Objectif :

Le but de ce document est de définir les notions de référentiel sémantique (ou référentiel linguistique), d’ontologie linguistique (ou de domaine) et ainsi de définir les relations sémantiques qui lient les mots dans un graphique d’analyse de « Text Mining » et leurs rôles dans l’analyse de textes et de données dans un environnement « Big Data ».

Nous étudierons également WordNet, un exemple de référentiel linguistique.

Définitions :

Référentiel sémantique (ou référentiel linguistique) :

*🡺 Trouver dans un article qui traite du « référentiel sémantique dans les SI en santé » mais la définition n’est pas mal est doit pouvoir être étendue.*

Représentation normalisée de l’information médicale par des concepts dont la signification est fixe et inaltérable.

But : palier à certains inconvénients du langage naturel : ambigüité (manque de consensus temporo-spatiale, polysémie, imprécision) et synonymie.

*🡺 Une autre définition (dans un document plus général) peut aussi servir à compléter ce qui est écrit ci-dessus :*

Des outils d’indexation tels que les thésaurus (référentiel linguistique) qui permettent le choix et l’attribution de descripteurs ou mots-clés décrivant de la façon la plus exhaustive possible le contenu conceptuel d’un document.

Pour que l’indexation, et donc la recherche d’informations, soit la plus performante possible, il est nécessaire de ne retenir que l’information utile. Pour y parvenir, plusieurs prétraitements sont nécessaires. Il s’agit notamment d’analyser les contenus pour éliminer les mots sans valeur sémantique et faire émerger l’intégralité des concepts pertinents, c’est-à-dire l’ensemble des groupes nominaux porteurs de sens.

Ontologie linguistique (ou de domaine) :

*🡺 Récupérer à cet endroit :* [*http://www.technolangue.net/imprimer.php3?id\_article=280*](http://www.technolangue.net/imprimer.php3?id_article=280)*, cette ensemble de définitions sous la forme de questions/réponses à Bruno Bachimont, directeur scientifique à l’INA peut s’avérer utile pour nous.*

Comment définiriez-vous une ontologie ?

On peut caractériser une ontologie comme une structuration des concepts d’un domaine. Ces concepts sont rassemblés pour fournir les briques élémentaires et exprimer les connaissances dont on dispose dans ce domaine.

Pouvez-vous faire la distinction entre une ontologie, un thésaurus et un index ?

Un thésaurus est l’ensemble des mots empruntés à la langue d’un domaine. On les structure par des liens d’hyponymie, d’hyperonymie et par des liens d’association, qui sont en général assez vagues puisque ce sont des liens du type « voir aussi » ou « voir également ».

Une ontologie va reprendre la structuration hyperonymique et hyponymique du thésaurus mais va considérablement la contraindre. La signification de la relation d’hyponymie, le long de l’arborescence de l’ontologie, devra être toujours la même, sans changement du point de vue de l’interprétation de cette relation. Par ailleurs, on définira beaucoup plus précisément les relations d’association qu’il y a entre les différents concepts de l’ontologie. Dernière distinction, c’est qu’une ontologie est une relation de concepts (ou notions) alors que dans un thésaurus ce sont davantage des mots plutôt que des concepts. Cela signifie par exemple qu’on peut avoir des concepts qui ne possèdent pas d’appellation habituelle dans la langue usuelle, alors on est obligé d’avoir des périphrases ou des locutions un peu élaborées pour pouvoir désigner un concept dans le cadre d’une ontologie.

On peut prendre un index au sens bibliographique, au sens de l’énumération de terme ou, plus généralement, au sens d’une liste alphabétique, avec simplement le lieu des occurrences dans un contenu donné. On peut considérer que l’ontologie peut servir de répertoire ou de ressources pour constituer un index. Cela donne les concepts que l’on va reprendre pour constituer l’index et voir quels sont les contextes d’apparition de ces concepts dans un contenu donné.

*🡺 Définition de Wikipédia :*

Une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Plus simplement, on peut aussi dire que l' « ontologie est aux données ce que la grammaire est au langage ».

Relations sémantiques :

*🡺 Une première définition trouvée par Google mais non créditée :*

Relation unissant les termes entre eux à l'intérieur d'un thésaurus. On distingue en général les relations d'équivalence, de hiérarchie et d'association.

*🡺 Une autre, trouvée ici :* [*http://www.adbs.fr/relation-semantique-18448.htm?RH=OUTILS\_VOC*](http://www.adbs.fr/relation-semantique-18448.htm?RH=OUTILS_VOC)

Relation mettant en évidence les liens de signification entre les termes ou les classes d'un langage documentaire. Les principales relations sémantiques sont la relation d'équivalence, la relation hiérarchique et la relation associative.

