Annotation sémantique : profilage textuel et lexical.

LEXICOGRAPHIE ET INFORMATIQUE : BILAN ET PERSPECTIVES

Colloque international à l'occasion du 50e anniversaire du lancement du projet du *Trésor de la Langue Française*

Dates : 23-25 janvier 2008

Lieu : Nancy, Campus Lettres et Sciences Humaines

Organisation : UMR ATILF (CNRS/Nancy-Université)

Mots-clés : sémantique textuelle, recherche d'informations, annotation sémantique, statistiques textuelles, TAL

Keywords: textual semantics, information retrievial, semantic annotation, textometry, NLP

Résumé : Cet article présente la plateforme d'annotation sémantique développée au sein du projet DIXEM. Celle-ci s'appuie sur un premier lexique extrait automatiquement depuis le TLFi. Différents aspects de cet outil sont abordés, son cadre théorique, les modes de représentation choisis, ses capacités actuelles ou encore les perspectives et les objectifs que nous suivons. Enfin, nous décrirons ses utilisations afin d'observer, sous un angle nouveau, deux corpus préparés à l'origine pour une étude linguistique, respectivement de la féminisation dans le vocabulaire français et du discours journalistique à propos de l'immigration en France.

Abstract : This article presents a semantic annotation system which has been built in the framework of the DIXEM project. It starts with a first automatically-extracted lexicon from the TLFi dictionary. Various aspects of this system are described: its theoritic framework, its interface, its present possibilities and the goals we would like to reach. Two experimentations are described: one about the feminisation in the french vocabulary and the second about immigration in France.

Introduction

La plateforme a été réalisée dans le but de pouvoir analyser des corpus de textes, les annoter sémantiquement et d'essayer d'en extraire des données sémantiquement informatives sur la base de procédures statistiques.

Nous optons pour une approche complémentaire en croisant sémantique interprétative et statistiques, et souhaitons offrir à la linguistique ainsi qu'au TAL un nouvel objet d'étude. Nous suivons en cela plusieurs travaux actuels [Rossignol & Sébillot 2006], [Enjalbert & Victorri 2005 : p. 82-83] et [Caillet, Pessiot, Amini & Gallinari 2004]. Au final, nous voulons isoler des informations sémantiques (et/ou thématiques) en étudiant les isotopies textuelles d'un texte par des méthodes statistiques.

La plateforme d'annotation a été réalisée en Python, langage de prototypage reconnu, qui permet d'écrire des programmes informatiques rapidement en respectant néanmoins la plupart des contraintes et normes des langages robustes tel Java. Elle se veut être

également un composant Python facilement réutilisable (et modifiable ou étendable) en se présentant sous la forme d'un paquetage indépendant et transportable.

L'outil s'utilise en ligne de commande (via des scripts python) ainsi qu'en manipulant un fichier de configuration. Dans un premier temps, aucune interface graphique ne vient accompagner le logiciel. Cela n'exclut pas la possibilité d'en développer par la suite . Il est fréquent que des applications graphiques soient simplement des « front-end » d'applications en ligne de commande, c'est-à-dire des interfaces graphiques pour des programmes non graphiques.

Enfin la première version de la plateforme souhaite, malgré sa jeunesse, constituer un socle logiciel cohérent, pérenne, efficace, avec le moins d'erreurs et le plus de contrôles possibles. Le projet cherche actuellement à établir des bases solides pour les enrichissements futurs. C'est dans ce but également qu'un manuel d'utilisation a été réalisé.

1 Présentation de la plateforme d'annotation « Sémy »

La plateforme s'appuie à la base sur les entrées du TLFi[1] dont voici un exemple :

> VERTEX, subst. masc.
>
> A. 1. ANAT., ANTHROPOL. Point le plus élevé de la voûte crânienne.
>
> 2. ENTOMOL. ,,Région de l'épicrâne située immédiatement derrière le front entre les yeux composés`` (SÉGUY 1967).
>
> B. 1. ASTRON. ,,Point représentatif, sur la sphère céleste, de la direction du vecteur vitesse d'un ensemble d'étoiles`` (Astron. 1980).
>
> 2. GÉOD. Point de latitude maximale d'une courbe qui est tracée sur une surface de révolution.

