EXCOM : Plate-forme d'annotation sémantique de textes multilingues

Motasem Alrahabi (1), Jean-Pierre Desclés (2)

Laboratoire LaLIC – Université Paris-Sorbonne
(1) motasem.alrahabi@paris4.sorbonne.fr
(2) jean-pierre.descles@paris-sorbonne.fr

Résumé Nous proposons une plateforme d'annotation sémantique, appelée « EXCOM ». Basée sur la méthode de l' « Exploration Contextuelle », elle permet, à travers une diversité de langues, de procéder à des annotations automatiques de segments textuels par l'analyse des formes de surface dans leur contexte. Les textes sont traités selon des « points de vue » discursifs dont les valeurs sont organisées dans une « carte sémantique ». L'annotation se base sur un ensemble de règles linguistiques, écrites par un analyste, qui permettent d'identifier les représentations textuelles sous-jacentes aux différentes catégories de la carte. Le système offre, à travers deux types d'interfaces (développeur ou utilisateur), une chaîne de traitements automatiques de textes qui comprend la segmentation, l'annotation et d'autres fonctionnalités de post-traitement. Les documents annotés peuvent être utilisés, par exemple, pour des systèmes de recherche d'information, de veille, de classification ou de résumé automatique.

Abstract We propose a platform for semantic annotation, called "EXCOM". Based on the "Contextual Exploration" method, it enables, across a great range of languages, to perform automatic annotations of textual segments by analyzing surface forms in their context. Texts are approached through discursive "points of view", of which values are organized into a "semantic map". The annotation is based on a set of linguistic rules, manually constructed by an analyst, and that enables to automatically identify the textual representations underlying the different semantic categories of the map. The system provides through two sorts of user-friendly interfaces (analyst or end-user) a complete pipeline of automatic text processing which consists of segmentation, annotation and other post-processing functionalities. Annotated documents can be used, for instance, for information retrieval systems, classification or automatic summarization.

Mots-clés : Plate-forme, Annotation automatique, Exploration Contextuelle, analyse sémantique, marqueurs discursifs, carte sémantique, multilinguisme.

Keywords: Platform, automatic annotation, Contextual Exploration, semantic analysis, discoursive markers, Semantic Map, multilingual approach.

Quelques principes théoriques

Les outils d'annotation sémantique existants sont basés de manière générale sur deux types de techniques : une technique à base de règles (ex. MUSE, SemTag), et une technique à base d'apprentissage (MnM, Ont-O-Mat/Amilcare…). Ces techniques nécessitent souvent une phase de prétraitement qui comporte une tokenisation, une analyse morpho-syntaxique, une reconnaissance d'entités nommées (voir la chaîne de traitement ANNIE de Gate). Notons aussi que l'approche 'sémantique' dans ces plateformes considère la langue comme un système de nomenclatures, 'nettoyée' des mots 'vides', et composé essentiellement de syntagmes nominaux. Notre approche, basée sur la méthode de l'Exploration Contextuelle (EC) (Desclés, 2006), consiste plutôt à appréhender la langue comme un système d'opérations portant sur des entités et sur des relations entre des entités. Ces relations sont exprimées par des prédicats verbaux, des prépositions, des déterminants, des adjectifs, etc. La plateforme que nous proposons, EXCOM (EXploration COntextuelle Multilingue), offre le moyen d'annoter des 'relations' sémantiques liées à la composante discursive de la langue et qui correspondent à des 'points de vue' (pdv) de fouille (*Qui a rencontré qui ?*, *Quelle est la cause de tel fait ?*, *Comment tel objet est-il apprécié par un tel public ?* etc.) ou à des pdv de mise en texte (idées principales d'un article, conclusions, hypothèses, résultats etc.). Les marqueurs des pdv discursifs sont des unités linguistiques observables, relativement indépendants d'un domaine particulier. Ils sont analysés dans leur contexte à l'aide d'un système de règles linguistiques écrites par l'analyste (linguiste ou expert du domaine). Notre technique ne se base pas sur des principes probabilistes ou d'apprentissage, et ne fait aucun appel à des prétraitements morpho-syntaxiques. Un pdv est mis en œuvre par l'analyse et la modélisation de besoins concrets d'utilisateurs, et par l'organisation des concepts et notions sous-jacents dans une 'carte sémantique' (une ontologie). Les nœuds de la carte sont interliés par différents types de relations (ex. spécification, composition, incompatibilité, etc.). Chaque nœud de la carte correspond à une valeur sémantique unique dont les occurrences sont représentées dans les textes par des unités linguistiques que nous appelons 'indicateurs' d'un pdv. Or, un indicateur n'est pas toujours suffisant pour souligner de manière biunivoque une relation sémantique (ambiguïté, indétermination, polysémie…), d'où le besoin de recourir à d'autres unités complémentaires dans le contexte, appelés 'indices'. Les indicateurs et indices peuvent être de toute catégorie lexicale (verbes, noms, adverbes…). Les indices sont soit 'positifs', soit 'négatifs', et peuvent ainsi, respectivement, confirmer ou infirmer la plausibilité -signalée par la présence de l'indicateur- du statut de la relation sémantique. L'analyse de textes permet de décrire l'agencement des indicateurs et des indices discursifs liés à une catégorie donnée par des constructions qui ne se réduisent pas à de simples schémas syntaxiques, et de les exprimer dans un formalisme de règles d'EC.

