Using LIDA. *Proceedings of 23rd International Conference on Software Engineering (ICSE 2001)*, Toronto, Canada, pp. 401-410.
<http://www.benoit-lavoie.ca/public/docs/lida-paper.pdf>
- Paaß, Gerhard; Jörg Kindermann; & Edda Leopold (2004). Learning prototype ontologies by hierarchical latent semantic analysis. *Proceedings of Knowledge Discovery and Ontologies*, pp. 193–205. <http://olp.dfki.de/pkdd04/paass-final.pdf>
- Protégé (2006). *Protégé' website*: <http://protege.stanford.edu>
- Serrano, Manuel (1991). Rgc: un generateur d'analyseurs lexicaux efficaces en Scheme. *Cahiers GUTenberg*, no. 9.
<http://www.sop.inria.fr/mimosal/Manuel.Serrano/publi/serrano-jfla92.ps.gz>
- SIG5 (2006). *SIG5' website* (Interest Group on Language Technologies and the Semantic Web): <http://ontoweb-lt.dfki.de>
- Sugiura, Naoki; Yoshihiro Shigeta; Naoki Fukuta; Noriaki Izumi; & Takahira Yamaguchi (2004). Towards On-the-fly Ontology Construction: Focusing on Ontology Quality Improvement. *Proceedings of the 1st European Semantic Web Symposium*, pp. 1–15.
<http://www.yamaguchi.comp.ae.keio.ac.jp/mmm/doddle/publication/esws2004.pdf>
- Wikipedia (2006c). *Distributional hypothesis*. Online article from Wikipedia encyclopedia.
http://en.wikipedia.org/wiki/Distributional_hypothesis
- Wikipedia (2006d). *Latent semantic indexing*. Online article from Wikipedia encyclopedia.
http://en.wikipedia.org/wiki/Latent_semantic_indexing
- Wikipedia (2006e). *tf-idf*. Online article from Wikipedia encyclopedia. <http://en.wikipedia.org/wiki/TFIDF>