Tous les mots sémantiquement pleins (verbes, noms, adjectifs, adverbes) de chacune des définitions sont considérés par hypothèse comme ses traits sémantiques. Un lexique de traits sémantiques, dont la structure est détaillé dans la partie 1.1, est donc crée de la sorte et nous permet d'annoter sémantiquement des textes.

1.1 Représentation du lexique

Comme l'illustre la figure 1, le lexique est structuré par association d'éléments. Ainsi, il contient une entrée 'vertex', qui contient une seule entrée 'commonNoun' (puisque l'entrée 'vertex' n'a qu'une seule forme lemmatique) à laquelle sont associées toutes les informations qui sont nécessaires : le nombre de définitions de l'entrée, les traits sémantiques extraits, les informations de domaines ou de crochets. Toutes ces données sont représentées de façon compacte mais sans perdre d'information.

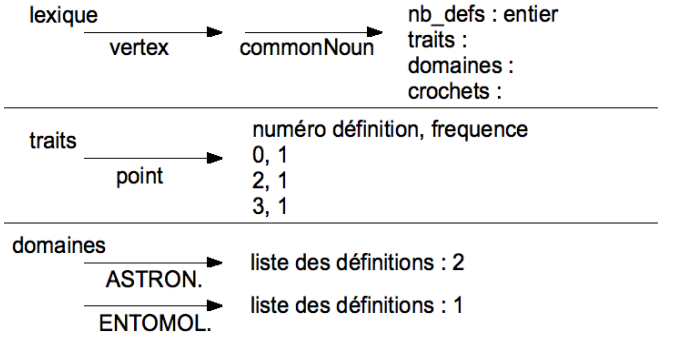

Figure 1 : La représentation du lexique

[1] En 2006, Etienne Pet à partir des entrées du TLFi, ce qui a permis mes utilisé, SEMEME. L'extraction consiste en la sélection des mots pleins des définitions associées aux mots vedettes du dictionnaire.

1.2 Représentation distributionnelle des textes

Pour chaque texte, deux index sont créés, l'un recensant tous les éléments textuels de surface[2], et l'autre étant l'index des traits sémantiques à savoir le résultat de l'annotation sémantique des items restants après le filtrage morphologique.

Au sein des index, chaque élément textuel ou trait sémantique est associé à sa représentation distributionnelle dans le texte et ses paragraphes[3]. Cela simplifie grandement, d'une part, l'analyse statistique qu'il est possible de réaliser sur le texte et ses items, et d'autre part, le développement de la structure logicielle. En effet, avec ces deux index (les seuls à être enregistrés), il est ensuite possible d'en générer d'autres comme par exemple l'index des lemmes d'un texte, mais aussi l'index d'un corpus, ou encore l'index des lemmes d'un corpus.

Cette représentation des corpus et des textes est synthétisée dans la figure 2.

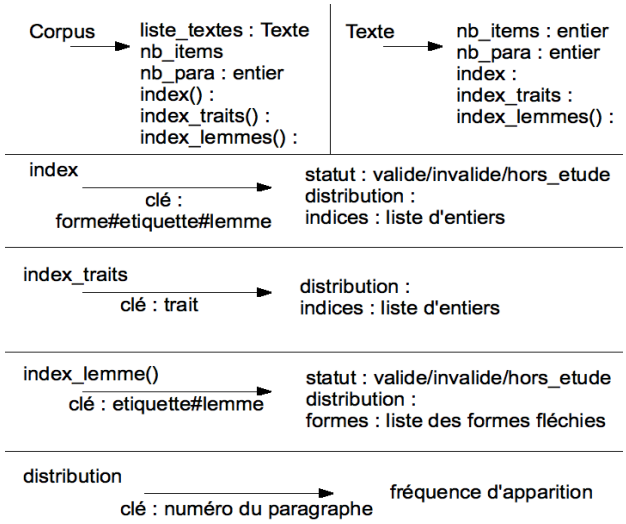

Figure 2 : La représentation des textes, des corpus et des indexs.