Chaîne de traitements

La nouvelle version d'EXCOM (Alrahabi, Desclés 2008)[1] est composée d'un moteur d'annotation, connecté à plusieurs modules supplémentaires de prétraitement et de post-traitement de données linguistiques (les techniques utilisées sont principalement JAVA et XML). Le prétraitement consiste à segmenter le corpus traité afin d'avoir une structure homogène de textes, dans laquelle sont définis les 'espaces de recherche' des unités linguistiques ainsi que les espaces destinés à être annotés. Le module de segmentation permet

[1] La première version d'EXCOM (Djioua et al., 2006) prend la suite des plateformes SERAPHIN et SAFIR pour le résumé automatique et de ContextO pour la fouille de textes. Voir par exemple (Desclés, Minel, 2004).

ainsi de définir, dans des fichiers en format de texte brut, les titres, sections, paragraphes et segments (*phrases*), et de produire des fichiers XML (DTD DocBook). La segmentation prend en compte le caractère multilingue des données, comme l'absence de majuscules ou d'espaces dans certaines langues ou le besoin de désambiguïser certains signes polysémiques, comme le point, dans d'autres. Les textes sont ensuite annotés selon la carte sémantique du pdv traité. La présence d'un indicateur dans un espace de recherche déclenche l'exécution des règles d'EC associées : les prémisses de la règle sont alors vérifiées (présence, absence, position et/ou ordre des indices dans le contexte). Si toutes les conditions sont satisfaites, une annotation est alors attribuée à l'espace destiné à être annoté. L'annotation porte des informations comme le nom de la catégorie sémantique et son positionnement dans la carte, les coordonnées du segment annoté dans le texte d'origine, etc. Le moteur d'EXCOM (Figure 1) manipule différents types de règles, en fonction du format de l'indicateur (unité linguistique, expression régulière ou annotation déjà attribuée) ; de l'espace de recherche continu ou discontinu (ex. entre titre et paragraphe suivant) et de l'espace destiné à être annoté (segment ou titre, ou bien un bloc supérieur). Un ordre de priorité peut être précisé dans le déclenchement des règles selon leur importance. Le système permet, après l'annotation du corpus, de visualiser le graphe de la carte avec, sur chaque nœud, un lien vers une base d'annotations attribuées à cette catégorie. Un autre outil (en cours) permettra la recherche de mots-clefs au sein de ces bases d'annotations. Deux types d'interface sont disponibles sous EXCOM : des interfaces dédiées à l'analyste qui effectue la construction et l'intégration des ressources linguistiques (marqueurs et règles au format XML), l'organisation de la carte et qui effectue les tests et évaluations du système ; d'autre part, EXCOM peut être utilisé par un utilisateur 'final', qui, à l'aide d'autres interfaces, exploite les pdv déjà implémentés. Celui-ci n'intervient pas dans la chaîne du traitement, il a juste à choisir un corpus dans une langue ou dans une autre, y lancer l'annotation en fonction des pdv disponibles dans le système, consulter les résultats du traitement et/ou les filtrer à l'aide de mots-clés. Une version basique d'EXCOM est librement accessible à l'adresse http://www.excom.fr/, où l'on trouve aussi des publications renvoyant aux différents pdv discursifs développés dans plusieurs langues (*causalité, événements, hypothèses, bibliosémantique, rencontres, résumés*, etc.). Notre démonstration portera notamment sur un pdv qui consiste à identifier et catégoriser les citations en arabe, en français et en coréen, selon une carte sémantique organisée autour de la notion de prise en charge énonciative.

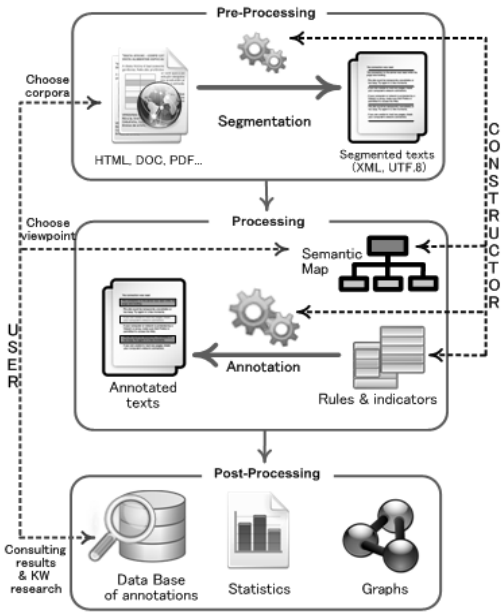

Figure 1 : architecture modulaire d'EXCOM

Références

DESCLES J.-P. (2006). Contextual Exploration Processing for Discourse Automatic Annotations of Texts. Actes de *FLAIRS 2006*, 281-284.

ALRAHABI M., DESCLES J.-P. (2008). Automatic annotation of direct reported speech in Arabic and French, according to semantic map of enunciative modalities. Actes de *GOTAL 2008*, 40-51.