Lexique multilingue dans le cadre du modèle linguistique Compreno développé par ABBYY

Elena Kozlova Maria Gontcharova Tatiana Popova
ABBYY, 2B rue Otradnaya, Moscou, Russie
Helen_Koz@abbyy.com maria_go@abbyy.com Tatiana_P@abbyy.com

RÉSUMÉ

Le lexique multilingue basé sur une hiérarchie sémantique universelle fait partie du modèle linguistique Compreno destiné à plusieurs applications du TALN, y compris la traduction automatique et l'analyse sémantique et syntaxique. La ressource est propriétaire et n'est pas librement disponible.

ABSTRACT

Multilingual lexical database in the framework of COMPRENO linguistic model developed by ABBYY

The multilingual lexical database based on the universal semantic hierarchy is part of Compreno linguistic model. This model is meant for various NLP applications dealing with machine translation, semantic and syntactic analysis. The resource is private and is not freely available.

MOTS-CLÉS : Lexique multilingue, hiérarchie sémantique universelle, traduction automatique.
KEYWORDS: Multilingual lexical database, universal semantic hierarchy, machine translation.

Nous présentons le composant sémantique du modèle linguistique Compreno. Ce modèle comprend 4 modules interdépendants : morphologique, sémantique, syntaxique, statistique, et dispose non seulement des mécanismes de désambiguïsation, mais aussi d'un large éventail d'outils pour traiter l'asymétrie translinguistique (Manicheva et al., 2012). Pour le moment la description de l'anglais (99000 classes lexicales) et du russe (87000 classes lexicales) est presque terminée ; la description du français (11500 classes lexicales), de l'allemand (13000 classes lexicales) et du chinois (8500 classes lexicales) est en cours. À présent le système assure la traduction de haute qualité de l'anglais en russe (Anisimovich et al, 2012). Les directions GE<->RU et FR<-> RU ont été également testées en version alpha.

Le pivot du modèle est une hiérarchie sémantique universelle (HS) qui sert de cadre pour des bases de données lexicales de différentes langues naturelles. La HS est organisée comme un arbre dont les nœuds, nommés classes sémantiques (CS), sont liés par des relations d'hypéronymie/hyponymie. Les CS correspondent à la notion de champs sémantique et sont reparties en 5 branches principales : ENTITY_LIKE_CLASSES, AREA_OF_HUMAN_ACTIVITY, CHARACTERISTIC_AND_VALUE, CONDITION et SITUATION. Chaque CS ne peut avoir qu'un seul ascendant direct et hérite les propriétés de son parent. Les CS universelles comportent des classes lexicales (CL), spécifiques à chaque langue. D'une part, les CL sont des éléments de la HS, c'est-à-dire elles sont des sens, d'autre part, elles comportent des lexèmes qui proviennent du module morphologique. Les CL peuvent contenir des lexèmes de différentes parties du discours. Vu la polysémie lexicale, le même lexème peut se trouver dans plusieurs CL et hériter de leurs propriétés. De ce point de vue, il est nommé dérivé grammaticale (DG). Encore un type de descendants des CL est nommé dérivé sémantique (DS) dont le sens se compose du sens du mot principal et d'un ou de plusieurs éléments de sens supplémentaires, comme dans lire-relire (répétitivité).

Les dépendances sémantiques sont décrites dans le modèle Compreno en termes de positions sémantiques. Il y a des positions sémantiques pour les actants verbaux, pour les modificateurs adverbiaux et adjectivaux, pour les compléments circonstanciels et pour beaucoup d'autres relations sémantiques (plus de 300 positions sémantiques au total). L'ensemble des positions sémantiques typiques de chaque CS constitue son modèle sémantique profond. Les sémantèmes sont porteurs des éléments de sens universels. ***Les sémantèmes distributionnels*** servent à regrouper des CS de différentes branches ayant des propriétés similaires pour mieux décrire la compatibilité (par exemple, <<Place>> dans la CS SPACE_AND_SPATIAL_OBJECTS et la CS ORGANIZATION). ***Les sémantèmes différentiels*** aident à distinguer de différentes CL au sein d'une CS (par exemple, dans la CS INTENSITY_OF_CONDITIONS_AND_CHARACTERISTICS 'subtil' diffère de 'léger' par <<Very_High_Degree>>). Possédant un jeu de sémantèmes similaires, les CL au sein d'une même CS sont des synonymes. Elles sont des antonymes si elles diffèrent par les sémantèmes de polarité (<<Polarity_Plus>> ou <<Polarity_Minus>>).

Le modèle Compreno prévoit la description des groupes de mots à l'aide des termes, idiomes et collocations. Les termes et les idiomes sont des variétés des CL et prennent part au choix lexical au même titre que les CL. Les collocations sont prévues pour chaque paire de langues concrètes, permettent d'améliorer la traduction et augmentent la possibilité de choix d'une classe correcte.

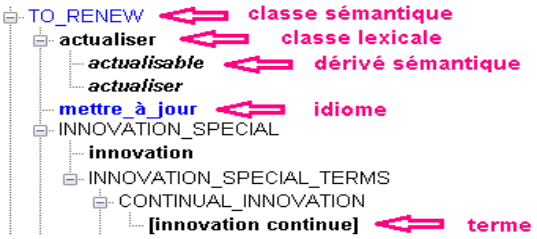

FIGURE 1 – Un fragment de la HS.

La HS a été créée à la base du russe et de l'anglais. Cependant l'ajout de nouvelles langues, même typologiquement différentes, a montré que la structure universelle ne demande pas de modifications profondes. Le mécanisme de représentativité des CS, c'est- à-dire de possibilité ou d'impossibilité pour une CS de chercher un équivalent de traduction dans son parent, permet d'éviter le regroupement infini des CS. L'ajout de nouvelles CS pour des notions uniques d'une langue donnée ne pose pas de problèmes puisque les groupes de mots décrivant de telles notions dans d'autres langues sont ajoutés comme termes ou idiomes.

Références

ANISIMOVICH, K. V., DRUZHKIN, K. Y., MINLOS, F. R., PETROVA M. A., SELEGEY V. P., ZUEV K.A. (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, *Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii 'Dialog' 2012* [Computational Linguistics and Intellectual Technologies: Proc. of the Internat. Conf. "Dialog 2012"], Bekasovo.

MANICHEVA E., PETROVA M., KOZLOVA E., POPOVA T. (2012). Compreno Semantic Model as Integral Framework for Multilingual Lexical Database. *In Proc. of the Workshop on Cognitive Aspects of the Lexicon (CogALex 2012),* Mumbai, India.