1.3 Fonctionnement

La première version permet de finaliser plusieurs opérations essentielles à l'outil en développement. Ainsi, à ce jour il permet principalement trois opérations (explicitées par la figure 3).

Il permet d'importer un texte depuis un fichier (ou un corpus depuis plusieurs fichiers) , en réalisant plusieurs sous-tâches séquentielles : le nettoyer, l'étiqueter[4] et le filtrer[5] morphologiquement, annoter sémantiquement tous les items textuels cibles (avec par défaut les expressions figées qui sont répertoriées et repérées dans les textes) et le représenter informatiquement et de façon distributionnelle.

Une fois cette première étape réalisée, il est possible d'exporter le texte (ou les textes) sous plusieurs formats, ainsi que leurs index et distributions. On peut également calculer et exporter des mesures statistiques sur les distributions.

---

[2] Tous les éléments de la forme brute du texte sont répertoriés à l'aide d'une clé unique : les mots, certaines expressions figées, la ponctuation, les changements de paragraphes sans perte d'information afin de pouvoir recréer la forme originale du texte à tout moment.

[3] Par exemple si le mot 'chat' apparaît cinq fois dans un texte de 3 paragraphes, une fois dans le n° 1 et quatre fois dans le n° 2, alors sa représentation distributionnelle est {0:1,1:4,2:0}.

[4] L'étiquetage morphologique s'effectue avec l'étiqueteur TreeTagger : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[5] Par défaut, les noms communs, les verbes, les adjectifs, les adverbes et les syntagmes sont étudiés, au contraire de la ponctuation ou des mots fonctionnels.

Enfin on peut comparer « distributionnellement » deux textes entre-eux, et exporter les résultats.

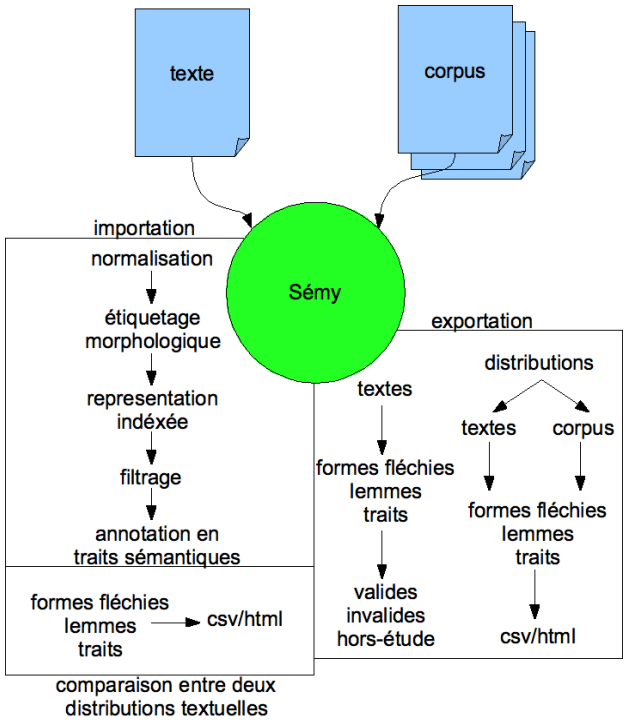

Figure 3 : schéma de fonctionnement

2 Applications sur corpus

La plateforme a déjà été utilisée pour annoter plusieurs corpus. Lors d'une première expérience, nous avons étudié les traits sémantiques d'une sélection d'articles du Monde Diplomatique [Auteur, 2007]. D'autre part, nous relatons une seconde expérience. La plateforme d'annotation a été utilisée sur un second corpus constitué de 300 titres de la presse française à propos de Ségolène Royal et Nicolas Sarkozy, soit un corpus assez homogène pour chacun des personnages, 9117 mots pour Sarkozy et 9477 pour Royal [Auteur2, 2007]. L'annotation sémantique en traits sémantiques assurée par la plateforme fait apparaître plusieurs éléments. Comme la plateforme permet d'étudier les fréquences tant des formes que des traits, il a été possible d'extraire facilement les verbes les plus fréquents autour de chaque personnage. Il apparaît alors une forte dysymétrie entre eux. Les verbes fréquents au côté de Royal décrivent des activités très générales (*révéler, montrer, dire, signifier, déclarer, signaler, énoncer*) alors que les verbes fréquents autour de Sarkozy décrivent des activités précises comme *embaucher, signer*. Du point de vue des champs sémantiques impliqués, on note une forte présence des verbes relationnels (*annoncer, susciter*) et de campagne (*élire, voter*) chez Royal, à comparer avec les champs sémantiques impliqués avec Sarkozy, la réalisation intellectuelle ou concrète (*réaliser, décider*) ou la présence d'un danger (*défendre, inquiéter*).

