Inbenta Semantic Search Engine : un moteur de recherche sémantique inspiré de la Théorie Sens-Texte

Manon Quintana
INBENTA FR 164 route de Revel 31400 Toulouse
mquintana@inbenta.com

RÉSUMÉ

Avec la digitalisation massive de documents apparaît la nécessité de disposer de systèmes de recherche capables de s'adapter aux habitudes de recherche des utilisateurs et de leur permettre d'accéder à l'information rapidement et efficacement.

INBENTA a ainsi créé un moteur de recherche intelligent appellé *Inbenta Semantic Search Engine* (ISSE). Les deux tâches principales de l'ISSE sont d'analyser les questions des utilisateurs et de trouver la réponse appropriée à la requête en effectuant une recherche dans une base de connaissances. Pour cela, la solution logicielle d'INBENTA se base sur la Théorie Sens-Texte qui se concentre sur le lexique et la sémantique.

ABSTRACT

Inbenta Semantic Search Engine: a semantic search engine inspired by the Meaning-Text Theory

The need to have search systems able to adapt themselves to the particular way users pose their questions so that they can get a quick and efficient access to information is increasingly relevant due to the huge digitalization of documents.

To cope with this reality, INBENTA has developed an intelligent search engine, called Inbenta Semantic Search Engine (ISSE). ISSE's main two tasks are analysing users' queries and finding the most appropriate answer to those questions in a knowledge-base. To carry out these tasks, INBENTA's software solution relies upon the Meaning-Text Theory, which focusses on the lexicon and semantics.

MOTS-CLÉS : Moteur de Recherche Sémantique, Théorie Sens-Texte, fonction lexicale
KEYWORDS: Semantic Search Engine, Meaning-Text Theory, lexical function

1 Théorie Sens-Texte

Nous considérons que le système idéal d'accès à l'information doit traiter le besoin de l'utilisateur à un niveau sémantique. La sémantique permet en effet d'améliorer la précision des résultats. En se basant sur les études faites par l'Observatoire de Linguistique Sens-texte (OLST) de l'Université de Montréal et le Laboratoire de Phonétique, Lexicologie et Sémantique (Flexsem) de l'Université Autonome de Barcelone, INBENTA a intégré les idées sous-jacentes à la théorie Sens-Texte de I. Mel'čuk en incorporant à ses descriptions linguistiques la notion de fonction lexicale.

Les Fonctions Lexicales sont spécialement désignées pour représenter formellement les relations entre les mots, et, par conséquent, elles nous permettent de formaliser et de décrire de manière simple le complexe réseau des relations lexicales que présente le langage.

2 Ressources linguistiques chez INBENTA

INBENTA travaille depuis plus de 8 ans dans le traitement automatique des langues et dispose d'une vaste base de connaissances et de données linguistiques de plus de 11 langues. Pour analyser le langage naturel, l'ISSE comprend une série de ressources linguistiques propres comme un correcteur orthographique, un module de désambiguisation, des dictionnaires électroniques génériques et spécialisés et un moteur de traitement de langage naturel.

2.1 Dictionnaires

Les dictionnaires éléctroniques d'INBENTA sont, sans nul doute, un des points clés du fonctionnement du moteur de recherche intelligent ISSE.

Au niveau de la macrostructure, celui-ci est uniquement composé d'unités lexicales. C'est-à-dire que chaque entrée du dictionnaire correspond à un triplet constitué d'une forme (ou un paradigme de formes), d'un sens et d'une combinatoire. Actuellement, le dictionnaire général du français d'INBENTA compte près de 20 481 unités lexicales qui donne lieu à 160 181 formes fléchies.

Au niveau microstructurel, chaque unité lexicale est associée à différents types d'information lexicographiques: le paradigme flexionnel, la catégorie grammaticale et une des informations essentielles et la plus novatrice, les champs d'information dédiés à la combinatoire de l'unité lexicale.

2.2 Fonctions lexicales

Actuellement, nos dictionnaires rassemblent des informations de type paradigmatique et syntagmatique sous les fonctions lexicales suivantes :

Syn: décrit les relations synonymiques entre les unités lexicales → ***Syn****(wifi)=internet*

N_0: représente la nominalisation des unités lexicales → ***N_0****(voler)=vol*

V_0: représente la verbalisation → ***V_0****(voyage)=voyager*

A_2: représente le dérivé sémantique adjectival → ***A_2****(ouvrir)=ouvert*

Oper: verbe support permettant de verbaliser un complément → ***Oper****(âme)=rendre*

Le moteur de recherche sémantique est capable de regrouper les signifiés équivalents ou proches indépendamment de leur signifiant. L'ISSE reconnaîtra ainsi que « véhicule », « auto », « voiture » sont sémantiquement liés et pourra les regrouper de façon efficace pour l'analyse.

3 Perspectives d'évolution de la solution

Il existe environ 70 fonctions lexicales standards dans la théorie Sens-texte. Notre principal objectif est de sélectionner les fonctions lexicales qui enrichiront la description de nos unités lexicales et amélioreront le moteur de recherche sémantique d'INBENTA.