Rôles dans l’analyse de « Text Mining » et dans les environnements « Big Data » :

Faire un paragraphe où l’on décrit (pour rappel) les graphiques que l’on trouve à la fin des analyses de textes (ex : avec Mahout, khCoder…).

Quelles améliorations apportent les référentiels à ces graphiques ?

Quels gains dans la compréhension du texte d’origine ?

Donner des exemples de référentiels (comment ils se présentent, comment on les utilise…)…

WordNet :

Pour rappel : WordNet est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'année. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Des versions de WordNet pour d'autres langues existent, mais la version anglaise est cependant la plus complète à ce jour.

La base de données ainsi que des outils sont disponibles pour téléchargement gratuit. Par rapport aux outils fournis, un développeur peut aussi accéder la base de données à partir des interfaces disponibles pour de nombreux langages de programmation.

WordNet est distribué sous une licence libre, permettant de l'utiliser commercialement ou à des fins de recherche.

Site web : <http://wordnet.princeton.edu/>

Download : <http://wordnet.princeton.edu/wordnet/download/>

Fichier : WordNet-2.1.exe (pour sa version Windows – 18 Mo). Version 3.0 en cours de préparation.

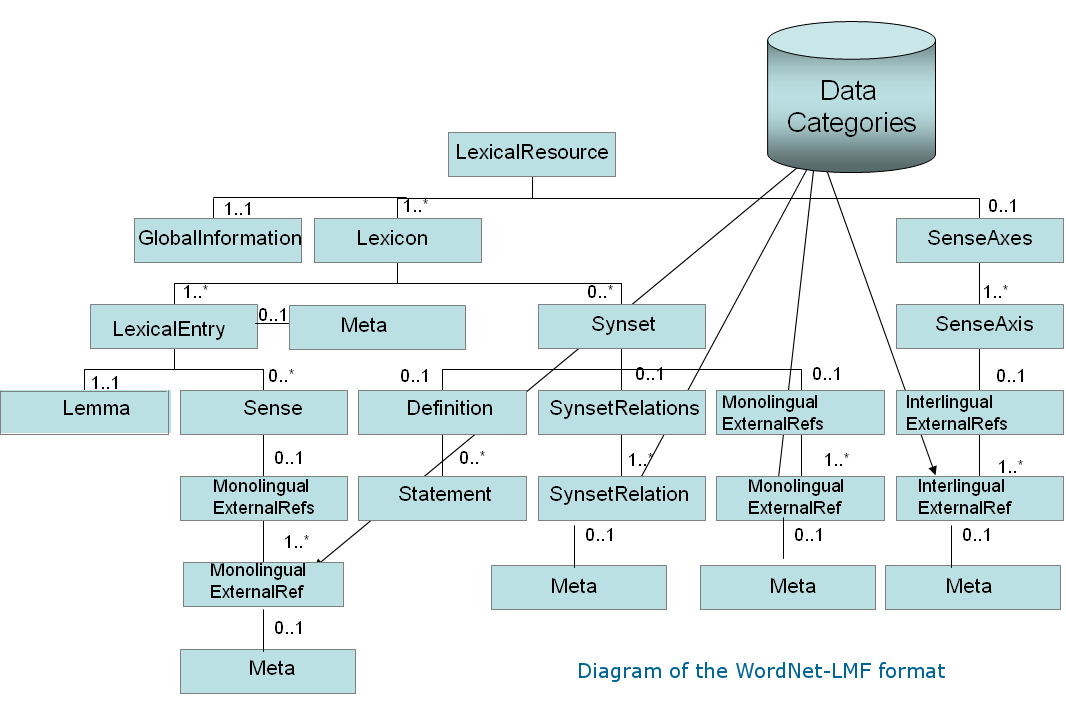
Version française : <http://alpage.inria.fr/~sagot/wolf.html>

Download : <https://gforge.inria.fr/frs/download.php/file/33496/wolf-1.0b4.xml.bz2>

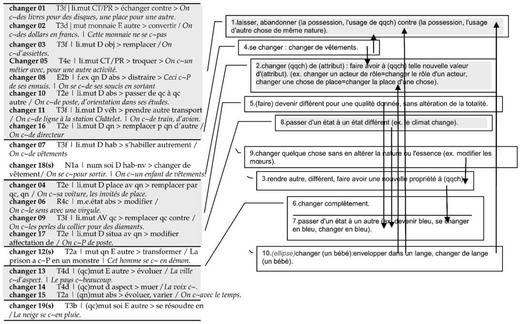
Fichier : wolf-1.0b4.xml (fichier unique – 39 Mo).

Exemples d’utilisation de WordNet (enrichissement des graphismes d’analyse de texte) :

Schéma conceptuel :



Exemple de liaison :



Exemple d’enrichissement du graphisme :