D'autre part, la même plateforme a permis d'étudier les fréquences de traits autour des deux personnages. Une comparaison des traits fréquents associés aux verbes fréquemment cooccurrents de chaque candidat entre en résonance avec les sondages d'opinion qui ont été largement diffusés à cette période : la compétence et la stature d'un homme d'état pour Sarkozy, l'écoute pour Royal. En analysant les fréquences de traits sémantiques calculées par la plateforme, Sarkozy apparaît comme un entrepreneur (/administrer/, /remplir/, /charger/, /exécuter/, etc) et Royal comme une militante, autrement dit comme caractérisée par une qualité associée habituellement à la condition féminine.

3 Perspectives et objectifs

Les développements futurs de la plateforme vont se concentrer, principalement autour du module de statistiques (en apprentissage, classification, reconnaissance, recherche et extraction d'informations), en restant cohérent avec les contraintes linguistiques issues de la sémantique interprétative et celles de la linguistique de corpus . Étudier l'aspect de la mise-à-jour (ou l'enrichissement ou encore le raffinage) du lexique mais aussi des textes et des corpus annotés, soit en d'autres termes de l'aspect évolutif de la plateforme elle-même, apparaît également comme intéressant du point de vue théorique. L'idée serait de se rapprocher itérativement du couple (lexique, corpus de référence) le plus cohérent, le plus efficace sur les plans de la sémantique, la linguistique, l'informatique et des statistiques.

A l'heure actuelle, la plateforme sait extraire du TLFi un premier lexique et annoter des textes avec celui-ci. Un stage a eu pour objet la réduction de l'hétérogénéité du lexique[6] sur les plans lexicologiques et lexicographiques. Nous envisageons des procédures pour classifier le lexique en s'appuyant sur [Valette et al, 2006], mais aussi les corpus (et les textes). A plus long terme, nous souhaitons détecter les isotopies textuelles[7] et lier de façon plus fine les corpus et le lexique.

Bibliographie

[Auteur1, 2007] Auteur1 (2007) :

[Auteur2, 2007] Auteur2. (2007) : « Féminisation des noms de métiers, discours journalistique : Une grande victoire ou une petite concession ? » *SILF 2007 : XXXIe colloque international de linguistique fonctionnelle*, Université Saint-Jacques de Compostelle, Espagne, Septembre 2007.

[M. Caillet, J-F. Pessiot, M-R. Amini, P. Gallinari, 2004] Unsupervised Learning with term clustering for Thematic segmentation of texts, In RIAO 2004, 26-28 Avril 2004, Avignon, France.

[P. Enjalbert, 2005] Sémantique et traitement automatique des langues Hermès Sciences.

[P. Enjalbert & B. Victorri, 2005] Les paliers de la sémantique. Chapitre 2 du document [Enjalbert, 2005].

[M. Rossignol & P. Sébillot, 2006] Acquisition sur corpus non spécialisés de classes sémantiques thématisées, In Jean-Marie Viprey, editor, 8èmes Journées internationales d'Analyse Statistiques des Données Textuelles (JADT 2006), Besançon, France.

[M. Valette & al, 2006] M. Valette, A. Estacio-Moreno, E. Petitjean, E. Jacquey, « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », Verbum ex machina, Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06). Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds). Cahiers du CENTAL, 2.1, UCL Presses Universitaires de Louvain. Volume 1. Pages 357-366).

[6] Dans le cadre d'un stage de Master 2 Recherche, été 2007, Egle Ramdani a jeté les premiers jalons de ce travail.

[7] Les traits sémantiques récurrents qui participent à la cohésion sémantique d'un texte.